Identificação de transferência gênica horizontal em procariotos, estudo do regulon SOS de *Caulobacter crescentus* e análise estatística de colocalização de posições genômicas.

Apuã César de Miranda Paquola

Tese apresentada ao Programa Interunidades em Bioinformática da Universidade de São Paulo para a obtenção do título de Doutor em Bioinformática.

São Paulo
São Paulo - Brasil
Dezembro - 2009

Identificação de transferência gênica horizontal em procariotos, estudo do regulon SOS de *Caulobacter crescentus* e análise estatística de colocalização de posições genômicas.

Apuã César de Miranda Paquola

Orientador: Prof. Dr. **Carlos Frederico Martins Menck**

Tese apresentada ao Programa Interunidades em Bioinformática da Universidade de São Paulo para a obtenção do título de Doutor em Bioinformática.

São Paulo
São Paulo - Brasil
Dezembro - 2009

À Pamela.

Agradecimentos

Agradeço imensamente ao prof. Menck por esta orientação, durante a qual ele sempre foi crítico, mas também foi muito paciente. Sempre gostou de ideias novas e sempre teve paixão pela ciência. Muito obrigado por todo o apoio, confiança e amizade.

Aos meus pais, pelo apoio, pelo carinho e por me ensinarem o valor de se ter um espírito crítico.

À CAPES, à AB^{3}C, à PRPG-USP e à CPG-Bioinformática pelo apoio financeiro.

À Patricia da CPG, que amansa a fera da burocracia.

Aos Profs. Sergio Verjovski-Almeida e Eduardo Reis, com quem tive o prazer e o privilégio de trabalhar no início da minha formação em Bioinformática.

Ao Prof. Carlos Bragança Pereira por suas sugestões com relação a análises estatísticas.

Ao Dr. Alysson Muotri pela oportunidade de fazer pós-doc em seu laboratório.

Aos meus colegas de laboratório pela convivência e pela amizade.

À Wanessa por muitas ideias e discussões sobre transferência gênica horizontal.

À Raquel e ao Rodrigo, pelo trabalho em conjunto no sistema SOS de *Caulobacter crescentus*.

Ao Stephano, à Ana Carolina, ao André, à Marinalva, ao Milton, ao Robson e ao Ricardo Vêncio pela amizade e por muitas discussões interessantes.

Ao Caiubi, à Raquel, à Vivi e ao Paulão também pela amizade e por muitas discussões interessantes.

À Cláudia, que teve um papel importante para que eu tivesse contato com a Bioinformática.

À Pamela, minha companheira de tudo, sempre sincera e forte, pelo amor e pelo carinho.

Sumário

Resumo

Esta tese consiste em três estudos relacionados com genomas procarióticos. O primeiro estudo trata do desenvolvimento e da aplicação de um método baseado em buscas de similaridade de sequências para a identificação em larga escala de genes potencialmente envolvidos em transferência gênica horizontal (TGH). O método usa a distância entre os rRNA 16S dos organismos em estudo como estimativa da distância filogenética entre eles. Genes com alvos de alta pontuação em organismos atipicamente distantes são considerados candidatos a TGH. A aplicação deste método a 408 genomas procarióticos anotados nos permite concluir que: (i) genes com funções operacionais estão significativamente enriquecidos, enquanto que genes com funções informacionais estão significativamente empobrecidos em candidatos a TGH; (ii) há uma forte correlação entre o tamanho do genoma e a proporção de genes candidatos a TGH, que indica que a proporção de TGH é menor em genomas pequenos e maior em genomas grandes; (iii) a hipótese da complexidade, que diz que genes cujos produtos participam de poucas interações proteicas são mais suscetíveis a TGH que aqueles que participam de muitas interações, está de acordo com nossas predições de TGH em dois estudos de interação proteína-proteína em *Escherichia coli*. O segundo estudo consiste na identificação de genes pertencentes ao regulon SOS de *Caulobacter crescentus* usando uma estratégia iterativa de predições *in silico* de sítios de ligação da proteína LexA no DNA e experimentos de laboratório para medir a expressão diferencial, entre a cepa selvagem e uma cepa deficiente em *lexA*, dos genes próximos aos sítios preditos de LexA. Esta estratégia nos permitiu identificar 37 genes pertencentes a este regulon. O terceiro estudo trata de um teste estatístico para a colocalização de posições genômicas. Dados dois conjuntos X e Y de posições genômicas, o método verifica se um número significativo de elementos de Y está posicionado próximo a elementos de X. Este método, aplicado às predições de sítios de ligação de LexA no genoma de *C. crescentus*, resultou na identificação de uma quantidade significativa de sítios preditos de LexA próximos aos sítios preditos de início de tradução deste genoma. Alguns destes sítios não haviam sido identifi-

cados no segundo estudo por terem pontuação inferior ao limiar efetivamente usado naquele estudo. O posicionamento destes sítios sugere que os genes correspondentes possam ter regulação dependente de LexA.

Palavras chave: transferência gênica horizontal, transferência gênica lateral, funções de genes, tamanho do genoma, hipótese da complexidade, *Caulobater crescentus*, regulon SOS, teste estatístico, colocalização, posições genômicas.

Abstract

This thesis comprises three studies related to prokaryotic genomes. The first study consists of the development and application of a sequence similarity-based method for large-scale identification of genes potentially involved in horizontal gene transfer (HGT). The method uses 16S rRNA sequence distances as estimates of phylogenetic distances between organisms. Genes having high-scoring similarity-search hits in atypically distant organisms are taken as HGT candidates. The application of this method to 408 annotated prokaryotic genomes allowed us to conclude that (i) operational categories (transport, metabolism, mobile elements and DNA restriction/modification) are significantly enriched in HGT candidates whereas informational categories (protein synthesis, protein fate, transcription, DNA metabolism) are significantly depleted in HGT candidates; (ii) there is a strong correlation between genome size and HGT proportion, indicating that the proportion of HGT candidates tend to be lower in smaller genomes and higher in bigger ones; (iii) the complexity hypothesis, which states that genes whose products have few interaction partners are more prone to HGT than those with many partners, is supported by our HGT predictions for data from two protein-protein interaction studies in *Escherichia coli*. The second study is about the identification of genes belonging to the SOS regulon in *Caulobacter crescentus*. We employed an iterative strategy of *in silico* predictions of LexA binding sites and differential gene expression measurements, between wild-type and *lexA*-deficient strains, for genes carrying predicted LexA binding sites in their promoter regions. This strategy allowed us to identify 37 genes belonging to this regulon. In the third study, we have developed a statistical test for the colocalization of genomic positions. Given two sets X and Y of genomic positions, the method verifies if a significant number of Y elements is positioned in close proximity to X elements. This method, applied to the LexA binding-site predictions from the second study, resulted in the identification of a significant number of predicted LexA binding sites near translation start sites in the *C. crescentus* genome. Some of these sites were not identified in the second study because their scores according to

the LexA binding site model were lower than the threshold effectively used in that study. The positioning of these sites suggests that the corresponding genes may be regulated in a LexA-dependent manner.

Keywords: horizontal gene transfer, lateral gene transfer, gene functions, genome size, complexity hypothesis, *Caulobater crescentus*, SOS regulon, statistical test, co-localization, genomic positions.

Prefácio

Esta tese consiste em três projetos de natureza e objetivos distintos: (1) análise em larga escala de transferência gênica horizontal em procariotos; (2) identificação de genes SOS de *Caulobacter crescentus*; e (3) um teste estatístico para a colocalização entre dois conjuntos de posições genômicas. Para maior clareza na apresentação, dividimos o conteúdo em capítulos independentes.

A ordem de apresentação não reflete necessariamente a ordem cronológica em que os projetos foram desenvolvidos. O primeiro capítulo não só apresenta uma maior complexidade de informação, como também é o que consideramos ter a maior contribuição científica, seja no desenvolvimento de ferramentas computacionais, seja na informação biológica gerada.

1 Identificação de transferência gênica horizontal em procariotos

1.1 Introdução

1.1.1 Transferência gênica vertical e horizontal

A transmissão de material genético de um organismo para sua progênie durante a reprodução é denominada transferência gênica vertical ou por descendência. Por ser esta a forma principal de herança de material genético, é possível conceber a organização dos seres vivos numa árvore genealógica na qual as espécies surgem pela divergência dos organismos a partir de seus ancestrais. Os esforços para se determinarem relações filogenéticas entre os seres vivos, que inicialmente baseavam-se em características morfológicas e fisiológicas, passaram a contar, a partir do desenvolvimento de métodos de sequenciamento, com uma fonte muito rica em informação filogenética: as sequências de DNA e proteínas.

O componente de RNA da subunidade menor do ribossomo (SSU rRNA, também chamado de rRNA 16S em procariotos ou rRNA 18S em eucariotos) foi usado por Woese e colaboradores num estudo que estabeleceu a separação filogenética entre bactérias, arqueas e eucariotos (Woese e Fox, 1977), e também na construção de uma filogenia de procariotos (Fox *et al.*, 1980). Esta molécula, que está universalmente presente nos organismos celulares e possui uma taxa lenta de evolução, é usada extensivamente em reconstruções filogenéticas e na estimativa da diversidade microbiológica em projetos de metagenomas. Em procariotos, árvores filogenéticas baseadas no rRNA 16S são geralmente consistentes com as obtidas por métodos que

combinam características extraídas de todo o genoma, o que sugere que as relações filogenéticas entre procariotos estejam geralmente bem representadas nas sequências do rRNA 16S (Delsuc *et al.*, 2005).

Árvores filogenéticas construídas a partir de sequências de proteínas individuais são, no entanto, muitas vezes incongruentes entre si ou com as árvores construídas a partir do rRNA 16S. Em geral, observa-se uma combinação de duas situações: a ausência da proteína estudada em parte das espécies e a incompatibilidade topológica entre as árvores. No exemplo hipotético da Fig. 1.1, a espécie Z tem posicionamento diferenciado nas duas árvores, formando, na do rRNA 16S, um grupo monofilético com a espécie Y, e, na do gene "g", com a espécie A. Neste caso, assumindo-se que a árvore do rRNA 16S siga a filogenia das espécies, uma explicação possível para a incongruência entre as árvores é a transferência (horizontal) do gene "g" da espécie A para a espécie Z, com a perda do ortólogo original, como está ilustrado na Fig. 1.2.

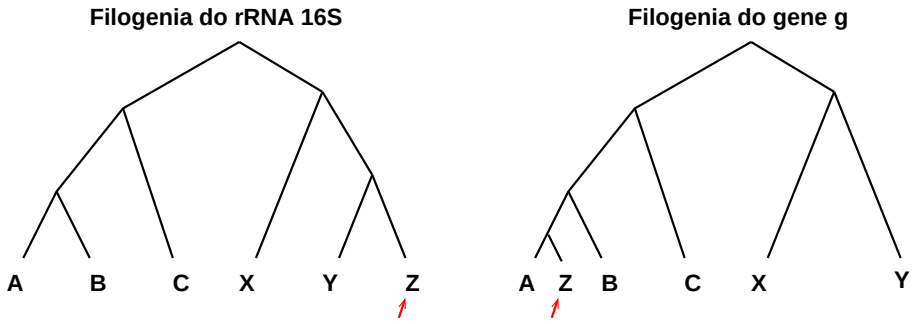

Figura 1.1 - Filogenias hipotéticas das espécies A, B, C, X, Y e Z para o rRNA 16S e o gene hipotético "g". As setas vermelhas indicam o posicionamento da espécie Z nestas duas árvores.

Transferência gênica horizontal (TGH) é a denominação dada às formas de transferência de material genético em que, diferentemente da descendência vertical, os organismos envolvidos não estão no papel de progenitor e progênie. Segundo esta

definição, a TGH pode envolver organismos de uma mesma espécie ou de espécies diferentes.

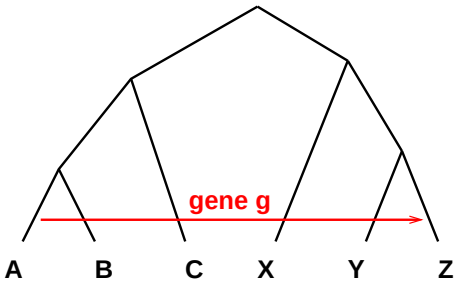

Figura 1.2 - Transferência horizontal do gene "g" da espécie A para a espécie Z.

1.1.2 Mecanismos de transferência gênica horizontal

A transferência horizontal de genes ocorre por meio de transformação, conjugação ou transdução (Jain *et al.*, 2002). Transformação é o processo pelo qual o organismo adquire DNA livre do meio externo e o incorpora ao seu genoma. A aquisição do DNA envolve uma série de etapas: ligação do DNA à superfície da célula, transporte através da membrana externa (em bactérias gram-negativas), fragmentação (em bactérias gram-positivas), degradação de uma das fitas e transporte da outra para o citoplasma (Chen e Dubnau, 2004). A competência genética, estado no qual o procarioto é capaz de adquirir DNA do ambiente, é, na maioria das espécies naturalmente competentes, induzida em condições ambientais específicas, embora possa ser também constitutiva, como no caso de *Neisseria gonorrhoeae* (Solomon e Grossman, 1996). Os sinais que promovem ou reprimem o desenvolvimento da competência variam de espécie para espécie, como também variam as respostas para cada sinal. Dentre estes sinais, podemos citar: carência nutricional, variação química do meio externo, entrada na fase estacionária, alta densidade populacional, danos no DNA e presença de antibióticos (Claverys *et al.*, 2006; Solomon e Grossman, 1996). O DNA adquirido pode ser usado pela célula como fonte de nutrientes (carbono, ni-

trogênio e fósforo), molde para reparo de DNA, material para recombinação, ou para a integração em seu genoma (Solomon e Grossman, 1996; Dubnau, 1999). Nos dois últimos casos, pode haver a incorporação de material genético externo no genoma do organismo receptor.

A conjugação é a transferência de material genético entre procariotos através do contato direto entre duas células, que assumem os papéis distintos de doadora e receptora. A célula doadora é aquela que possui um elemento genético transmissível, geralmente um plasmídeo ou transposon conjugativo, que é o material a ser transferido para a célula receptora. A conjugação do plasmídeo F de *Escherichia coli*, um dos mecanismos de conjugação mais estudados, pode ser resumida nas seguintes etapas: (i) formação do *pilus* na célula doadora, (ii) ligação do *pilus* à célula receptora, (iii) aproximação das duas células, (iv) transferência de uma das fitas do plasmídeo conjugativo da célula doadora para a receptora, (v) recircularização dos plasmídeos e síntese das fitas complementares em ambas as células. Como as proteínas que compõem a maquinaria de conjugação são codificadas em genes no próprio plasmídeo F, a célula receptora torna-se uma doadora após a conjugação (Griffiths *et al.*, 2004). A mobilização de elementos transponíveis e a recombinação entre um cromossomo da célula receptora e o plasmídeo conjugativo podem fazer com que genes cromossomais também sejam transferidos por este mecanismo (Thomas e Nielsen, 2005).

Transdução é a transferência de DNA entre dois organismos procarióticos por meio de vírus. Dois exemplos bem estudados são os dos fagos P1 e λ de *E. coli*. No ciclo lítico do fago P1, as partículas virais formadas encapsulam, tipicamente, o DNA do próprio vírus, mas podem, ocasionalmente, encapsular fragmentos aleatórios do DNA bacteriano. O fago λ, no ciclo lisogênico, encontra-se integrado num sítio específico do cromossomo bacteriano. No ciclo lítico, a excisão errônea do profago pode levar à formação de partículas virais contendo DNA bacteriano contíguo ao sítio de inserção. A transdução se completa quando uma destas partículas virais injeta o DNA bacteriano em outra célula e este se integra ao genoma hospedeiro. As

formas de transdução observadas nos fagos P1 e λ são chamadas, respectivamente, de generalizada, pela qual fragmentos aleatórios de DNA bacteriano são incorporados em partículas virais, e de especializada, pela qual regiões específicas do DNA são preferencialmente incorporadas (Griffiths *et al.*, 2004). Os chamados agentes de transferência gênica (ATG) constituem um tipo especial de profago, e estão presentes em diversos grupos de bactérias e arqueas. Diferentemente de um profago típico, as partículas virais de um ATG em ciclo lítico não têm preferência por empacotar o próprio genoma: elas contêm, em sua maioria, fragmentos do genoma hospedeiro. Além disso, a quantidade de DNA empacotada por um ATG é muito pequena, de modo que seu genoma completo não cabe em uma única partícula viral (Stanton, 2007). Segundo Koonin e Wolf (2008), os ATGs podem ser vistos como entidades funcionais especializadas em transferência gênica horizontal.

Uma vez integrado no genoma receptor, e independentemente do mecanismo de transferência, a fixação do DNA adquirido numa população de microorganismos está sujeita à seleção natural e à deriva genética e depende do quanto este DNA é deletério, neutro ou vantajoso para seu portador. Este cenário, no entanto, pode ser influenciado por elementos presentes no próprio DNA transferido. Alguns plasmídeos conjugativos, por exemplo, usam sistemas toxina/antitoxina, cujo efeito é o de matar células que perdem o plasmídeo, como estratégia para aumentar sua frequência na população (Hayes, 2003). Outra estratégia de efeito similar é a usada por um grande número de bacteriófagos temperados: quando em ciclo lisogênico, estes fagos provêm à bactéria hospedeira imunidade contra infecções líticas causadas por fagos do mesmo tipo (Griffiths *et al.*, 2004).

1.1.3 Consequências evolutivas da TGH

A evolução se dá, segundo o modelo correntemente aceito, pelo surgimento de indivíduos com modificações aleatórias no DNA (mutações pontuais, duplicações, rearranjos e deleções de trechos do genoma) e pela seleção natural dos indivíduos mais aptos à sobrevivência e à reprodução. Dentro deste paradigma, o surgimento

de um novo gene pode decorrer da duplicação de genes pré-existentes, de rearranjos entre suas partes, combinando domínios funcionais, e de mutações em sua sequência nucleotídica. Um novo gene que permita, por exemplo, a proliferação de um organismo em um habitat inexplorado, pode ser o fator inicial para a separação deste organismo de sua linhagem parental. A separação de habitats mais o acúmulo, ao longo do tempo, de modificações no DNA, podem fazer com que a linhagem originada deste organismo tenha características bem distintas da parental, a ponto de ser reconhecida como uma nova espécie. Num modelo que considere apenas a transferência vertical, a evolução, em um organismo, conta somente com rearranjos e mutações no próprio genoma como matéria-prima para a inovação genética. Com a TGH, conjuntos de genes "prontos" podem ser transferidos de um organismo a outro, conferindo-lhe imediatamente novas funções biológicas, que podem ser, como dito, o fator inicial para um processo de especiação. Com a TGH, a matéria-prima para inovação genética passa a ser, teoricamente, o conjunto de todos os genomas existentes (Lawrence, 2002; Lawrence e Retchless, 2009), o que, por si só, ressalta a enorme importância evolutiva da TGH, sobretudo em procariotos.

1.1.4 Métodos de detecção de TGH

As diferentes condições ambientais em que vivem os organismos, como temperatura, concentrações de sal, nutrientes e substâncias tóxicas, além fatores genéticos, como a capacidade de transportar e metabolizar certos nutrientes e a abundância dos diferentes tRNAs têm influência sobre o viés mutacional de cada genoma, que reflete em parâmetros mensuráveis de composição nucleotídica. Tais parâmetros, como conteúdo G+C, frequência de dinucleotídeos, frequência de uso de códons, e modelos de Markov para frequências de nucleotídeos, podem ser vistos como assinaturas de cada genoma e usados na identificação de regiões genômicas com potencial origem horizontal (Karlin, 2001). Uma região do genoma que tenha sido transferida recentemente terá características nucleotídicas mais semelhantes às do organismo doador que às do receptor. Com base nisto, e explorando diferentes

parâmetros nucleotídicos, foram desenvolvidos diversos métodos para a detecção de potencial transferência horizontal por meio da identificação de regiões com parâmetros nucleotídicos distintos dos da média global do genoma (Karlin, 2001; Vernikos e Parkhill, 2006; Reva e Tümmler, 2005). Estes métodos têm, em geral, baixo custo computacional e necessitam somente de sequências do próprio genoma em estudo. Alguns deles, porém, tentam identificar os potenciais doadores das regiões transferidas, procurando num banco de dados de genomas, aqueles que tenham parâmetros nucleotídicos semelhantes (Nakamura *et al.*, 2004; Merkl, 2004; Waack *et al.*, 2006). Em particular, no trabalho de Nakamura *et al.* (2004), são usados preditores de ORFs baseados em modelos de Markov para a identificação de candidatos a TGH. Cada preditor de ORFs é treinado em um genoma específico e aplicado nos demais genomas em estudo. Genes com pontuação significativamente maior segundo o modelo de outro genoma que segundo o do próprio genoma em estudo são tidos como candidatos a TGH.

Após a aquisição horizontal, a região transferida passa a sofrer as mesmas pressões mutacionais do genoma hospedeiro e tende, gradualmente, ao longo do tempo evolutivo, a adquirir as características nucleotídicas deste. Tal fenômeno é denotado pela palavra inglesa *amelioration* (Lawrence e Ochman, 1997). Os métodos de detecção de TGH por composição nucleotídica são, por natureza, incapazes de identificar regiões transferidas entre genomas com parâmetros nucleotídicos semelhantes, ou regiões de transferência antiga, que já se adaptaram às características do genoma hospedeiro. Além disso, desvios em parâmetros de composição nucleotídica em algumas regiões genômicas podem ser devidos a características funcionais dos genes ali presentes e não a eventos de transferência horizontal. As proteínas ribossomais de *E. coli*, por exemplo, têm frequências de uso de códons distintas das da média do genoma (Karlin *et al.*, 1998), o que motiva alguns métodos de predição de TGH, como o de Tsirigos e Rigoutsos (2005), a excluírem esta classe de genes dos candidatos a TGH.

Para que se detectem TGHs antigas ou entre organismos com composição

nucleotídica semelhante são necessários métodos que analisem relações filogenéticas entre os organismos, buscando incongruências entre a árvore das espécies e as árvores dos genes (Fig. 1.1). Alguns destes métodos fazem reconstruções explícitas de árvores filogenéticas para cada gene do genoma em estudo, comparando-as com uma árvore de referência. Como exemplo, citamos o trabalho de Beiko *et al.* (2005), no qual se constrói, para cada família de genes ortólogos, uma árvore filogenética bayesiana. Em seguida, constrói-se uma árvore de referência com base nas arestas com bom suporte estatístico (probabilidade *a posteriori* $> 95\%$) destas árvores. Árvores de famílias de genes que têm bom suporte estatístico e que discordam da árvore de referência são tomados como indicadores de TGH.

O alto custo computacional das reconstruções de árvores filogenéticas motivou vários grupos a desenvolverem métodos alternativos, que usam, por exemplo, alvos de BLAST ou informação de presença/ausência de genes nas espécies, como critérios aproximados para a detecção de TGH. Há na literatura uma grande diversidade nas técnicas específicas usadas por estes métodos, alguns dos quais citamos aqui.

Em Pál *et al.* (2005), para cada família de genes ortólogos, mapeiam-se, numa árvore filogenética construída a partir do rRNA 16S, as espécies que possuem um representante nesta família. A seguir, para cada família, assumindo-se probabilidades *a priori* para eventos de perda gênica e de TGH, calcula-se qual o cenário mais parcimonioso: aquele que envolve múltiplas perdas gênicas ou o que envolve transferência horizontal.

Em Podell e Gaasterland (2007), usa-se uma medida de distância entre duas espécies baseada no número de nós que elas têm em comum na árvore taxonômica do NCBI (http://www.ncbi.nlm.nih.gov/). Genes com alvos de BLASTP (BLAST entre sequências de proteínas, Altschul *et al.* (1997)) com pontuação alta em organismos distantes, segundo esta medida, são considerados como potencialmente transferidos horizontalmente.

Em Kanhere e Vingron (2009), analisam-se correlações entre distância de

rRNA 16S e distâncias entre sequências de proteínas para famílias gênicas representadas no banco de dados COG (*Clusters of Orthologous Groups*, Tatusov *et al.*, 2003). Para genes com transmissão principalmente vertical é esperado que estas distâncias tenham correlação positiva. A identificação de candidatos a TGH se dá pela detecção de *outliers* na análise de correlação entre estas distâncias.

1.1.5 Funções transferidas horizontalmente e a hipótese da complexidade

Uma questão importante no estudo de TGH é a de determinar quais são as categorias funcionais de genes mais propensas à transferência horizontal. Analisando sequências de genes ortólogos em quatro genomas (*Saccharomyces cerevisiae*, *Synechocystis 6803*, *Escherichia coli* e *Methanococcus jannaschii*), Rivera *et al.* (1998) observaram que duas classes funcionais de genes, a dos chamados genes *informacionais* e a dos *operacionais*, apresentam reconstruções filogenéticas dissimilares e distribuições dissimilares de distância entre as sequências. Com base nessas análises, concluíram que os genes informacionais – grupo que inclui genes envolvidos em tradução, transcrição, replicação e síntese de tRNA, além das ATP-ases e GTP-ases vacuolares – são provavelmente menos transferidos que os genes operacionais – envolvidos em biossíntese de cofatores, envelope celular, metabolismo energético, metabolismo intermediário, síntese de ácidos graxos e fosfolipídios, síntese de nucleotídeos e funções regulatórias.

Com base na constatação de que proteínas envolvidas em tradução proteica participam, em geral, de uma rede complexa de interações com outras proteínas e RNAs, em oposição a certas enzimas de metabolismo, que participam de poucas interações, o mesmo grupo propôs a *hipótese da complexidade* como explicação para a ocorrência preferencial de TGH entre os genes operacionais. Segundo esta hipótese, genes cujos produtos participam de muitas interações têm menos chance de ser transferidos que genes cuja rede de interações é pequena (Jain *et al.*, 1999). Esta hipótese baseia-se na ideia de que a fixação de um gene transferido é mais provável quando este confere alguma vantagem seletiva aos organismos que o possuem. Uma proteína

capaz de executar uma função por si só, tem, provavelmente, uma boa chance de ser funcional no organismo receptor. Já uma proteína que precisa interagir com muitas outras pode não ser funcional ou mesmo ser deletéria no organismo receptor devido à provável incompatibilidade ou à compatibilidade parcial com os sítios de interação das proteínas nativas. Usando dados de interação proteína-proteína de *E. coli*, do estudo de Arifuzzaman *et al.* (2006), Wellner *et al.* (2007) constataram enriquecimento em predições de TGH entre genes cujos produtos interagem com poucas proteínas.

Além dos estudos de transferência horizontal em genes individuais, a análise de regiões contíguas dos genomas permitiu a descoberta das chamadas "ilhas genômicas": regiões de DNA provavelmente adquiridas por TGH contendo genes que conferem ao organismo portador capacidades específicas que aumentam seu nível de adaptação ao meio (Hacker e Kaper, 2000). Dentre as ilhas genômicas, são descritas ilhas de patogenicidade, de simbiose, de metabolismo e de resistência, dependendo das funções nelas codificadas. As ilhas genômicas são relativamente grandes (tamanho de 10 kb a 100 kb), têm composição nucleotídica distinta da do restante do genoma e carregam, muitas vezes, elementos genéticos móveis, genes de tRNA e repetições diretas (Gal-Mor e Finlay, 2006; Koonin e Wolf, 2008).

Em nosso laboratório, a pesquisa com TGH teve início com o trabalho de doutorado de Wanessa C. Lima, que consistiu no mapeamento de ilhas genômicas em *Xanthomonas campestris* e *Xanthomonas axonopodis pv. citri* e anotação das funções biológicas nelas representadas. Usando uma abordagem para detecção de TGH que combina métodos de composição nucleotídica, BLAST e reconstrução filogenética atípica, Wanessa identificou 35 ilhas em cada um destes genomas (Lima *et al.*, 2005) e classificou-as em potencialmente antigas ou recentes dependendo da proporção de genes com características nucleotídicas atípicas (Lima *et al.*, 2008). Funções relacionadas a virulência, patogenicidade e metabolismo secundário foram encontradas em considerável proporção nestas ilhas, especialmente nas de potencial transferência recente. Nas ilhas de potencial transferência antiga foram encontradas funções relacionadas ao metabolismo celular primário. A investigação mais detalhada

de duas destas ilhas genômicas levou à identificação de potencial transferência horizontal do operon de biossíntese de arginina no grupo das Xanthomonadales (Lima e Menck, 2008); e da via da quinurenina para síntese de NAD, tipicamente presente em eucariotos, nos grupos das Xanthomonas e das Flavobacteriales (Lima *et al.*, 2009).

Os estudos em larga escala, que analisam de dezenas a centenas de genomas, nem sempre concordam em quais categorias funcionais são mais presentes entre os candidatos a TGH. No trabalho de Nakamura *et al.* (2004), as categorias funcionais principalmente enriquecidas em TGH são: funções relacionadas a fagos, plasmídeos e transposons; envelope celular; patogênese; funções regulatórias e processos celulares. Em Kanhere e Vingron (2009), a categoria de metabolismo está enriquecida em TGH para as transferências entre bactérias e arqueas, enquanto que, curiosamente, para transferências entre bactérias, as funções relacionadas a tradução estão enriquecidas em TGH. No trabalho de Choi e Kim (2007), não foi detectada preferência em TGH para nenhuma categoria funcional específica.

1.1.6 TGH, o tamanho dos genomas e o conjunto "flexível" de genes

O tamanho dos genomas dos procariotos, de acordo com o que foi sequenciado até o momento, varia de $\sim$ 144 kb (do simbionte intracelular *Candidatus Hodgkinia cicadicola* (McCutcheon *et al.*, 2009)) até $\sim$ 13 Mb (da bactéria de solo *Sorangium cellulosum* (Schneiker *et al.*, 2007)). O tamanho do genoma está no centro do conflito entre duas pressões seletivas: genomas pequenos são replicados com maior rapidez enquanto que genomas grandes podem ter um repertório gênico maior para lidar com diferentes condições ambientais. Genomas de organismos de vida livre tendem a ser maiores que os de parasitas obrigatórios devido, provavelmente, a uma maior variedade de desafios ambientais a que os primeiros estão expostos. Os simbiontes intracelulares têm, em geral, genomas reduzidos (< 1 Mb), provavelmente em função de extensivas perdas gênicas devidas ao fato destes organismos viverem em um ambiente relativamente homogêneo e rico em metabólitos produzidos pela célula hospedeira.

Num estudo comparativo de genomas procarióticos, Koonin e Wolf (2008) analisaram o número de genes classificados em diferentes categorias funcionais em função do número total de genes no genoma. As categorias de metabolismo, transdução de sinal e reguladores de transcrição apresentaram um aumento consideravelmente mais rápido no número de genes que as de tradução, replicação e reparo de DNA, com o aumento do número total de genes no genoma. Estes resultados estão de acordo com a visão de que há um núcleo de genes informacionais, amplamente presentes nos procariotos, e um conjunto "flexível" de genes, relacionados à exploração de nichos metabólicos ou ambientais específicos e presentes, cada qual, em grupos específicos de procariotos. Os genes do conjunto "flexível" estão amplamente sujeitos a perdas, em função da variação nas condições ambientais, e transferências, em função das vantagens adaptativas que podem conferir aos organismos receptores. Com base nisso, e nos resultados de Koonin e Wolf (2008), é de se esperar que genomas maiores tenham uma fração maior de seus genes no conjunto "flexível" e entre os que participam de TGH. Jain *et al.* (2003), usando métodos filogenéticos, num estudo com 8 genomas, constataram que as transferências horizontais ocorrem mais frequentemente entre organismos com genomas de tamanhos similares. Nakamura *et al.* (2004), usando um método de composição nucleotídica num estudo com 116 genomas procarióticos, observaram forte correlação entre a proporção de candidatos a TGH e o número de ORFs no genoma.

1.2 Objetivos

O crescimento exponencial no número de genomas procarióticos sequenciados (Koonin e Wolf, 2008), a discordância entre diferentes estudos quanto às categorias funcionais mais propensas a TGH, e as questões da hipótese da complexidade e da relação entre tamanho do genoma e transferência horizontal motivaram o desenvolvimento do presente trabalho, que teve os seguintes objetivos:

- Desenvolver um método de baixo custo computacional para a detecção de

TGH em procariotos, com base na identificação de alvos de BLAST em organismos atipicamente distantes.

- Analisar as categorias funcionais mais e menos representadas no grupo de genes potencialmente envolvidos em TGH.

- Analisar a correlação entre número de genes no genoma e proporção de genes potencialmente envolvidos em TGH.

- Testar a hipótese da complexidade, usando dados de interações proteína-proteína obtidos da literatura.

1.3 Métodos

1.3.1 Dados de genomas

A principal fonte de dados usada neste estudo é a versão 22 do banco de dados Omniome, que faz parte do *website* "Comprehensive Microbial Resource" do J. Craig Venter Institute (Peterson *et al.*, 2001), obtido de `http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi` em outubro de 2008 e importado num gerenciador de bancos de dados MySQL. Este banco de dados contém sequências genômicas de 421 bactérias e 31 arqueas, predições de genes que codificam proteínas (também chamadas de ORFs[1]), predições de genes de RNA, categorização funcional dos genes e classificação taxonômica dos organismos.

Também estão no banco de dados comparações de sequência entre todas as ORFs obtidas pelo programa BLAST-Extend-Repraze (BER, `http://ber.sourceforge.net/`). O programa BER faz, inicialmente, uma busca BLASTP (Altschul *et al.*, 1997) de cada ORF contra uma base contendo todas as ORFs do banco de dados. Os alvos de BLAST que tiverem pontuação acima de um limite de corte são armazenados em uma minibase de sequências. Para garantir robustez frente a erros de predição das posições de início e fim da ORF de busca, a

[1]Quadro de leitura aberto, do inglês *open reading frame*.

sequência genômica de 300 nucleotídeos a 5' até 300 nucleotídeos a 3' desta ORF é alinhada com as sequências da minibase usando uma variante do algoritmo Smith-Waterman (Smith e Waterman, 1981) que traduz uma das sequências de entrada para proteína. O programa retorna as pontuações e medidas de significância (valores p) destes alinhamentos.

1.3.2 Distâncias entre os rRNA 16S dos genomas procarióticos

As distâncias entre os rRNA 16S dos genomas estudados, expressas em substituições nucleotídicas por sítio, são calculadas da seguinte forma: (i) os rRNA 16S são alinhados com o programa Kalign2 (Lassmann *et al.*, 2009), usando penalidade de *gap* terminal de 3.94, (ii) são geradas 100 réplicas de *bootstrap* do alinhamento com o programa `seqboot` do pacote PHYLIP (`http://evolution.genetics.washington.edu/phylip.html`), (iii) para cada réplica, as distâncias entre os rRNA 16S são calculadas com o programa `dnadist` do pacote PHYLIP, utilizando o modelo de distância F84 (Felsenstein e Churchill, 1996).

1.3.3 Predição de genes envolvidos em transferência horizontal

O método aqui desenvolvido tem por objetivo inferir se um gene que codifica proteína está potencialmente envolvido em TGH com base nas distâncias filogenéticas (medidas pelo rRNA 16S) aos organismos com os quais tenha alvos BER. O princípio de funcionamento do método consiste em classificar um gene como envolvido em TGH caso ele tenha (i) alvos BER somente em organismos distantes ou (ii) alvos BER em organismos distantes com pontuação maior que aqueles em organismos próximos.

Numa etapa de pré-processamento, são retidos apenas os alvos BER que correspondem a potenciais ortólogos do gene em análise, excluindo-se os alvos em parálogos ou devidos a alinhamentos de regiões pequenas, como domínios proteicos. Para isso, um alvo BER deve satisfazer os seguintes critérios:

- corresponder ao melhor alvo recíproco: caso a proteína Pa do genoma A

tenha como melhor alvo BER no genoma B a proteína Pb, o alvo Pb só será considerado se o melhor alvo de Pb no genoma A for Pa.

- possuir cobertura da sequência de busca de no mínimo 60%.
- possuir valor p da busca BER inferior a 10^{-25}

A seguir, o gene em estudo é classificado como típico, atípico ou indeterminado, com base em seus alvos BER, em ordem decrescente de pontuação, e em dois parâmetros fornecidos pelo usuário: d_{self} e $d_{distant}$ (a Fig. 1.3 exemplifica este critério de classificação para quatro genes distintos). Sejam: X um alvo BER do gene em estudo; $\bar{d_X}$ a média, entre as réplicas de *bootstrap*, das distâncias de rRNA 16S entre o organismo em estudo e o do alvo X; e A o primeiro alvo "não-*self*", ou seja, o primeiro alvo, nesta ordem decrescente de pontuação, que satisfaça $\bar{d_A} > d_{self}$. O gene é considerado:

- **típico** (de provável origem vertical) caso $\bar{d_A} \leq d_{distant}$ (exemplo na Fig. 1.3 A),
- **atípico** (candidato a TGH) caso $\bar{d_A} > d_{distant}$ (exemplos na Fig. 1.3, B e C),
- **indeterminado** caso não exista este alvo A, com $\bar{d_A} > d_{self}$ (exemplo na Fig. 1.3 D).

O parâmetro d_{self} tem por função excluir genomas muito próximos ao genoma em estudo. Numa análise em que se quer considerar, por exemplo, transferências horizontais anteriores à divergência de duas cepas de uma mesma espécie, d_{self} deve ser superior à distância entre estas duas cepas. Caso contrário, um alvo BER de uma cepa na outra já faria com que o gene fosse classificado como típico. O parâmetro $d_{distant}$ é escolhido pelo usuário como a distância mínima de interesse para se avaliar transferência horizontal. Com estes parâmetros, pode-se controlar o foco da análise em transferências mais recentes ou mais antigas (parâmetro d_{self}); e entre

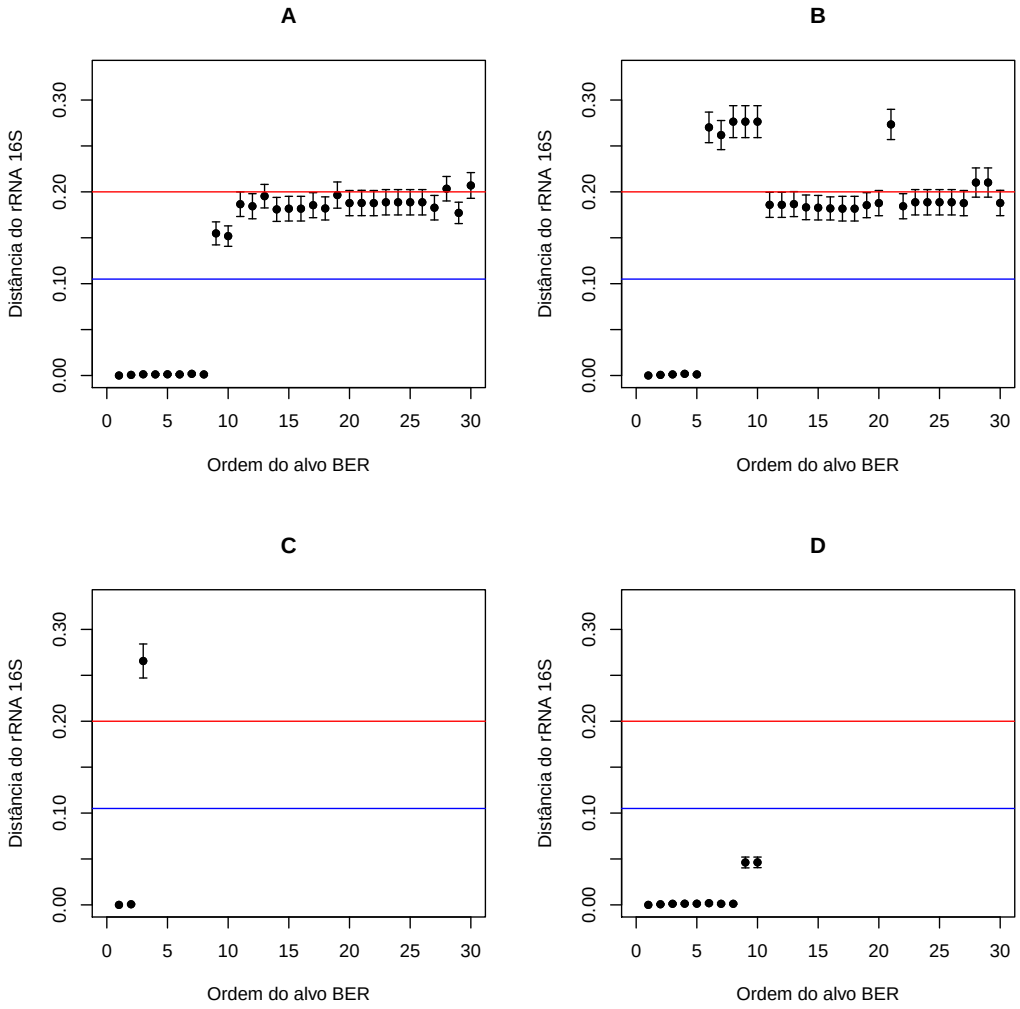

Figura 1.3 - Exemplos do funcionamento do método de detecção de TGH. Cada um dos gráficos trata da análise de TGH para um gene específico. No eixo X está representado, o número de ordem de cada alvo BER do gene em estudo e, no eixo Y, a distância de rRNA 16S entre o organismo e o organismo correspondente ao alvo BER. As linhas azuis e vermelhas correspondem, respectivamente, aos parâmetros d_{self} e $d_{distant}$. A classificação destes quatro genes quanto a transferência horizontal é, para (A), provável origem vertical, para (B) potencial envolvimento em TGH (classe 1), para (C) potencial envolvimento em TGH (classe 2), e para (D), indeterminada. As barras de erro correspondem aos desvios-padrões das réplicas de *bootstrap* das distâncias de rRNA 16S.

organismos mais próximos ou mais distantes (parâmetro $d_{distant}$). No entanto, para que seja preservada a qualidade das predições de TGH, a escolha destes parâmetros deve levar em conta a representatividade dos genomas no banco de dados.

Os genes atípicos são separados em duas classes, com base nos alvos BER subsequentes ao primeiro alvo "não-self" A. O gene é atribuído à:

- **classe 1:** caso exista algum alvo B subsequente a A (na ordem decrescente de pontuação) com d_B significativamente inferior a d_A (exemplo na Fig. 1.3B);
- **classe 2:** caso contrário (exemplo na Fig. 1.3C).

Um gene atribuído à classe 1 está potencialmente envolvido na substituição de um gene por um homólogo adquirido horizontalmente, enquanto que um gene atribuído à classe 2 pode estar envolvido na aquisição de uma função biológica nova para o genoma receptor.

A condição para se atribuir um gene à classe 1 é verificada caso a distância até B seja menor que a distância até A em no mínimo 99% das réplicas de *bootstrap* das medidas de distância. A fim de diminuir o custo computacional desta verificação, usamos uma abordagem indireta, que parte do pressuposto de que as réplicas de *bootstrap* das distâncias sejam normalmente distribuídas, condição que foi posteriormente verificada com o teste de normalidade de Shapiro-Wilk (Shapiro e Wilk, 1965) para os alvos com $\bar{d_X} > 0.003$. Com isso, a distância entre o organismo em estudo e o do alvo X pode ser expressa pela variável aleatória d_x, com distribuição:

$$d_X \sim N(\bar{d_X}, s_X^2)$$

onde s_X^2 é a variância amostral das réplicas de *bootstrap*. Na abordagem indireta, a condição (ii) é verificada se $P(d_B < d_A) \geq 0.99$. A probabilidade $P(d_B < d_A)$ é estimada com base no fato de que:

$$d_A - d_B \sim N(\bar{d_A} - \bar{d_B}, s_A^2 + s_B^2)$$

segundo a propriedade da distribuição da diferença entre duas variáveis aleatórias independentes normalmente distribuídas.

É importante ressaltar que, por este método não examinar reconstruções filogenéticas, suas predições de TGH não são definitivas: fatores como taxas de evolução díspares, perdas gênicas em linhagens intermediárias ou convergência evolutiva podem levar a uma falsa detecção de TGH. É importante também mencionar que o método identifica um gene como candidato a TGH, mas não identifica quais genomas estão no papel de doador e receptor. Para responder a esta pergunta e para se ter mais confiança na predição de TGH, recomenda-se fazer uma reconstrução filogenética da família gênica de interesse.

1.3.4 Identificação de genes com conteúdo G+C atípico

Um gene é considerado atípico quanto ao conteúdo G+C se sua contagem de G e C tem um desvio significativo (calculado por meio de um teste binomial bicaudal) da proporção destes nucleotídeos no genoma. A fim de controlar a taxa de falsos positivos em múltiplos testes de hipóteses, aplicamos o método *qvalue* (Storey e Tibshirani, 2003) aos valores p obtidos do teste binomial. Diferentemente do valor p, que expressa a probabilidade de um resultado individual ser um falso positivo, o valor q expressa a proporção esperada de falsos positivos entre os resultados considerados significativos. Genes com valor q inferior a 0.01 são classificados como atípicos e os demais como típicos.

1.3.5 Análise de enriquecimento de categorias funcionais entre os genes potencialmente envolvidos em TGH

No banco de dados Omniome, os genes estão anotados em 21 categorias funcionais, divididas em 121 subcategorias. Como algumas subcategorias de interesse em estudos de transferência horizontal pertenciam a categorias com uma ampla gama de funções, decidimos separá-las em categorias à parte. Além disso, para melhor apresentação, decidimos unificar as categorias relacionadas a genes não classificados.

As modificações feitas no esquema de categorias funcionais foram:

- Criação da categoria “Patogenicidade, produção de toxinas e resistência”, com as subcategorias “Patogenicidade e “Produção de toxinas e resistência”, que estavam originalmente na categoria “Processos celulares”.
- Retirada da subcategoria “Restrição/modificação de DNA” da a categoria “Metabolismo de DNA” e colocação em uma categoria à parte.
- Unificação das categorias “Não classificados”, “Rejeitos do Glimmer”, “Quadro de leitura rompido” e “Função desconhecida” na categoria “Não classificados”.

Para cada categoria X e para cada genoma, monta-se uma tabela 2x2 com as contagens de: (i) genes atípicos anotados na categoria X, (ii) genes típicos anotados na categoria X, (iii) genes atípicos não anotados na categoria X, e (iv) genes típicos não anotados na categoria X. Os genes com atipicidade indeterminada não são computados (ver tabela 1.1).

Tabela 1.1. Tabela de contingência para análise de enriquecimento funcional. O fator de enriquecimento de genes atípicos na categoria X é dado por $\frac{AD}{BC}$.

	genes atípicos	genes típicos
genes na categoria X	A	B
genes fora da categoria X	C	D

Com esta tabela, calcula-se o fator de enriquecimento (ou depleção) de genes atípicos nesta categoria, e a significância estatística (valor p) deste enriquecimento (ou depleção). O fator de enriquecimento, expresso pela razão de chances da

tabela de contingência, corresponde à taxa de variação da razão entre o número de genes anotados na categoria X e os não anotados nesta categoria, quando passamos do conjunto dos genes típicos ao dos atípicos (ver tabela 1.1). O valor p é calculado por meio do teste χ^2, usando simulações de Monte Carlo (função `chisq.test` do ambiente estatístico R (Hope, 1968)).

Nas análises que envolvem múltiplos genomas e múltiplas categorias funcionais, aplicamos o método *qvalue* (Storey e Tibshirani, 2003) para estimar a taxa de falsos positivos em múltiplos testes de hipóteses. Este método computa os valores q, que correspondem às taxas de descobertas falsas, a partir dos valores p obtidos para cada categoria funcional e genoma.

1.3.6 Teste da Hipótese da Complexidade

Para testar se a hipótese da complexidade é suportada por nossas predições, usamos dados de três estudos independentes de interação proteína-proteína em escala genômica, dois em *Escherichia coli* (Butland *et al.*, 2005; Arifuzzaman *et al.*, 2006) e um em *Helicobacter pylori* (Rain *et al.*, 2001). Para cada gene analisado nestes estudos, anotamos o número de parceiros de interação proteína-proteína e sua classificação quanto a transferência horizontal, ou seja, se o gene é típico, atípico ou indeterminado. A seguir, verificamos, para cada estudo, se o número parceiros de interação proteica é estocasticamente menor entre os genes atípicos que entre os genes típicos, com base no teste de Mann-Whitney (Mann e Whitney, 1947). A fim de minimizar a influência de fatores experimentais de cada um destes estudos, analisamos em separado os genes nos papéis de presa e de isca.

1.3.7 Implementação

Os programas que implementam os métodos aqui desenvolvidos foram escritos Perl e fazem uso de um banco de dados MySQL e de rotinas do ambiente estatístico R.

1.4 Resultados e Discussão

1.4.1 Identificação de candidatos a TGH

Aplicamos o método de identificação de TGH descrito na seção 1.3.3, ao conjunto de todas as ORFs preditas de 408 dos 452 genomas procarióticos do banco de dados Omniome para os quais pudemos identificar anotações de rRNA 16S. Usamos, nesta análise, os parâmetros $d_{self} = 0.105$ e $d_{distant} = 0.2$ substituições nucleotídicas por sítio. O parâmetro d_{self} foi escolhido de modo a que fiquem situadas dentro deste limite de distância a maior parte das espécies de um mesmo gênero. Já o parâmetro $d_{distant}$ corresponde, aproximadamente, à distância média entre representantes de proteobactérias do grupo α e do grupo γ. Os números totais de genes classificados como atípicos (candidatos a TGH), típicos (de provável descendência vertical) e indeterminados, são mostrados na Tabela 1.2. Como será visto na seção 1.4.2, este método de detecção de TGH tem melhor desempenho quando o genoma em estudo possui um número mínimo de "vizinhos" filogenéticos representados no banco de dados. Em razão disto, os mesmos totais são também mostrados para os genomas com ≥ 20 vizinhos. A lista completa dos candidatos a TGH pode ser obtida no endereço: `http://www.icb.usp.br/~mutagene/apua/hgt_candidates.zip`.

Aplicamos também o método descrito na seção 1.3.4 para encontrar genes com conteúdo G+C significativamente distinto da média do genoma. Os totais encontrados são mostrados na Tabela 1.3. A fim de verificar se a classificação de um gene como típico ou atípico pelo método de BLAST é ou não independente de sua classificação por conteúdo G+C, compilamos, na Tabela 1.4, as contagens totais dos genes classificados por ambos os métodos, para os genomas com ≥ 20 vizinhos. Aplicando-se o teste χ^2 a estes dados, verifica-se que o conjunto dos genes classificados como atípicos segundo o método baseado em BLAST está significativamente enriquecido em genes com conteúdo G+C atípico ($p < 10^{-6}$), com fator de enriquecimento, expresso pela razão de chances da tabela de contingência, de 1,31. No entanto, as predições feitas pelos dois métodos concordam para apenas 67% dos ge-

Tabela 1.2. Números totais de genes classificados quanto ao potencial envolvimento em TGH, segundo o método baseado em BLAST, para os 408 genomas estudados e para os 256 genomas com ≥ 20 vizinhos filogenéticos (ver seção 1.4.2).

	Todos os genomas		Com ≥ 20 vizinhos	
Classificação	Contagem	Fração	Contagem	Fração
Atípico	257087	0.20	151867	0.17
Típico	527570	0.41	424457	0.47
Indeterminado	504011	0.39	325999	0.36
Total	1288668	1.00	902323	1.0

Tabela 1.3. Números totais de genes classificados quanto ao conteúdo G+C, para os 408 genomas estudados e para aqueles com ≥ 20 vizinhos filogenéticos

	Todos os genomas		Com ≥ 20 vizinhos	
Classificação	Contagem	Fração	Contagem	Fração
Atípico	249038	0.19	164745	0.18
Típico	1039630	0.81	737578	0.82
Total	1288668	1.0	902323	1.0

nes. Tal divergência pode ser devida ao fato do método de conteúdo G+C identificar potenciais transferências recentes entre organismos com conteúdo G+C distinto. Já as transferências antigas ou entre organismos com composição nucleotídica parecida podem ser detectadas apenas pelo método baseado em BLAST.

Tabela 1.4. Números totais de genes classificados quanto ao potencial envolvimento em TGH segundo o métodos baseados em BLAST e conteúdo G+C. Os genes classificados com indeterminados pelo método de BLAST não são considerados nesta análise. O conjunto de genes com BLAST atípico está significativamente enriquecido em genes com conteúdo G+C atípico ($p < 10^{-6}$ e razão de chances de 1,31, usando-se o teste χ^2 com 10^6 simulações de Monte Carlo). A proporção de genes com classificação igual pelos dois métodos (a diagonal principal desta tabela) é de 67%.

	BLAST atípico	BLAST típico
G+C atípico	32446	72844
G+C típico	119421	351613

1.4.2 A influência representatividade dos genomas

Os métodos de detecção de TGH por meio de BLAST são particularmente sensíveis à representatividade dos genomas no banco de dados de busca. Um organismo com poucos "vizinhos" filogenéticos pode ter uma proporção considerável de genes incorretamente preditos como envolvidos em TGH por não terem alvos de BLAST nos poucos organismos próximos representados no banco. A fim de analisar este efeito sobre nosso método, aplicamos a seguinte estratégia: (i) para cada genoma, definimos o número de vizinhos como sendo o número de genomas que têm, com este, distância entre d_{self} e $d_{distant}$; (ii) plotamos, na Fig. 1.4, a proporção de

genes atípicos em função do número de vizinhos. Na região de 0 a 20 vizinhos, observam-se uma alta variabilidade e uma diminuição na proporção de genes atípicos, conforme aumenta o número de vizinhos. A partir de 20 vizinhos, embora a variabilidade diminua com o aumento do número de vizinhos, não há tendência de queda na proporção de genes atípicos, que se estabiliza em aproximadamente 14%, de acordo com a curva de regressão local.

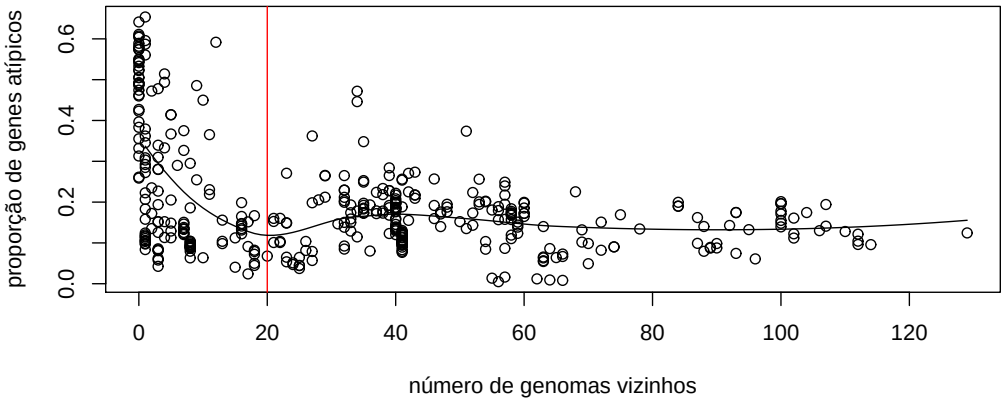

Figura 1.4 - Proporção de genes atípicos (candidatos a TGH) de cada genoma em função do número de genomas "vizinhos", i.e., com distância de rRNA 16S entre d_{self} e $d_{distant}$. A linha sólida corresponde a uma regressão local obtida com o método LOESS.

A grande variabilidade e as altas proporções de candidatos a TGH detectados para genomas com < 20 vizinhos, sugerem que a representatividade de genomas no banco de dados seja de fato importante no desempenho do método de detecção de TGH, de modo que genomas com poucos vizinhos podem ter proporções de TGH superestimadas. No entanto, a estabilização da proporção média de candidatos a TGH a partir de 20 vizinhos em um valor diferente de zero sugere que o método

detecta uma fração basal dos genes como envolvida em TGH, e que esta fração não diminui com o aumento do número de genomas vizinhos.

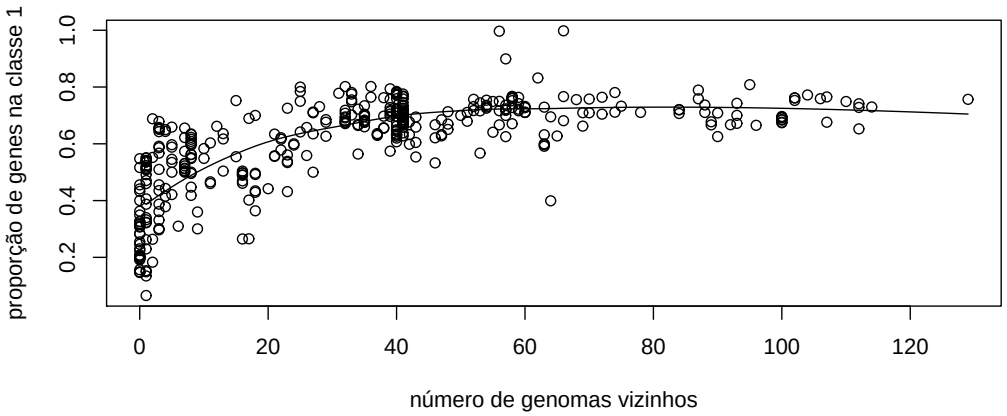

Figura 1.5 - Proporção de genes atípicos (candidatos a TGH) atribuídos à classe 1 (genes que possuem alvos melhores em organismos distantes que em organismos próximos) em função do número de vizinhos. A linha sólida corresponde a uma regressão local obtida com o método LOESS.

Os candidatos a TGH são atribuídos à classe 1 caso tenham alvos BER subsequentes ao primeiro alvo "não-self" em organismos com distância significativamente inferior à do primeiro alvo "não-self"; e na classe 2 caso contrário (ver seção 1.3.3 para mais detalhes). Como visto, esta classificação pode ter a seguinte interpretação biológica:

- classe 1: potencial substituição de um gene por um homólogo adquirido horizontalmente.
- classe 2: potencial aquisição de um gene novo.

Na Fig. 1.5, é plotada a proporção de genes na classe 1, dentre o total

dos genes atípicos, em função do número de genomas vizinhos. Observa-se um crescimento nesta proporção, com estabilização em torno de 72%, conforme aumenta o número de vizinhos. A estabilização em um valor diferente de 1 sugere que uma fração basal (28%) dos candidatos a TGH pertença efetivamente à classe 2, ou seja, a classe dos genes envolvidos em potenciais transferências de novas funções biológicas ao organismo receptor.

Nas análises subsequentes apresentadas neste trabalho, a não ser quando explicitamente mencionado em contrário, o conjunto dos candidatos a TGH se refere àqueles detectados em genomas com ≥ 20 vizinhos.

1.4.3 Análise de enriquecimento funcional entre os candidatos a TGH

A partir das predições de TGH feitas com o método baseado em BLAST descrito na seção 1.3.3, analisamos quais categorias funcionais estão enriquecidas ou empobrecidas em candidatos a TGH. Os resultados da análise global considerando todos os 256 genomas com ≥ 20 vizinhos são mostrados na Tabela 1.5. As categorias de restrição/modificação de DNA, transporte e elementos genéticos móveis foram as que tiveram os maiores fatores de enriquecimento, seguidas das categorias de patogenicidade e metabolismo central intermediário. Em contraste, as categorias de síntese de proteínas, transcrição, metabolismo de DNA e síntese de purinas, pirimidinas e nucleotídeos tiveram fatores de enriquecimento consideravelmente menores que 1, o que significa que estas categorias estão empobrecidas em candidatos a TGH.

A mesma análise, porém feita para a detecção de TGH por meio de conteúdo G+C atípico, é mostrada na Tabela 1.6. Assim como na análise anterior, a categoria de restrição/modificação de DNA foi a que teve o maior fator de enriquecimento, com valor comparável ao daquela análise. Para as outras categorias, porém, os fatores de enriquecimento variaram pouco (0.88 a 1.35) em comparação com os da análise anterior (0.29 a 1.72). Para algumas categorias, os fatores de enriquecimento vão em sentidos opostos nas duas análises. Por exemplo: metabolismo de DNA (0.44 e 1.35) e transporte (1.72 e 0.93). Para outras, os sentidos concordam: síntese de

Tabela 1.5. Análise de enriquecimento funcional em candidatos a TGH detectados por BLAST nos 256 genomas com ≥ 20 vizinhos. Genes: número de genes na categoria; %Atípicos: porcentagem de genes atípicos na categoria; FE: fator de enriquecimento (ver seção 1.3.5, pág. 19). $*$: FE significativamente diferente de 1 ($p < 0.05$ no teste χ^2). Os genes indeterminados não são considerados nesta análise.

Categoria funcional	Genes	%Atípicos	FE	$\log_2 FE$	
Síntese de proteínas	34468	9.8	0.29	-1.79	$*$
Processamento de proteínas	29099	18.5	0.62	-0.69	$*$
Metabolismo de DNA	23558	13.9	0.44	-1.19	$*$
Metabolismo de nucleotídeos	15081	17.3	0.58	-0.79	$*$
Biossíntese de cofatores, grupos prostéticos e carreadores	25514	18.2	0.61	-0.71	$*$
Processos celulares	24267	22.6	0.81	-0.30	$*$
Transcrição	9495	11.9	0.37	-1.42	$*$
Biossíntese de aminoácidos	21722	19.1	0.65	-0.62	$*$
Metabolismo de fosfolipídeos e ácidos graxos	16223	28.6	1.12	0.16	$*$
Funções regulatórias	46146	28.4	1.12	0.16	$*$
Transdução de sinais	6951	27.4	1.05	0.08	
Patogenicidade, produção de toxinas e resistência	13110	34.4	1.48	0.56	$*$
Metabolismo energético	74723	28.9	1.16	0.21	$*$
Proteínas hipotéticas conservadas	39440	27.7	1.07	0.10	$*$
Envelope celular	40501	26.3	1.00	-0.01	
Não classificados	96292	29.8	1.23	0.30	$*$
Metabolismo central intermediário	27556	35.0	1.54	0.62	$*$
Proteínas hipotéticas não conservadas	9177	33.2	1.40	0.48	$*$
Funções relacionadas a elementos genéticos móveis	7933	37.8	1.71	0.78	$*$
Proteínas de transporte e ligação	76840	36.2	1.72	0.78	$*$
Restrição/modificação de DNA	1691	43.2	2.13	1.09	$*$
Total	576324	26.4			

Tabela 1.6. Análise de enriquecimento funcional para genes com conteúdo G+C atípico nos 256 genomas com ≥ 20 vizinhos. Genes: número de genes na categoria; %Atípicos: porcentagem de genes atípicos na categoria; FE: fator de enriquecimento (ver seção 1.3.5, pág. 19). $*$: FE significativamente diferente de 1 ($p < 0.05$ no teste χ^2). Os genes indeterminados não são considerados nesta análise.

Categoria funcional	Genes	%Atípicos	FE	$\log_2 FE$	
Síntese de proteínas	41506	16.8	0.90	-0.15	$*$
Processamento de proteínas	38419	18.2	1.00	-0.00	
Metabolismo de DNA	30193	22.9	1.35	0.43	$*$
Metabolismo de nucleotídeos	16081	17.0	0.92	-0.12	$*$
Biossíntese de cofatores, grupos prostéticos e carreadores	28024	19.3	1.07	0.10	$*$
Processos celulares	33288	16.8	0.90	-0.15	$*$
Transcrição	11575	16.9	0.91	-0.14	$*$
Biossíntese de aminoácidos	23487	18.9	1.04	0.06	$*$
Metabolismo de fosfolipídeos e ácidos graxos	19185	16.4	0.88	-0.19	$*$
Funções regulatórias	64702	16.7	0.89	-0.17	$*$
Transdução de sinais	8818	18.1	0.99	-0.01	
Patogenicidade, produção de toxinas e resistência	20758	21.6	1.24	0.31	$*$
Metabolismo energético	88259	20.4	1.17	0.22	$*$
Proteínas hipotéticas conservadas	120152	16.0	0.84	-0.26	$*$
Envelope celular	65338	20.1	1.14	0.19	$*$
Não classificados	144115	17.1	0.91	-0.14	$*$
Metabolismo central intermediário	31848	17.9	0.97	-0.04	
Proteínas hipotéticas não conservadas	86106	22.0	1.30	0.37	$*$
Funções relacionadas a elementos genéticos móveis	27221	20.9	1.19	0.25	$*$
Proteínas de transporte e ligação	90573	17.3	0.93	-0.10	$*$
Restrição/modificação de DNA	2191	35.6	2.48	1.31	$*$
Total	902323	18.3			

proteína (0.27 e 0.90) e patogenicidade (1.48 e 1.24).

Os padrões globais de enriquecimento funcional entre os candidatos a TGH foram também observados em genomas individuais. Os resultados, para o método baseado em BLAST, são mostrados na Fig. 1.6 (para os 256 genomas com $\geq$ 20 vizinhos) e na Fig. 1.7 (para os 374 genomas com pelo menos 1 vizinho). Para a análise de conteúdo G+C, as respectivas figuras são a 1.8 e a 1.9. Nestas figuras, cada coluna corresponde a uma categoria funcional e cada linha a um genoma. Uma célula é marcada em vermelho se a categoria funcional correspondente está significativamente enriquecida em candidatos a TGH (valor $q < 0.01$ e fator de enriquecimento ≥ 1.5) em relação á média do genoma; uma célula é marcada em azul se a categoria correspondente está empobrecida em candidatos a TGH (valor $q < 0.01$ e fator de enriquecimento $\leq 1.5^{-1}$). As células marcadas em branco correspondem a enriquecimentos/empobrecimentos não significativos e/ou de magnitude ≤ 1.5. Os genomas são postos em ordem de acordo com sua classificação taxonômica, levando-se em conta todos os níveis taxonômicos presentes no banco de dados. Assim, todos membros de um mesmo grupo taxonômico aparecem como um bloco contíguo em cada uma destas figuras.

Como pode ser observado na Fig. 1.6, os genomas têm, em geral, padrão de enriquecimento de candidatos a TGH similar ao observado na análise global (Tabela 1.5). Em muitos casos, porém, não se atinge o limiar estabelecido de significância estatística. Um exemplo notável é o da categoria de restrição/modificação de DNA, que apresenta o maior fator de enriquecimento na análise global mas que, devido ao pequeno número de genes classificados nesta categoria, não apresenta enriquecimento significativo nas análises de genomas individuais. É interessante notar que, para uma mesma coluna, a grande maioria das células não-brancas têm a mesma cor, o que mostra que uma mesma categoria funcional tem, em geral, padrão de enriquecimento em candidatos a TGH consistente entre os genomas.

Na Fig. 1.7, em que não são retirados os genomas com $<$ 20 vizinhos, nota-se padrão de enriquecimento similar ao da Fig. 1.6, mesmo entre genomas com

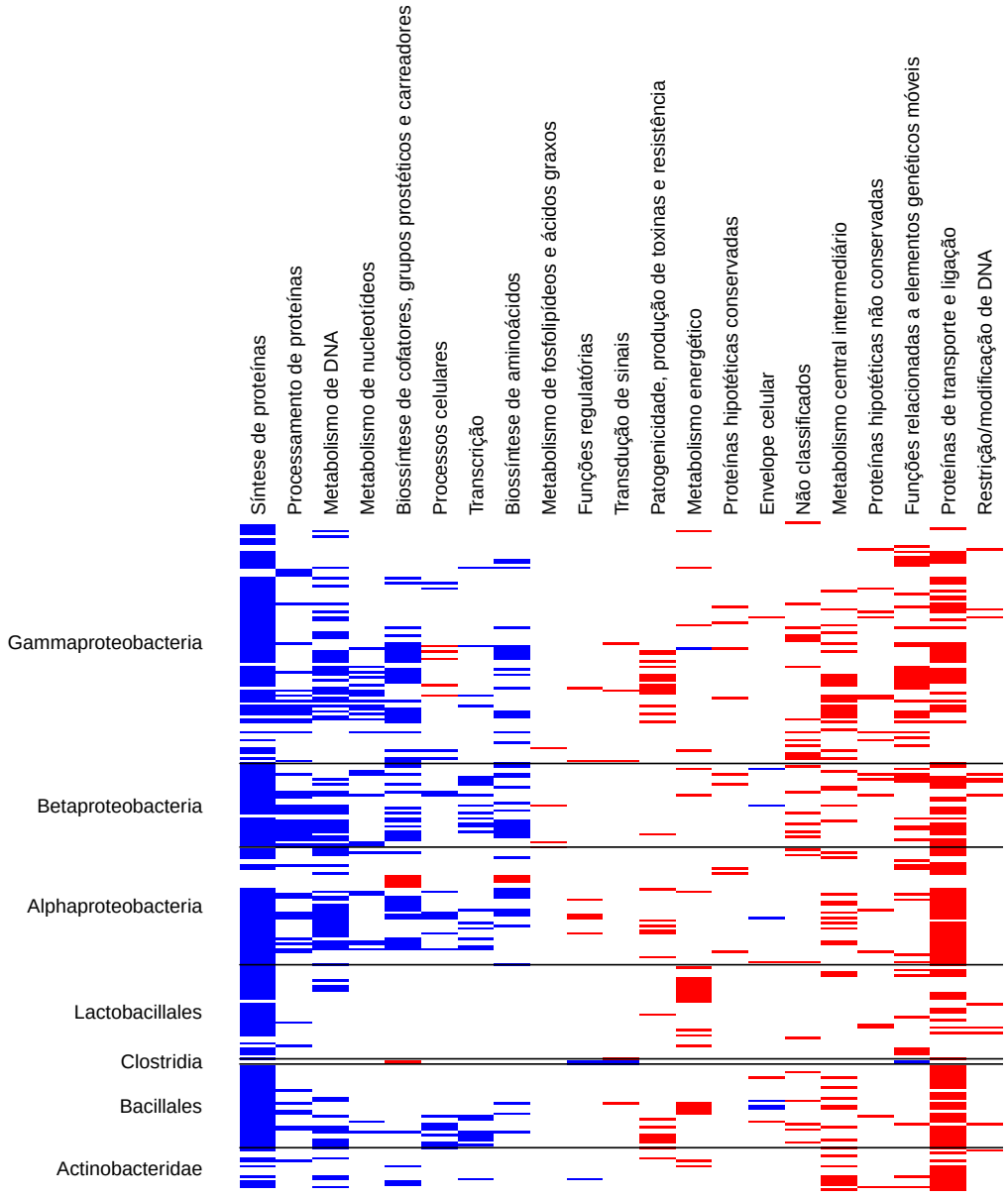

Figura 1.6 - Mapa de enriquecimento funcional para candidatos a TGH detectados por BLAST para os 256 genomas com ≥ 20 vizinhos. Cada linha representa um genoma e cada coluna um categoria funcional. Uma célula é marcada em vermelho(azul) se a proporção de candidatos a TGH naquela categoria é significativamente maior(menor) que a média do genoma (Teste χ^2 com 10^6 simulações de Monte Carlo, com valor $q < 0.01$ e fator de enriquecimento/empobrecimento ≥ 1.5).

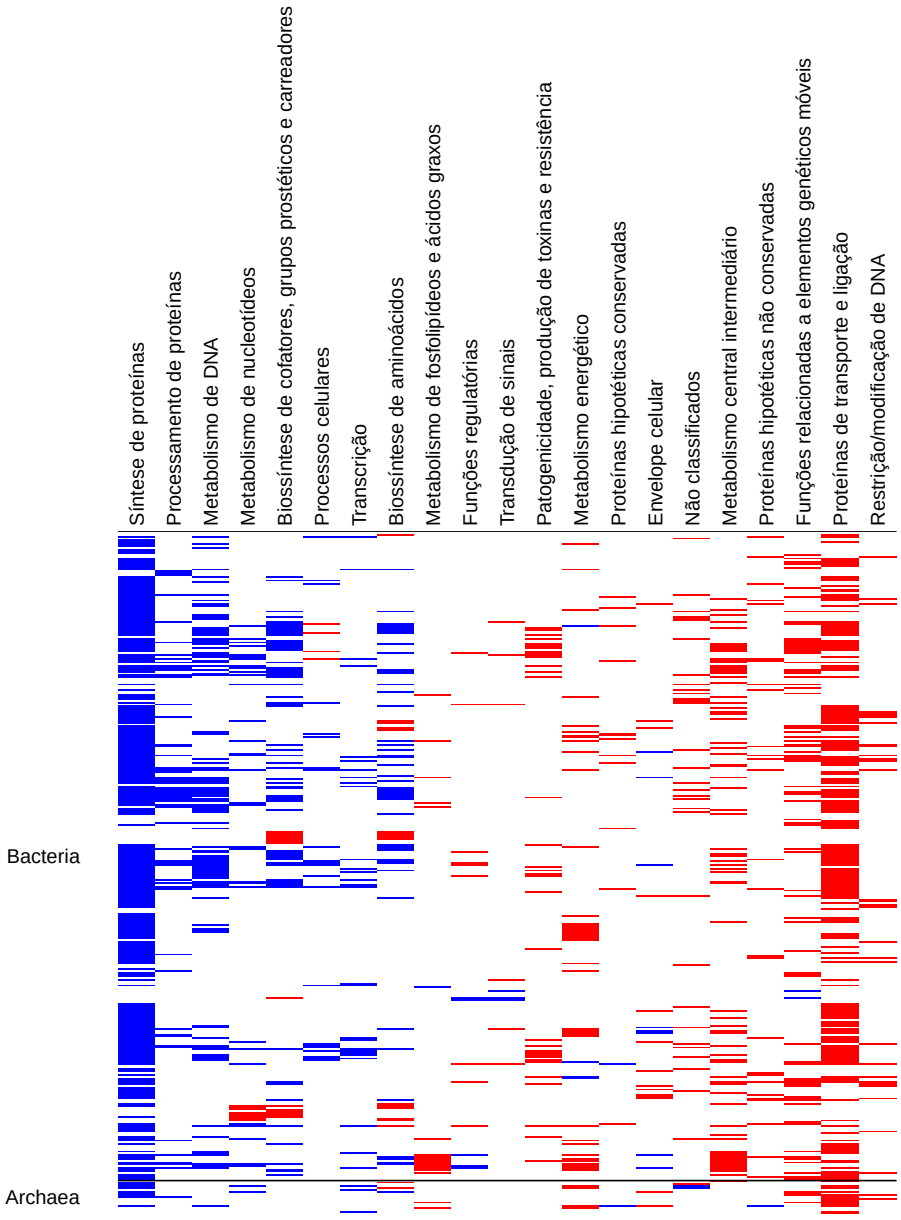

Figura 1.7 - Mapa de enriquecimento funcional para candidatos a TGH detectados por BLAST para os 374 genomas com pelo menos 1 vizinho. Cada linha representa um genoma e cada coluna um categoria funcional. Uma célula é marcada em vermelho(azul) se a proporção de candidatos a TGH naquela categoria é significativamente maior(menor) que a média do genoma (Teste χ^2 com 10^6 simulações de Monte Carlo, com valor $q < 0.01$ e fator de enriquecimento/empobrecimento ≥ 1.5).

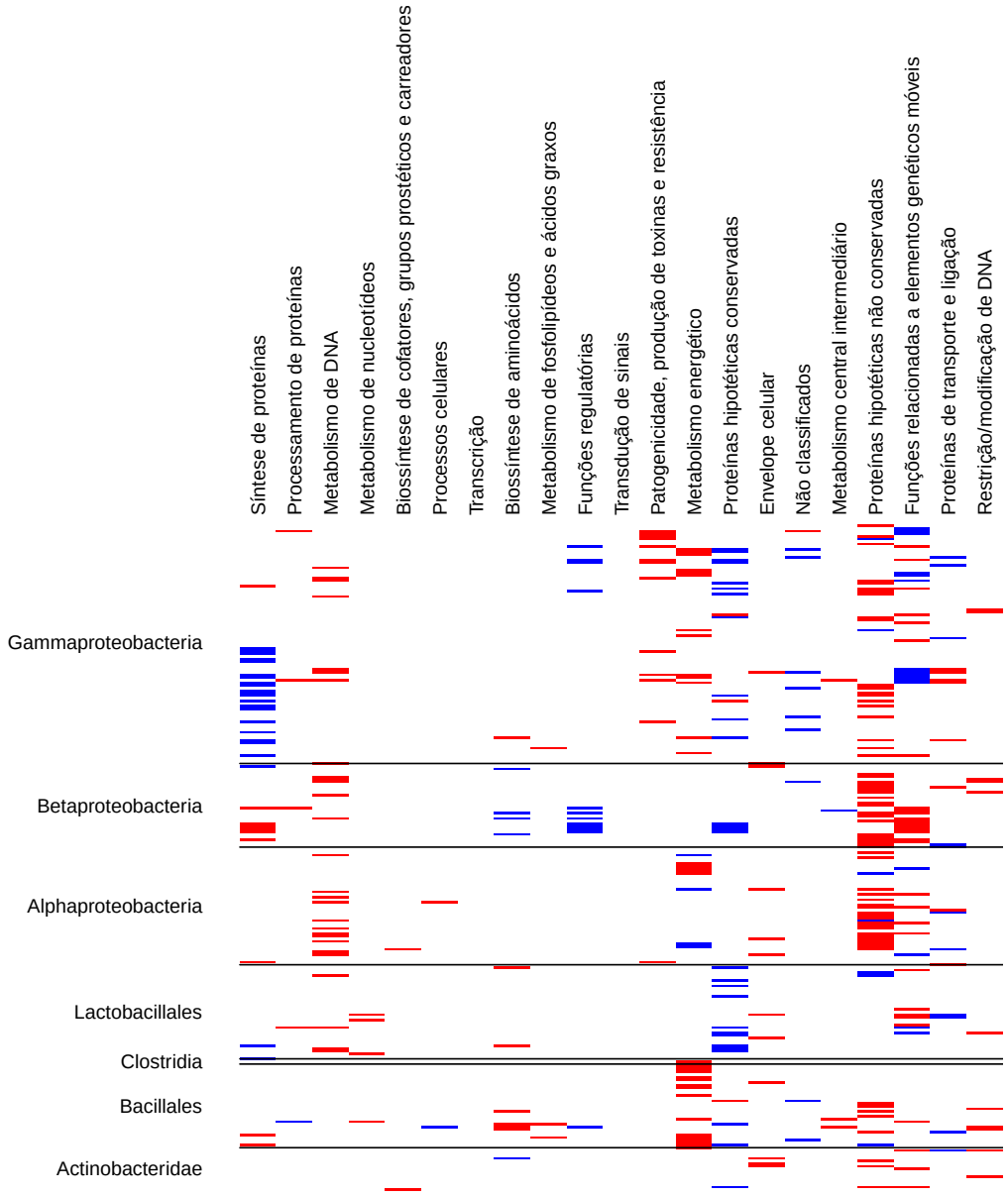

Figura 1.8 - Mapa de enriquecimento funcional para genes com conteúdo G+C atípico para os 256 genomas com ≥ 20 vizinhos. Cada linha representa um genoma e cada coluna um categoria funcional. Uma célula é marcada em vermelho(azul) se a proporção de candidatos a TGH naquela categoria é significativamente maior(menor) que a média do genoma (Teste χ^2 com 10^6 simulações de Monte Carlo, com valor $q < 0.01$ e fator de enriquecimento/empobrecimento ≥ 1.5).

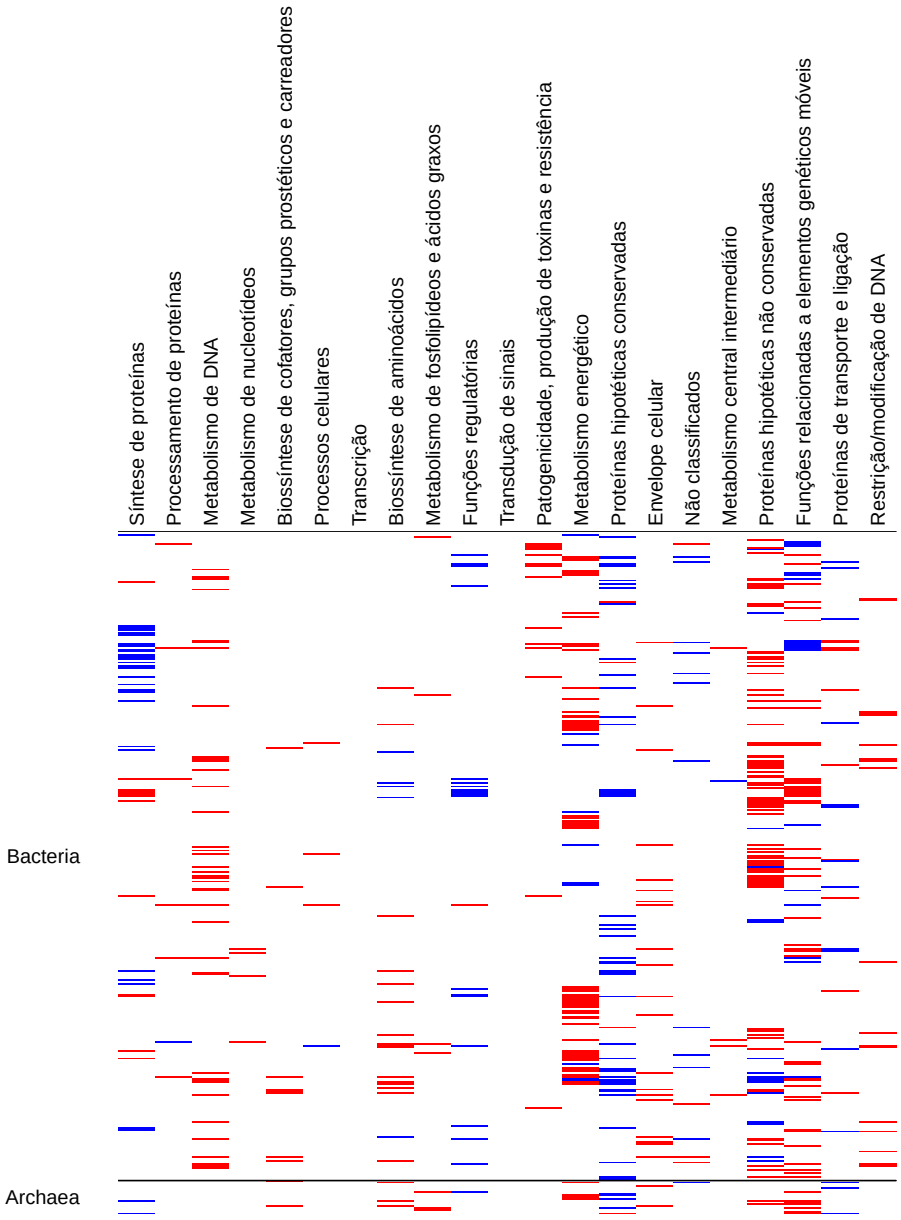

Figura 1.9 - Mapa de enriquecimento funcional para genes com conteúdo G+C atípico para os 374 genomas com pelo menos 1 vizinho. Cada linha representa um genoma e cada coluna um categoria funcional. Uma célula é marcada em vermelho(azul) se a proporção de candidatos a TGH naquela categoria é significativamente maior(menor) que a média do genoma (Teste χ^2 com 10^6 simulações de Monte Carlo, com valor $q < 0.01$ e fator de enriquecimento/empobrecimento ≥ 1.5).

$<$ 20 vizinhos, como é o caso das arqueas. No entanto, as exceções à observação mencionada no parágrafo anterior (ou seja, células azuis e vermelhas numa mesma coluna) são mais frequentes, provavelmente devido à menor qualidade na predição de TGH.

Nas análises de conteúdo G+C atípico (Figs. 1.8 e 1.9), as categorias funcionais têm, em geral, padrões de enriquecimento pouco consistentes entre os genomas, se comparados aos da análise baseada em BLAST (Fig. 1.6).

Os sistemas de restrição/modificação, enriquecidos em candidatos a TGH segundo os dois métodos usados neste trabalho, são conhecidos por participarem de transferência horizontal e por estarem associados a fagos, plasmídeos e elementos genéticos móveis. De maneira análoga aos sistemas toxina/antitoxina, estes sistemas comportam-se como elementos genéticos egoístas, aumentando sua frequência na população por meio da eliminação das células que venham a perdê-los (Kobayashi, 2001).

Na análise baseada em BLAST (Figs. 1.6 e 1.7) destacam-se as categorias de genes relacionadas a transporte (como altamente enriquecida em genes envolvidos em TGH) e de genes relacionados a síntese proteica (como principalmente provenientes de descendência vertical). O fato da categoria de transporte estar consideravelmente enriquecida em candidatos a TGH (segundo o método baseado em BLAST) é coerente com a ideia de que proteínas situadas na periferia da rede metabólica são mais propensas a TGH. No estudo de Pál *et al.* (2005), que analisa a proporção de candidatos a TGH em função da posição de cada proteína na rede metabólica de *E. coli*, e cujo método de detecção de TGH foi mencionado na seção 1.1.4, foram encontradas proporções decrescentes de candidatos a TGH à medida em que se consideraram proteínas classificadas em: transporte, primeira reação, reações intermediárias e produção de biomassa. Com relação à categoria de síntese proteica, há provavelmente uma pressão seletiva muito grande para que genes envolvidos nesses processos não possam ser facilmente substituídos (por ortólogos provenientes de TGH, por exemplo), dada a função essencial que estas proteínas desempenham na

célula. Além disso, boa parte dessas proteínas atuam de modo extremamente coordenado, requisitando um alto número de interações em complexos, o que também, provavelmente, dificultam a fixação de eventos de TGH em genes desta categoria. Curiosamente, o trabalho recente de Kanhere e Vingron (2009) propõe que esta categoria funcional está enriquecida em TGH, quando se consideram transferências entre espécies bacterianas filogeneticamente distantes. Este trabalho será discutido em mais detalhe na seção 1.4.4.

Os genes anotados como hipotéticos não conservados, aqueles cuja função é desconhecida e que não apresentam conservação em um número significativo de espécies, também se mostraram enriquecidos em candidatos a TGH segundo ambos os métodos (Tabelas 1.5 e 1.6). Uma observação semelhante havia sido feita no estudo com 63 genomas de Hsiao *et al.* (2005) que, usando detecção de TGH por análise de uso de códons, identificaram uma proporção significativamente alta de genes hipotéticos não conservados em regiões preditas como ilhas de transferência horizontal. Estas observações, além do fato dos genes hipotéticos conservados não apresentarem estas mesmas proporções de TGH, estão de acordo com a ideia de que o conjunto "flexível" de genes (seção 1.1.6) é mais propenso à transferência horizontal.

1.4.4 Crítica ao trabalho de Kanhere e Vingron (2009)

Kanhere e Vingron (2009) propuseram um método de identificação de TGH que se baseia na análise de correlação entre distâncias entre sequências de proteínas e distâncias entre os rRNA 16S das espécies correspondentes. Supondo-se que as famílias gênicas têm descendência principalmente vertical, com alguns poucos eventos de TGH, é de se esperar uma correlação positiva entre estas duas distâncias: quanto maior a separação filogenética entre as espécies (inferida pela dissimilaridade na sequência do rRNA 16S) maior se espera que seja a dissimilaridade entre as sequências de proteína. Os pares de proteínas que figuram como *outliers* nesta análise de correlação são selecionados para a próxima fase da análise, na qual as proteínas devem ainda passar por dois critérios para serem classificadas como potencialmente

envolvidas em TGH: ter alta similaridade de sequência com uma proteína de um organismo filogeneticamente distante e árvore filogenética significativamente diferente da árvore de referência das espécies. Aplicando este método a 63 genomas procarióticos representados no banco COG (Tatusov *et al.*, 2003), os autores concluíram que genes transferidos entre bactérias e arqueas são, em sua maioria, relacionados a metabolismo, enquanto que genes transferidos entre bactérias são principalmente envolvidos em tradução (síntese proteica). Esta última afirmação é oposta aos resultados do presente trabalho, no qual observamos que a categoria de síntese de proteínas está significativamente empobrecida em candidatos a TGH. A fim de entender tal disparidade, analisamos cada passo do trabalho de Kanhere e Vingron (2009) quanto ao enriquecimento de genes na categoria J do COG (tradução, estrutura do ribossomo e biogênese) entre os candidatos a TGH. Estes passos são resumidos a seguir:

1. Pré-seleção de COGs para os quais o método é aplicável: para cada grupo de genes ortólogos (COG) faz-se uma análise de correlação entre as distâncias entre sequências de proteínas e as distâncias entre os rRNA 16S das respectivas espécies. São retidos para os passos posteriores somente os COGs que têm pelo menos 8 proteínas, coeficiente de correlação > 0.3 e valor $p < 10^{-5}$.

2. Detecção de *outliers*: pares de genes com distância de Cook (Cook, 1979) acima de certo limiar são considerados *outliers* nas análises de correlação e retidos para o próximo passo.

3. Seleção de genes que com alvos de alta similaridade em organismos distantes: são selecionados para o próximo passo os genes que têm 40% de identidade em sequência de aminoácidos com alvos em organismos com distância de rRNA 16S superior a 0.3 substituições nucleotídicas por sítio.

4. Reconstruções filogenéticas: são selecionados como candidatos a TGH os genes cujas árvores filogenéticas de proteína são significativamente diferentes da árvore de rRNA 16S, segundo testes estatísticos aplicados a estas árvores.

5. Seleção de transferências entre bactérias: são separados os genes potencialmente transferidos entre bactérias daqueles transferidos entre arqueas e bactérias.

A Tabela 1.7 mostra o número de genes selecionados após cada passo de análise, a fração, dentre estes, pertencente à categoria J do COG (tradução, estrutura do ribossomo e biogênese), a taxa de variação nesta fração desde o passo anterior e a significância estatística desta variação, calculada por meio de um teste χ^2 com simulações de Monte Carlo. Não há informação, no artigo, sobre o número de genes na categoria J após o passo 3. Portanto, no passo 4, a taxa de variação é calculada a partir do passo 2. Observa-se, nesta tabela, que houve uma duplicação na fração de genes pertencentes à categoria J após a pré-seleção de COGs para os quais o método é aplicável (passo 1). Após o passo 2 houve um aumento pequeno, porém significativo, na fração de genes na categoria J. Após o passo 4, no qual é definida a lista final contendo 171 candidatos a TGH, observa-se uma diminuição nesta fração, porém sem significância estatística. Selecionando-se os 53 genes potencialmente transferidos de uma bactéria para outra (passo 5), observa-se um incremento significativo, de fator 2.55, na fração de genes na categoria J.

No passo 1, 61% dos genes são descartados do conjunto de análise por não satisfazerem condições pré-estipuladas para a aplicabilidade do método (coeficiente de correlação > 0.3 e número de genes na família ≥ 8). A não satisfação destas condições não significa, entretanto, que estes genes não possam estar envolvidos em TGH. Curiosamente, o descarte destes genes levou a uma duplicação estatisticamente significativa da representatividade dos genes relacionados à tradução (Tabela 1.7), o que mostra que os critérios usados nesta pré-seleção introduziram um viés funcional no conjunto de análise. O passo 3, no qual se observa redução considerável no conjunto de genes em análise, consiste na eliminação dos genes que não têm 40% de identidade de aminoácidos com um alvo em um organismo filogeneticamente distante, com o objetivo de reter apenas aqueles genes envolvidos em transferência horizontal recente. Algumas categorias funcionais, como por exemplo a de tradução, apresen-

tam famílias gênicas com alto grau de conservação de sequência e taxa de evolução lenta ao longo do tempo evolutivo. Uma transferência horizontal antiga de um gene de uma destas famílias poderia resultar na observação de tal nível de identidade de aminoácidos, enquanto que, para uma família de genes com taxa de evolução mais rápida, níveis similares só seriam atingidos para transferências recentes. Deste modo, o passo 3 pode selecionar um número proporcionalmente maior de genes com taxa de evolução lenta, já que genes envolvidos tanto em transferências antigas quanto recentes poderiam satisfazer este critério. Após os passos 4 e 5, temos dois conjuntos de candidatos a TGH: aqueles potencialmente transferidos entre bactérias e arqueas e aqueles transferidos entre diferentes espécies bacterianas. Para o primeiro conjunto, os genes são, em sua maioria, relacionados a metabolismo, cabendo lembrar que, no banco de dados COG, as funções de transporte estão distribuídas em diferentes categorias de metabolismo. Para este conjunto, podemos concluir que houve concordância com os dados de enriquecimento funcional do presente trabalho. Para o segundo conjunto, a categoria mais representada entre os candidatos a TGH é a de tradução, porém as de metabolismo também têm um número considerável de representantes (Fig. 5 de Kanhere e Vingron (2009)). É possível que para transferências entre bactérias, o número de genes na categoria de tradução esteja superestimado em função da detecção de eventos antigos de transferência, enquanto que para as categorias de metabolismo só seriam detectadas as transferências recentes. O mesmo não ocorreria para transferências entre bactérias e arqueas, já que as grandes distâncias filogenéticas entre elas poderiam implicar na não satisfação do critério de 40% de identidade de aminoácidos para genes de transferência antiga.

Isto posto, a principal crítica que fazemos a este trabalho é que a aplicação de critérios restritivos de pré-filtragem e detecção de TGH pode ter descartado da análise uma porção considerável de genes envolvidos em transferência horizontal. Além disso, como mostramos para o passo 1, a aplicação destes critérios pode ter levado à seleção de um conjunto funcionalmente enviesado de genes.

Tabela 1.7. Análise de enriquecimento de genes na categoria J do COG (tradução, estrutura do ribossomo e biogênese) ao longo dos passos de análise do trabalho de Kanhere e Vingron (2009). Variação cat. J: variação na proporção de genes na categoria J em relação ao passo anterior. Valor p: significância estatística desta variação, calculada por meio de um teste χ^2 com 10^6 simulações de Monte Carlo.

Passo	Genes	Genes cat. J	Fração cat. J	Variação cat. J	Valor p
início	144320	10573	0.07	-	-
1	56266	8265	0.15	2.00	$< 10^{-6}$
2	4178	711	0.17	1.16	1.7×10^{-4}
3	257	-	-	-	-
4	171	19	0.11	0.65	0.074
5	53	15	0.28	2.55	7.0×10^{-5}

1.4.5 Relação entre o número de genes no genoma e a proporção de candidatos a transferência horizontal

A Figura 1.10 mostra a proporção de candidatos a TGH, identificados pelo método baseado em BLAST, em função do número de genes no genoma, para os organismos com $\geq$ 20 vizinhos no banco de dados. Como pode ser observado nesta figura, há uma forte correlação positiva entre o número de genes no genoma e a proporção de candidatos a TGH. Os pontos distantes da linha de regressão, circulados em vermelho, correspondem aos organismos *Buchnera aphidicola Sg*, *Wigglesworthia glossinidia brevipalpis*, *Buchnera aphidicola (Baizongia pistaciae)*, *Candidatus Blochmannia floridanus*, *Baumannia cicadellinicola*, *Candidatus Blochmannia pennsylvanicus str. BPEN*, que compartilham as características de serem simbiontes intracelulares de insetos e terem genomas reduzidos. Na Figura 1.11, observa-se uma correlação pequena, porém significativa, entre o número de genes no genoma e a proporção de candidatos a TGH atribuídos à classe 2, isto é, aqueles genes para os quais não há ortólogos detectados na vizinhança filogenética do genoma em estudo e que correspondem, potencialmente, à transferência de funções novas para o genoma receptor (ver seções 1.3.3 e 1.4.2).

Já na figura 1.12, em que as predições de TGH foram feitas por análise de conteúdo G+C, não se observa correlação significativa entre proporção de candidatos a TGH e número de genes no genoma. A baixa correlação obtida para o método de conteúdo G+C, além da pouca regularidade nos padrões de enriquecimento funcional (seção 1.4.3), sugerem que a atipicidade no conteúdo G+C possa ser um preditor pobre de TGH. Uma possível explicação para isso é que transferências antigas ou entre genomas com conteúdo G+C semelhantes deixem de ser detectadas, e que detecções falsas possam ocorrer pelo fato de que alguns genes podem ter conteúdo G+C atípico devido a características funcionais, que podem implicar em frequências incomuns no uso dos aminoácidos. Métodos que usem características nucleotídicas evolutivamente neutras podem ter melhor desempenho na detecção de TGH.

Os dados obtidos pelo método baseado em BLAST concordam com dados

do estudo de Nakamura *et al.* (2004), em que também se observa uma forte correlação entre o número de genes no genoma e a proporção de TGH. O método usado por este grupo é porém, de natureza distinta: emprega modelos de Markov aplicados às sequências nucleotídicas. Estes dados, em conjunto, suportam a ideia de que genomas maiores têm uma fração maior de genes envolvidos em TGH. Os dados da Figura 1.11 sugerem também que os genomas maiores têm uma leve tendência a terem mais genes envolvidos em transferências de funções novas para o organismo receptor.

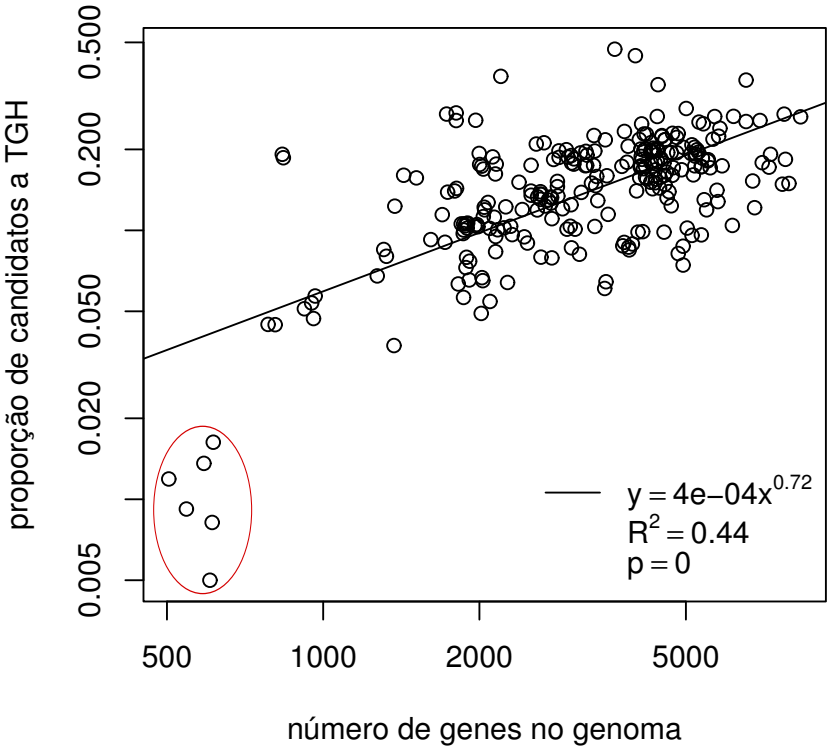

Figura 1.10 - Proporção de candidatos a TGH identificados pelo método baseado em BLAST em função do número de genes no genoma. Os pontos circulados em vermelho correspondem a organismos que são simbiontes intracelulares de insetos e têm genomas reduzidos.

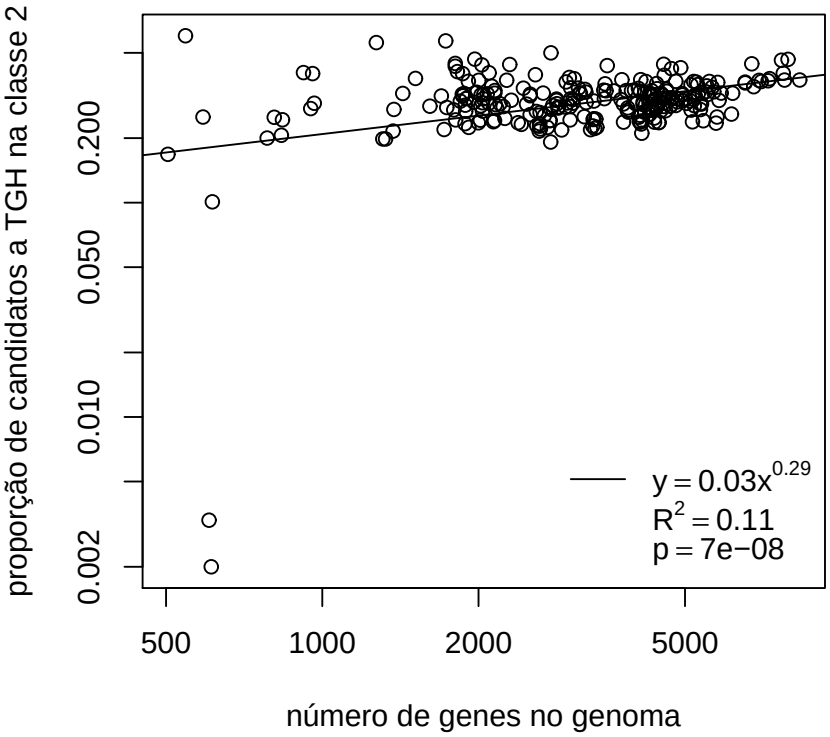

Figura 1.11 - Proporção de candidatos a TGH pertencentes à classe 2 (ver seção 1.3.3) em função do número de genes no genoma.

1.4.6 Teste da Hipótese da Complexidade

A existência de estudos em larga escala de interações proteína-proteína nos permite testar se nossas predições de TGH suportam a hipótese da complexidade. Para isto, usamos os dados de três estudos publicados, que usam diferentes técnicas experimentais para a identificação de interações proteína-proteína. Dois dos estudos são feitos em *E. coli*: um deles (Butland *et al.*, 2005) na linhagem DY330, derivada da W3110, e o outro (Arifuzzaman *et al.*, 2006) na própria W3110, que está representada no banco de dados Omniome. O estudo de Rain *et al.* (2001) é feito em *H. pylori*

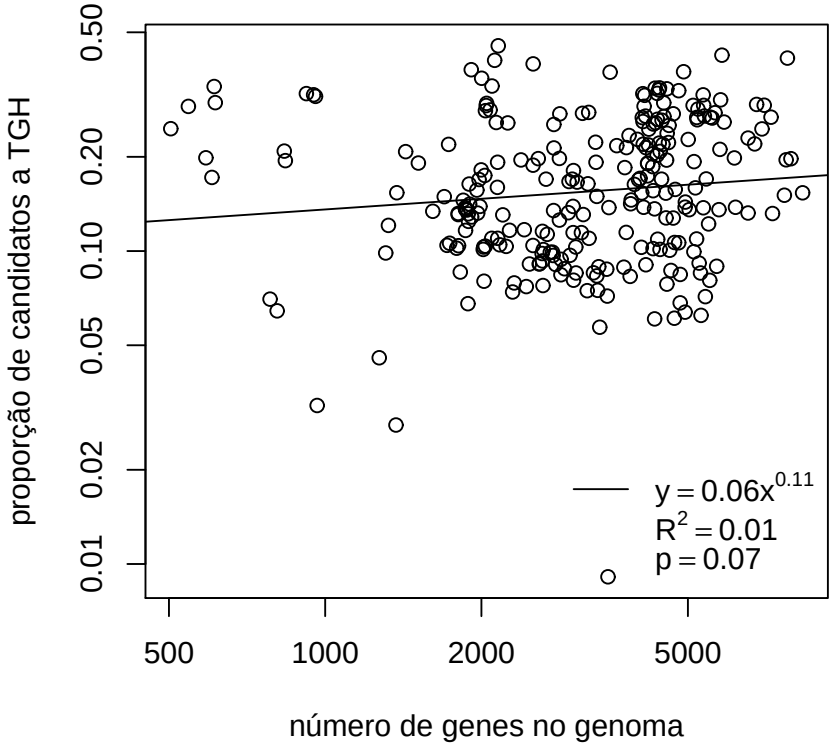

Figura 1.12 - Proporção de candidatos a TGH identificados por conteúdo G+C atípico em função do número de genes no genoma.

26695, que também está representada no banco de dados.

Os estudos em *E. coli* usam construções de proteínas "isca", pela adição de uma cauda purificável por afinidade, e posterior identificação das "presas", as proteínas que interagem com as iscas, por espectrometria de massa. O estudo com *H. pylori* usa técnica baseada no sistema duplo-híbrido de levedura para identificar interações proteína-proteína.

O teste aqui proposto para a hipótese da complexidade consiste em verificar se os candidatos a TGH têm significativamente menos parceiros de interação proteína-proteína que os genes com provável descendência vertical. A fim de evitar vieses devidos a dados gerados em condições experimentais diferentes, analisamos em separado os genes no papel de isca e de presa. Nas Figs. 1.13, 1.14 e 1.15 são mostrados histogramas do número de parceiros de interação proteica para os trabalhos de Butland *et al.* (2005), Arifuzzaman *et al.* (2006) e Rain *et al.* (2001), respectivamente. Estão também anotados nesta figura os resultados de testes de Mann-Whitney, que verificam se o número de parceiros é estocasticamente menor entre os genes atípicos que entre os típicos. Observa-se, nas Figs. 1.13 e 1.14, que os histogramas referentes aos genes atípicos (candidatos a TGH) concentram-se mais à esquerda em comparação com os histogramas referentes aos genes típicos (de provável descendência vertical) e que os testes de Mann-Whitney apresentam valores significativos ($p < 0.05$). Na Fig. 1.15, não se observam a mesma concentração à esquerda nos histogramas, nem valores significativos no teste estatístico.

Nas Figs. 1.13 e 1.14, observa-se menos distinção entre os histogramas dos genes atípicos e típicos para as análises com genes no papel experimental de isca. Os genes no papel de isca têm, pela construção experimental, mais parceiros potenciais (o conjunto das presas, que abrange praticamente todo o proteoma do organismo) que os genes no papel de presa, cujas interações detectáveis são somente aquelas com as iscas. Isto pode fazer com que os histogramas (tanto para genes típicos como atípicos) concentrem-se mais à direita em comparação com os das presas, como é de fato observado nestes gráficos, o que sugere que os números de interações proteicas

não sejam diretamente comparáveis entre genes no papel de presa e isca.

É interessante notar que no trabalho de Butland *et al.* (2005), para o qual se observa diferença mais significativa entre os histogramas dos genes atípicos e típicos (Fig. 1.13), há uma etapa de pré-tratamento com nucleases para minimizar a detecção de interações entre proteínas mediadas por ácidos nucleicos. É possível que nos outros dois trabalhos, que não mencionam tal tratamento, interações mediadas por ácidos nucleicos, e que, portanto, não dependem da compatibilidade física direta das proteínas envolvidas, sejam também contadas como interações proteína-proteína. Estas interações indiretas não seriam, em teoria, impedimento para a TGH, já que a proteínas codificadas pelos genes transferidos teriam que interagir somente com ácidos nucleicos, que, no caso do DNA e de certas moléculas de RNA, têm sua estrutura conservada ao longo da evolução.

Considerando-se as variações provavelmente devidas a particularidades experimentais de cada estudo, os resultados apresentados nesta seção para os trabalhos de Butland *et al.* (2005) e Arifuzzaman *et al.* (2006) sugerem que a hipótese da complexidade é compatível com as predições de TGH deste trabalho.

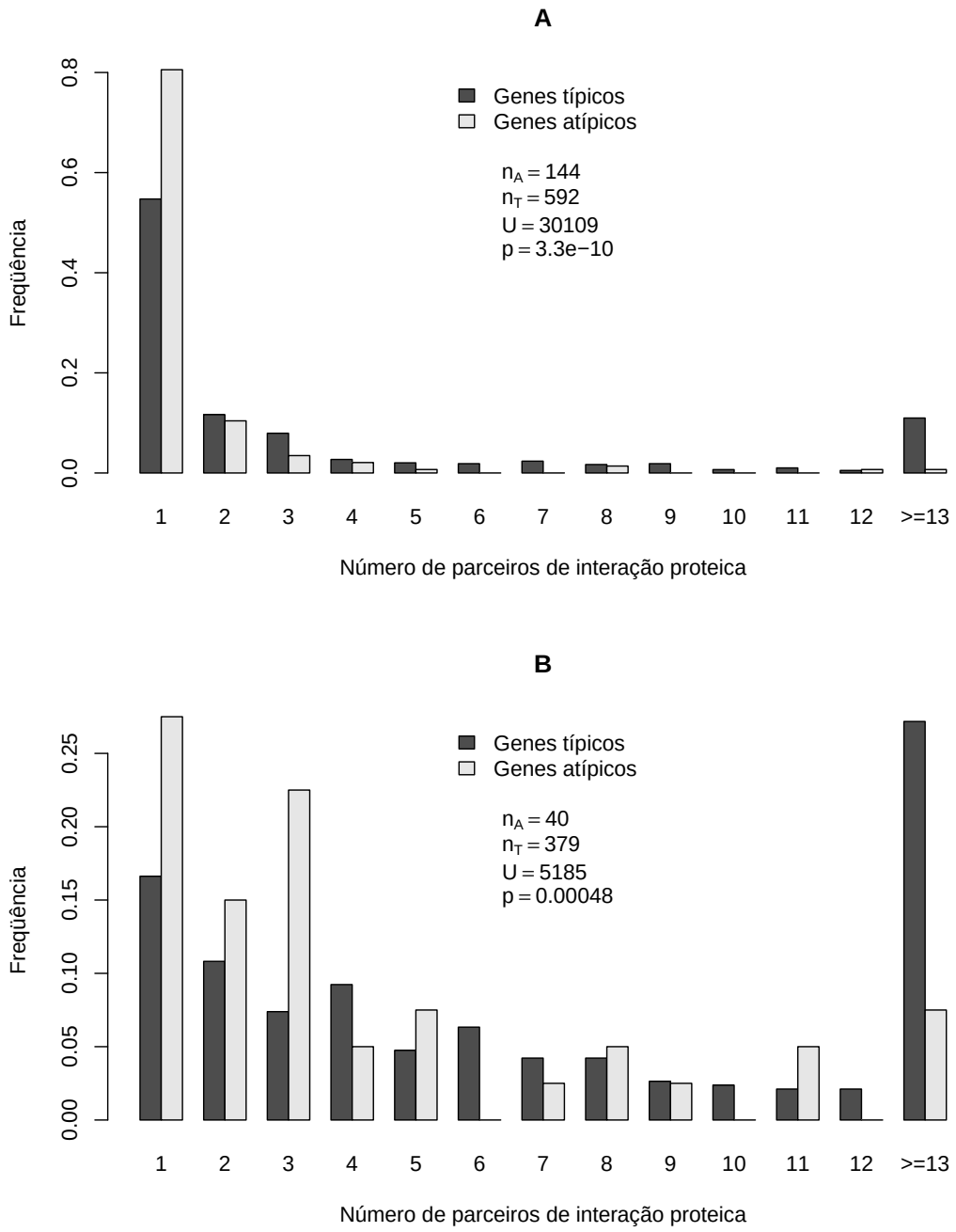

Figura 1.13 - Distribuições dos números de parceiros em interações proteína-proteína, obtidas do estudo em *E. coli* de Butland *et al.* (2005), para os genes atípicos (candidatos a TGH) e típicos (de provável descendência vertical). As distribuições são mostradas separadamente para os genes no papel experimental de presa (gráfico A) e de isca (gráfico B). A afirmação de que os genes atípicos têm significativamente menos parceiros que os típicos é verificada por meio de um teste de Mann-Whitney unicaudal. n_A: número de genes atípicos; n_T: número de genes típicos; U: estatística do teste de Mann-Whitney; p: valor p.

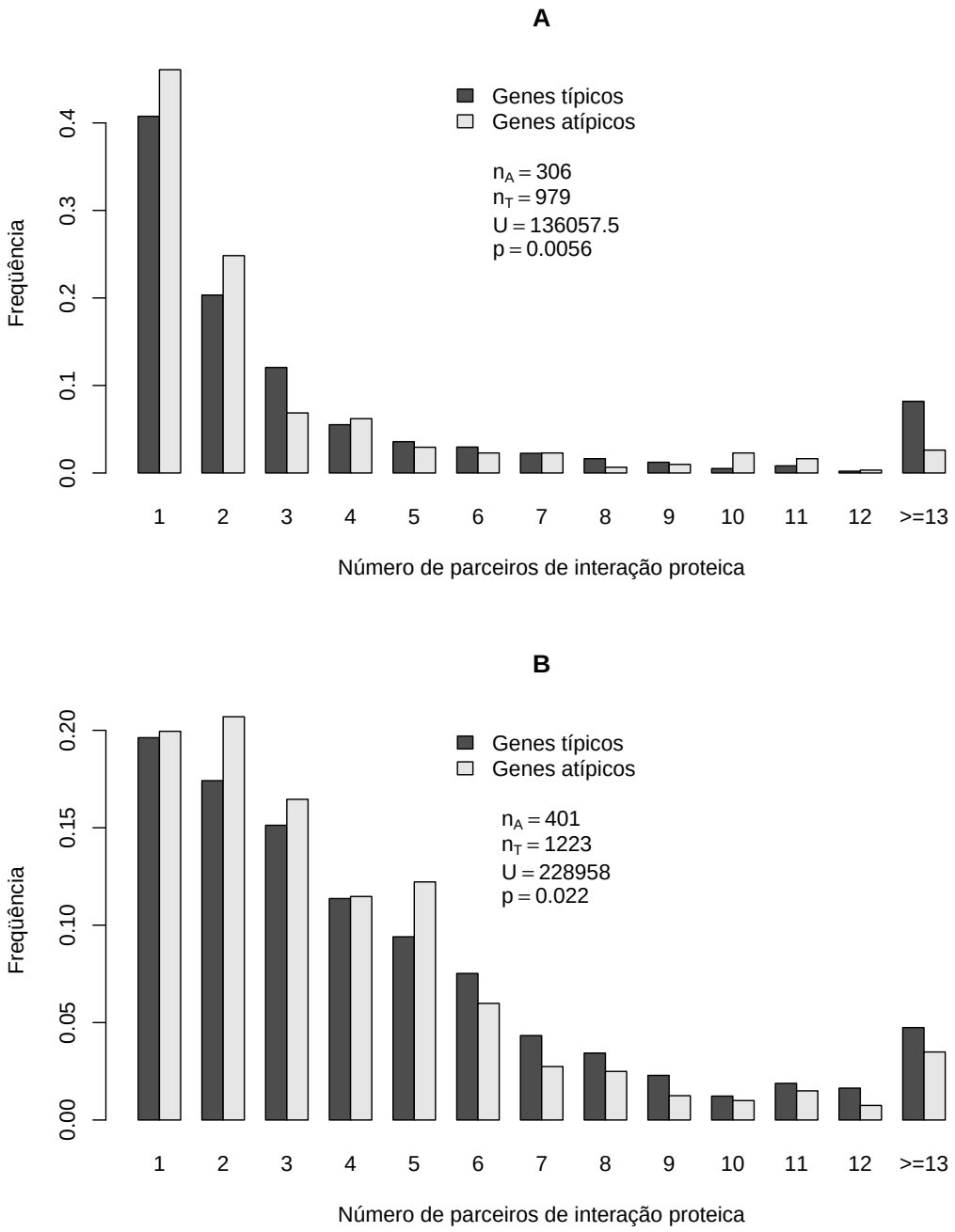

Figura 1.14 - Distribuições dos números de parceiros em interações proteína-proteína, obtidas do estudo em *E. coli* de Arifuzzaman *et al.* (2006), para os genes atípicos e típicos. As distribuições são mostradas separadamente para os genes no papel experimental de presa (gráfico A) e de isca (gráfico B). A afirmação de que os genes atípicos têm significativamente menos parceiros que os típicos é verificada por meio de um teste de Mann-Whitney unicaudal. n_A: número de genes atípicos; n_T: número de genes típicos; U: estatística do teste de Mann-Whitney; p: valor p.

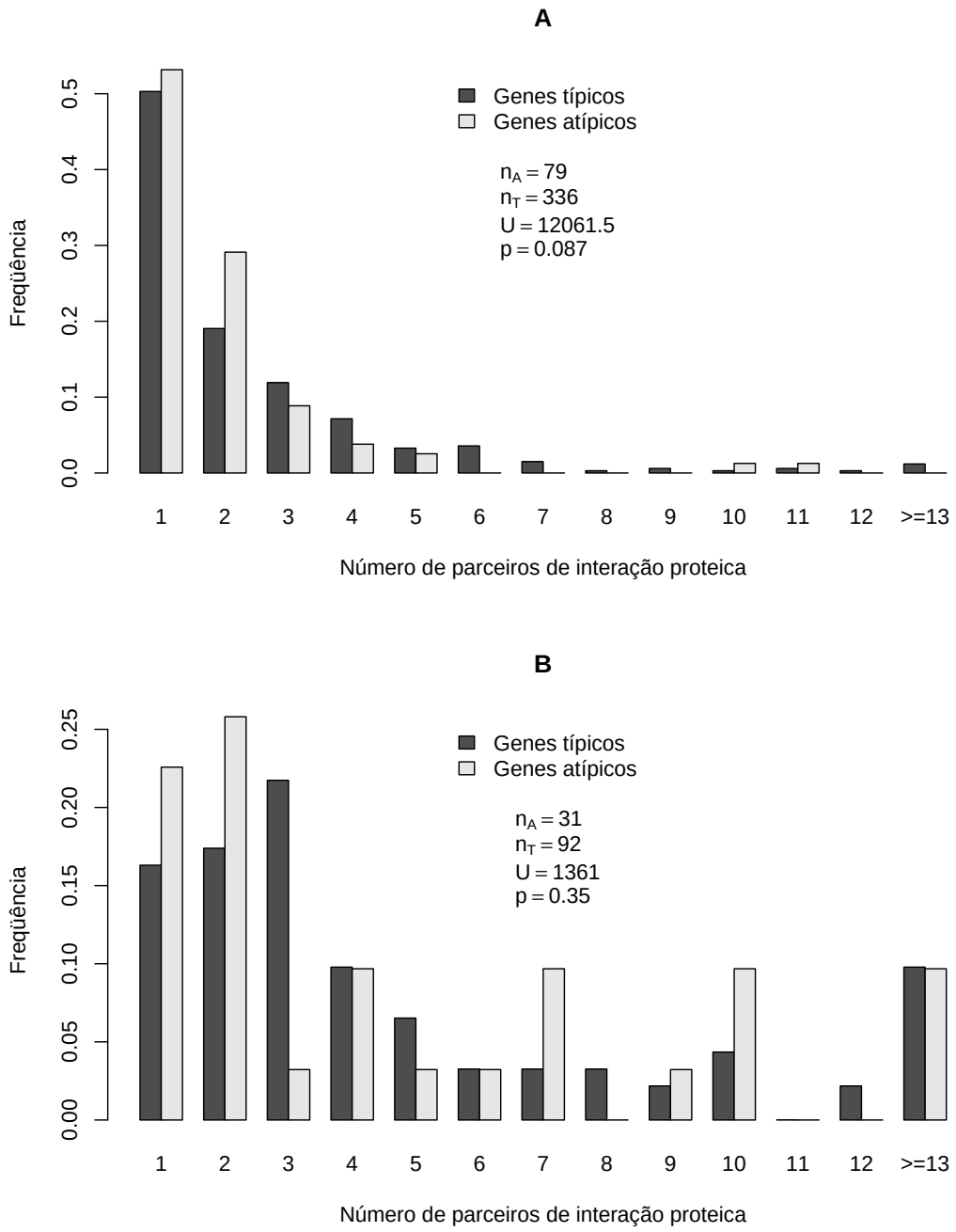

Figura 1.15 - Distribuições dos números de parceiros em interações proteína-proteína, obtidas do estudo em *H. pylori* de Rain *et al.* (2001), para os genes atípicos e típicos. As distribuições são mostradas separadamente para os genes no papel experimental de presa (gráfico A) e de isca (gráfico B). A afirmação de que os genes atípicos têm significativamente menos parceiros que os típicos é verificada por meio de um teste de Mann-Whitney unicaudal. n_A: número de genes atípicos; n_T: número de genes típicos; U: estatística do teste de Mann-Whitney; p: valor p.

1.5 Conclusões

Neste trabalho, desenvolvemos um método de baixo custo computacional para a detecção de TGH, que se baseia na análise dos alvos de BLAST do gene em estudo e nas distâncias de rRNA 16S aos organismos correspondentes. Por ser este método baseado em busca de similaridade de sequência, cabe lembrar que as predições de TGH não são definitivas pois taxas de evolução díspares, múltiplas perdas gênicas ou a baixa representatividade de genomas vizinhos no banco de dados podem levar a falsas detecções de TGH. Por outro lado, pelo fato do método usar recursos computacionais modestos, é possível fazer análises em larga escala com pouco tempo de processamento. O estudo aqui apresentado é, de acordo com nosso conhecimento, o estudo de TGH em maior escala feito até o momento, dentre aqueles que analisam enriquecimento funcional entre os genes potencialmente transferidos.

Dadas as características e limitações do método, podemos concluir que as predições de TGH feitas neste trabalho para 408 genomas procarióticos estão de acordo com as seguintes afirmações:

- que genes relacionados a transporte são os mais enriquecidos em candidatos a TGH na maioria dos genomas estudados;
- que genes relacionados a síntese proteica são os mais empobrecidos em candidatos a TGH na maioria dos genomas estudados;
- que, com base nos parâmetros usados, aproximadamente 14% dos genes procarióticos estão envolvidos em transferência horizontal;
- que a maior parte (72%) dos candidatos a TGH correspondem potencialmente a substituições de genes ortólogos, e cerca de 28% correspondem a aquisições de funções novas.
- que genes com funções operacionais participam mais de transferência horizontal que genes com funções informacionais;

- que as categorias funcionais mais enriquecidas ou empobrecidas em candidatos a TGH são geralmente as mesmas entre os genomas;
- que genomas grandes têm uma proporção maior de genes envolvidos em TGH que os genomas pequenos;
- que a hipótese da complexidade se verifica em dois de três estudos independentes de interações proteína-proteína.

2 Estudo da composição do regulon SOS de *Caulobacter crescentus*

2.1 Introdução

Esta seção trata da identificação de genes pertencentes ao regulon SOS do organismo modelo *Caulobacter crescentus*. Os resultados aqui apresentados são fruto de um trabalho conjunto que envolveu análises *in silico* e experimentos de laboratório, estes últimos conduzidos pela doutoranda Raquel P. Rocha e pelo pesquisador Rodrigo S. Galhardo. Este trabalho foi publicado no periódico científico "Journal of Bacteriology", no artigo "Characterization of the SOS regulon of *Caulobacter crescentus*" (da Rocha *et al.*, 2008).

2.1.1 Regulação do sistema SOS em *Escherichia coli*

O sistema SOS é um mecanismo amplamente presente nos procariotos que regula respostas adaptativas da célula frente a danos no DNA. Em *Escherichia coli*, organismo no qual é mais estudado, o regulon SOS consiste num conjunto de mais de 40 genes, com funções biológicas diversas, que são induzidos na presença de danos no DNA (Friedberg *et al.*, 2006).

O mecanismo de regulação do sistema SOS de *E. coli* é ilustrado na Fig. 2.1. Em células que não sofreram danos no DNA, a expressão dos genes do regulon SOS é reprimida pela ligação de dímeros da proteína LexA em sequências operadoras, localizadas, cada qual, na região promotora de um gene ou operon do regulon SOS. As sequências operadoras às quais LexA se liga são chamadas de caixas SOS. O

dímero de LexA, ligado à caixa SOS, dificulta a ligação da RNA polimerase à região promotora, reprimindo assim a expressão do gene correspondente. A repressão destes genes não é, no entanto, absoluta: mesmo no estado não induzido, os genes do regulon SOS são expressos em níveis basais.

Quando a célula tenta replicar uma região danificada do DNA ou quando a replicação normal é interrompida, há a geração de sequências de fita simples de DNA próximas à forquilha de replicação. A proteína RecA liga-se a estas regiões de fita simples, resultando na formação de filamentos em forma de hélice, de DNA e RecA. Estes filamentos interagem com a proteína LexA, promovendo a autoclivagem de LexA em um sítio específico, próximo ao meio de sua sequência proteica. A clivagem de LexA a torna incapaz de atuar como repressora, além de expor sítios que a tornam alvo para degradação. Como consequência, os níveis da proteína LexA diminuem, e os genes SOS, que incluem as próprias LexA e RecA, têm sua expressão aumentada.

Com o restabelecimento da replicação normal do DNA, por meio da atuação de mecanismos reparo de DNA ou de tolerância a lesões, as regiões de DNA de fita simples desaparecem. Como consequência, há um novo acúmulo da proteína LexA intacta e o retorno ao estado não induzido do regulon SOS.

O bloqueio da replicação do DNA constitui uma forma robusta da célula detectar condições adversas ao seu crescimento. Entre os tratamentos que causam indução do sistema SOS em *E. coli*, alguns, como luz ultravioleta (UV), radiação ionizante, mitomicina e peróxido de hidrogênio, causam danos no DNA, que, por sua vez, bloqueiam a maquinaria de replicação. Outros tratamentos, como deprivação de timidina (em bactérias auxotróficas para timidina) ou introdução de análogos de bases, têm o efeito direto de impedir o progresso da replicação do DNA. Em ambos os casos, há a formação de regiões de fita simples de DNA, que é o sinal inicial para a indução do regulon SOS.

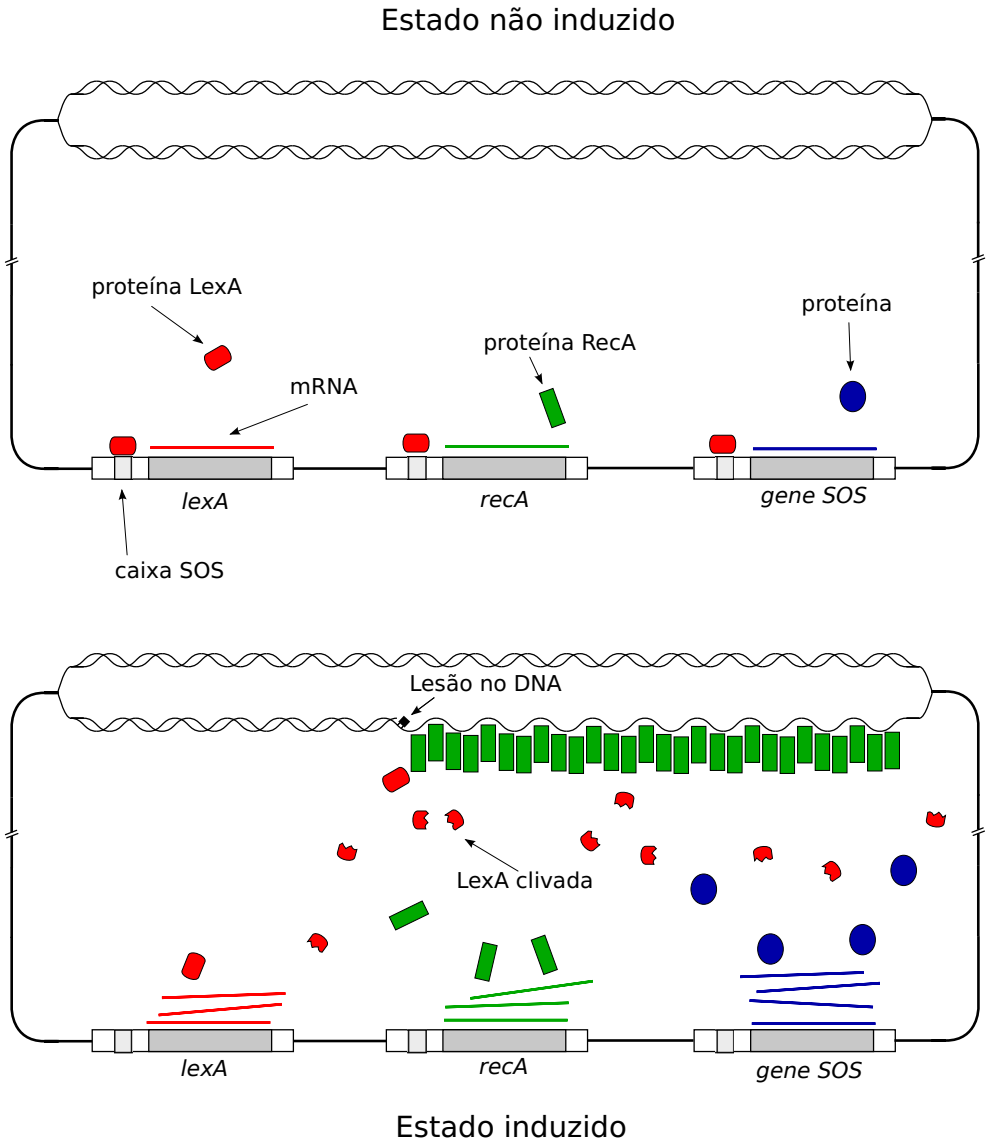

Figura 2.1: Mecanismo de regulação do sistema SOS. No estado não induzido, a proteína LexA age como repressora dos genes SOS, que são transcritos em níveis basais. Na presença de lesões no DNA, há um acúmulo de regiões de fita simples no DNA, que se associam à proteína RecA em forma de filamentos. Os filamentos RecA-DNA promovem a autoclivagem da proteína LexA, que perde a capacidade de se ligar às caixas SOS e reprimir a expressão dos genes correspondentes. A diminuição nos níveis de LexA intacta faz com que os genes do regulon SOS tenham sua expressão aumentada (estado induzido). Quando a replicação é restabelecida em função da atuação de mecanismos reparo de DNA ou de tolerância a lesões, o sinal indutor do sistema SOS (DNA de fita simples) deixa de existir, permitindo um novo acúmulo da proteína LexA intacta, o que leva o regulon SOS novamente ao estado não induzido. (Adaptado de Friedberg *et al.*, 2006).

2.1.2 Identificação do regulon SOS em *Escherichia coli*

Os estudos que contribuíram para a identificação dos genes do regulon SOS em *E. coli* variam grandemente quanto à metodologia utilizada. A seguir, citamos três destes estudos, que usaram abordagens em escala genômica.

Usando um bacteriófago modificado, Kenyon e Walker (1980) construíram mutantes de *E. coli* com inserções aleatórias de uma sequência que carrega os genes estruturais do operon repórter *lac*, mas não seu promotor. Quando esta sequência é inserida em um gene no quadro de leitura correto, cria-se uma fusão transcricional, e o operon *lac* passa a ser expresso sob o controle do promotor do gene hospedeiro. Aplicando tratamentos de luz UV e mitomicina C aos mutantes, os autores puderam identificar um conjunto de *loci* que são expressos em resposta a estes agentes que causam dano ao DNA. Além disso, verificaram que estes *loci* não são induzidos em mutantes deficientes em *recA* ou em mutantes *lexA* não induzíveis (*lexA*(Ind-)), aqueles nos quais a proteína LexA não tem a atividade de autoclivagem.

Fernandez de Henestrosa *et al.* (2000) usaram uma medida chamada índice de heterologia (Lewis *et al.*, 1994) para identificar sequências similares ao consenso da caixa SOS: 5'-TACTG$(TA)_5$CAGTA-3' ao longo do genoma de *E. coli*. Os genes correspondentes às potenciais caixas SOS encontradas tiveram sua expressão analisada por meio de ensaios de Northern *blot* na cepa selvagem, na cepa *lexA*(Ind-), na qual o regulon SOS não é induzível, e na cepa *lexA*(Def), na qual o regulon SOS é expresso constitutivamente. Com esta estratégia, os autores puderam identificar 7 novos genes regulados por LexA além dos conhecidos até então.

Courcelle *et al.* (2001) usaram microarranjos de DNA contendo 95.5% dos genes preditos de *E. coli* para analisar o perfil transcricional das cepas selvagem e *lexA*(Ind-) após irradiação com luz UV. Com este método, os autores confirmaram a indução dependente de *lexA* para os 26 operons (incluindo operons de 1 gene) previamente documentados e identificaram 17 novos operons que respondem de maneira dependente de *lexA*.

2.1.3 Principais funções do regulon SOS

Entre os genes induzidos na resposta SOS em *E. coli*, são listados, na Tabela 2.1, alguns dos genes relacionados a reparo e tolerância a lesões no DNA (Friedberg *et al.*, 2006).

Tabela 2.1. Genes do regulon SOS de *E. coli* com funções relacionadas a reparo e tolerância a lesões no DNA

Genes	Função
uvrA e *uvrB*	Reparo por excisão de nucleotídeos.
dinB e *polB*	DNA polimerases de síntese translesão, de baixa fidelidade.
umuC e *umuD*	O complexo UmuC+UmuD' (UmuD clivada) forma uma DNA polimerase de síntese translesão, de baixa fidelidade.
sulA	Inibidor da divisão celular.
recN	Participa no reparo de quebras duplas no DNA.
recA	Participa no reparo recombinacional do DNA e na regulação do sistema SOS.
ruvA	Participa como subunidade do complexo RuvABC na resolução de junções de Holliday durante a recombinação homóloga.

As proteínas de reparo por excisão de nucleotídeos e de reparo recombinacional agem, por mecanismos distintos, no reparo das lesões presentes no DNA. Caso este reparo tenha sucesso, a replicação do DNA pode prosseguir, o que faz com que a célula volte ao estado não induzido do sistema SOS.

As polimerases de síntese translesão, capazes de polimerizar DNA mesmo que a fita molde tenha lesões, são um recurso importante para aumentar a chance de sobrevivência da célula com o DNA danificado. Estas polimerases são, no entanto, mutagênicas: incorporam nucleotídeos não pareados com uma frequência maior que a polimerase replicativa. A indução destas polimerases pelo sistema SOS tem também um papel evolutivo: com a taxa de mutações aumentada, é maior a probabilidade de

surgirem mutações que tornem a célula resistente às condições adversas a que está submetida, como no caso das mutações que geram resistência a antibióticos.

A proteína SulA inibe a polimerização da proteína FtsZ e, consequentemente, a formação do septo e a divisão celular. Tal efeito gera uma condição observável ao microscópio: a formação de filamentos de bactérias que não completaram a divisão. A indução do gene sulA como resposta SOS, funciona como um *checkpoint* baseado em danos no DNA, que impede que as duas cópias-filhas do cromossomo sejam separadas pela divisão celular.

2.1.4 O modelo *Caulobacter crescentus*

Caulobacter crescentus é uma bactéria não patogênica, de vida livre, classificada no grupo das proteobactérias-α, e é encontrada predominantemente em ambientes aquáticos. Seu genoma, sequenciado em 2001, é composto de um cromossomo circular, com 4.016.942 pares de bases, que codifica 3767 genes (Nierman *et al.*, 2001). *C. crescentus* é, no presente, o principal modelo bacteriano para estudos de diferenciação celular em função da característica assimétrica de sua divisão e do acoplamento entre ciclo celular e diferenciação (Brown *et al.*, 2009). Da divisão desta bactéria originam-se duas células morfologicamente distintas: uma célula flagelada, capaz de se locomover, e uma célula que possui um talo em um de seus polos, e é capaz de se fixar a superfícies sólidas por meio do pedúnculo, estrutura adesiva localizada na extremidade do talo. A Fig. 2.2 ilustra o ciclo celular de *C. crescentus*. A célula móvel, que não pode se dividir, diferencia-se em uma célula talo, processo que envolve a perda do flagelo e dos *pili* e a formação de um talo no polo anteriormente ocupado pelo flagelo. Em seguida, a célula talo inicia a divisão celular, processo em que ocorrem a formação de um flagelo no polo oposto ao do talo, a separação dos cromossomos e a separação das células filhas. A célula talo filha pode iniciar diretamente a replicação do DNA, enquanto que a célula móvel precisa antes iniciar sua diferenciação em célula talo. A replicação do cromossomo, a formação das estruturas polares e a divisão celular são processos coordenados que dependem da

produção, ativação, localização e degradação de proteínas regulatórias em tempos precisos durante o ciclo celular (Skerker e Laub, 2004; Brown *et al.*, 2009).

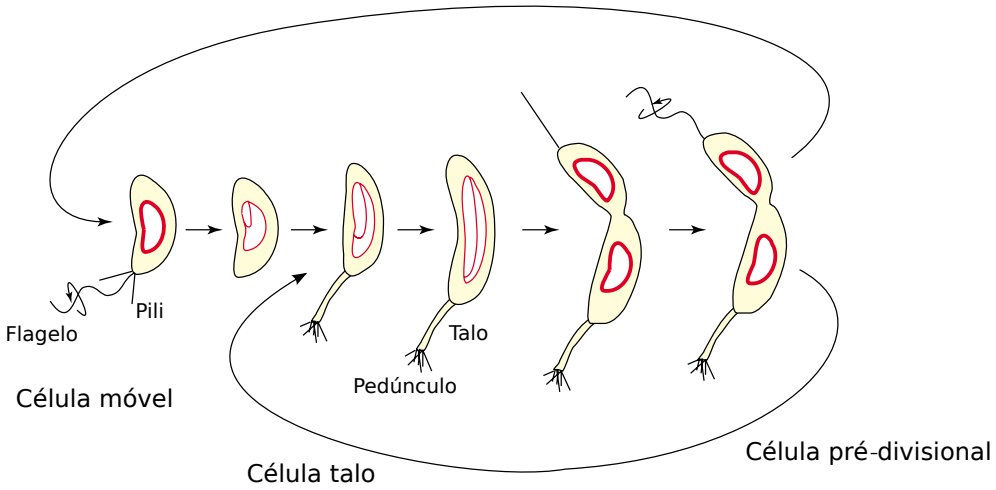

Figura 2.2: Ciclo celular de *Caulobacter crescentus*. A divisão celular origina uma célula móvel, flagelada, e outra séssil, que possui um talo em um de seus polos, permitindo a fixação da bactéria em superfícies sólidas. A célula talo filha pode iniciar um novo ciclo reprodutivo, com a replicação do DNA. Já a célula móvel, que não é capaz de replicar seu DNA, precisa antes iniciar sua diferenciação em célula talo, processo que envolve a perda do flagelo e dos *pili* e a formação do talo. (Adaptado de Quardokus e Brun, 2003).

2.2 Objetivos

Um dos objetivos gerais da pesquisa em nosso laboratório é o de investigar mecanismos de reparo de DNA e tolerância a lesões em *Caulobacter crescentus* e estudar as possíveis interações entre estes mecanismos e a regulação do ciclo celular. Nesta linha, Galhardo *et al.* (2005) identificaram um operon regulado pelo sistema SOS de *C. crescentus*, cujos produtos são necessários para a síntese de DNA de baixa fidelidade, tolerante a lesões. Este operon é composto dos genes hipotéticos *imuA* e *imuB* – este último codifica uma proteína similar às polimerases de síntese translesão da família Y – e do gene *dnaE2*, que codifica uma segunda cópia da

subunidade catalítica da polimerase III. Este operon é responsável pela maioria das mutações induzidas por luz UV e mitomicina C em *C. crescentus*. Também dentro deste mesmo objetivo geral, Martins-Pinheiro *et al.* (2007) identificaram genes de *C. crescentus* potencialmente envolvidos em reparo de DNA por análise de similaridade com sequências de proteína de outras espécies nas quais mecanismos de reparo de DNA estão melhor descritos.

O presente trabalho tem por objetivo a identificação, em escala genômica, de genes pertencentes ao regulon SOS de *C. crescentus* por meio da predição computacional de caixas SOS e análise da expressão dos genes correspondentes em um mutante deficiente em *lexA*.

2.3 Métodos

Neste estudo, usamos uma estratégia iterativa que permite expandir, incrementalmente, o conjunto de genes atribuídos ao regulon SOS de *C. crescentus*. Esta estratégia consiste na repetição iterada dos seguintes passos: (1) construção de um modelo para a caixa SOS a partir de regiões promotoras de genes pertencentes ao conjunto correntemente atribuído ao regulon SOS; (2) predição, usando este modelo, de quais genes têm caixas SOS em suas regiões promotoras; e (3) análise experimental da expressão relativa destes genes em uma cepa de *C. crescentus* deficiente em LexA (*lexA(Def)*), em comparação com a cepa selvagem. Uma vez que um gene é confirmado como sendo regulado por LexA, sua região promotora é usada, na próxima iteração, na construção de um novo modelo, mais refinado, para a caixa SOS.

O modelo utilizado neste trabalho para a caixa SOS é o da matriz posicional de probabilidades de nucleotídeos. Este modelo representa um sítio de ligação com proteína por uma matriz M, de 4 linhas e w colunas, onde w é o número (fixo) de posições nucleotídicas compreendidas pelo sítio. Cada coluna especifica probabi-

lidades de ocorrência de A, C, G e T para uma posição no sítio. Matematicamente:

$$\mathbf{P} = \begin{pmatrix} p_{\mathrm{A},1} & p_{\mathrm{A},2} & \cdots & p_{\mathrm{A},w} \\ p_{\mathrm{C},1} & p_{\mathrm{C},2} & \cdots & p_{\mathrm{C},w} \\ p_{\mathrm{G},1} & p_{\mathrm{G},2} & \cdots & p_{\mathrm{G},w} \\ p_{\mathrm{T},1} & p_{\mathrm{T},2} & \cdots & p_{\mathrm{T},w} \end{pmatrix}$$

Seja $x = x_1\, x_2 \ldots x_w$ uma sequência de DNA correspondente a um potencial sítio de ligação da proteína, sua pontuação, segundo o modelo, é dada por:

$$S(x) = \sum_{i=1}^{w} \log \frac{p_{x_i,i}}{b_{x_i}} \tag{2.1}$$

o onde $p_{x_i,i}$ é a probabilidade do nucleotídeo x_i na posição i do sítio e b_{x_i} é a probabilidade geral de ocorrência do nucleotídeo x_i no genoma.

Este modelo é bastante popular entre os métodos de busca de sítios de ligação de proteínas no DNA por ser simples o suficiente para ser estimado com poucos exemplos de sítios e por permitir que se dê, para cada sítio, uma pontuação teoricamente relacionada com a energia de sua ligação com a proteína (Stormo, 2000; Berg e von Hippel, 1987).

Na Fig. 2.3, temos um exemplo de *sequence logo* (Schneider e Stephens, 1990), ferramenta de visualização comumente empregada para a matriz posicional de probabilidades de nucleotídeos. A altura total de cada coluna é dada por $2 - H_i$, onde $H_i = -\sum_{y \in \{A,C,G,T\}} p_{y,i} \log_2 p_{y,i}$ é a entropia da i-ésima coluna da matriz. A altura de cada letra é proporcional à probabilidade $p_{y,i}$ de cada nucleotídio em cada posição. Esta representação gráfica permite uma visualização direta de quais sítios são mais conservados (valor alto de $2 - H_i$), e portanto mais importantes na ligação com a proteína, além de informar qual ou quais os nucleotídeos mais prováveis em cada posição do modelo.

O esquema de pontuação da Eq. 2.1 supõe que a contribuição de cada nucleotídeo para a força de ligação DNA-proteína seja independente de quais nucleotídeos estão nas outras posições. Embora em alguns casos esta suposição não seja válida, ela provê simplicidade ao modelo sem sacrificar muito sua acurácia (Stormo,

2000). Modelos mais complexos que preveem dependências entre posições nucleotídicas foram desenvolvidos (por exemplo, Georgi e Schliep, 2006). A estimação de seus parâmetros exige, no entanto, mais exemplos de sítios de ligação DNA-proteína.

Usamos, para a estimação da matriz no passo (1), o programa Gibbs Motif Sampler (Thompson *et al.*, 2003), fornecendo a este as regiões de -250 a +50 nucleotídeos do códon de início do conjunto corrente de genes regulados por LexA. A matriz resultante é usada, no passo (2), para a identificação de sítios com alta pontuação, segundo a Eq. 2.1, na sequência genômica de *C. crescentus*. Os genes imediatamente a jusante dos sítios de pontuação mais alta são selecionados para a análise experimental de expressão por meio de reação em cadeia da polimerase em tempo real. Genes que têm expressão relativa medida maior que 2 ou menor que 0.5 são considerados como parte do regulon SOS e introduzidos no passo (1) da próxima iteração. Para a construção do modelo inicial da caixa SOS, usamos as regiões promotoras dos genes lexA, recA, uvrA, recN (cujos ortólogos são parte do sistema SOS de *E.coli*) e imuA (CC_3213), para o qual foi verificada, em nosso laboratório, sua regulação por LexA (Galhardo *et al.*, 2005).

A parte experimental deste trabalho, desenvolvida pela doutoranda Raquel P. Rocha e pelo pesquisador Rodrigo S. Galhardo, consistiu na construção da cepa *lexA(Def)*, por meio da deleção da maior parte da região codificante do gene *lexA*, e na análise da expressão relativa dos genes candidatos a SOS, por meio da reação em cadeia da polimerase em tempo real (*real-time RT-PCR*). As medidas de expressão gênica por *real-time RT-PCR* foram feitas em 3 a 7 réplicas biológicas, usando-se como referência o gene *rho*, cuja expressão não varia com a indução do sistema SOS. (Mais detalhes sobre os experimentos encontram-se na referência: da Rocha *et al.*, 2008). A análise estatística da expressão diferencial entre as cepas selvagem e *lexA(Def)* foi feita da seguinte maneira: para cada réplica biológica, calcula-se o nível de expressão do gene candidato em unidades de expressão do gene *rho* para as duas cepas. A seguir, verifica-se, por meio do teste t de Student pareado, se a diferença de expressão entre as duas cepas é significativa.

2.4 Resultados e discussão

Após um número de iterações da estratégia experimental descrita na seção anterior, obtivemos as caixas SOS e as medidas de expressão relativa mostradas na Tabela 2.2. De um total de 44 genes analisados experimentalmente, 35 tiveram expressão aumentada em mais que duas vezes na cepa *lexA(Def)*, indicando o papel de LexA na repressão destes genes. Dois dos genes analisados tiveram sua expressão diminuída em mais que duas vezes nesta cepa, o que sugere que sua transcrição seja estimulada na presença de LexA. Estes 37 genes que tiveram expressão alterada por um fator maior que 2, estão distribuídos em 32 unidades transcricionais preditas (transcritos monocistrônicos ou operons). As regiões promotoras destas 32 unidades transcricionais foram usadas para se construir um modelo para a caixa SOS de *C. crescentus*. A Tabela 2.3 mostra o modelo obtido em forma de matriz posicional de probabilidade de nucleotídeos e a Fig. 2.3 mostra este modelo em forma de *sequence logo*.

Tabela 2.2: Medidas da expressão relativa de genes de *C. crescentus* entre as cepas *lexA(Def)* e selvagem. P: pontuação da caixa LexA de acordo com o modelo de matriz posicional de probabilidade de nucleotídeos; ER: média da razão de expressão do gene entre as cepas *lexA(Def)* e selvagem obtidas através de 2 a 5 réplicas experimentais ± o desvio-padrão destas medidas; D: posição relativa (em nucleotídeos) da caixa LexA predita em relação o códon de início predito. ∗: nível de expressão significativamente diferente entre as cepas selvagem e *lexA*- ($p < 0.05$, usando teste t de Student pareado entre as réplicas biológicas).

Gene	Função	Caixa SOS	P	D	ER	
CC_1902	*lexA*	AATGTTCTCCTGGTGTTCC	14.3	-51	43.2	
CC_0627	Proteína hipotética	AAAGTTCGCGTTATGTTCT	18.8	-9	40.3	
CC_3467	Proteína hipotética conservada	CATGTTCCAGCTTTGTTCG	13.6	-20	37.5	
CC_3518	Proteína hipotética conservada	GATGTTCATGTATTGTTCT	18.8	-3	27.4	
CC_2332	Proteína hipotética conservada	ATCGTTCTTGATTTGTTCT	18.7	-13	18.8	∗
CC_2333	Proteína relacionada à Uracil DNA glicosilase	*em operon com CC_2332*			8.3	
CC_1926	*dnaE*/DNA polimerase III, subunidade alfa	CATATTCCGGTTTTGTTCT	16.4	-165	1.6	
		AGATTTCTTGTTTTGTTCC	11.4	-181		
		TCTGTTCACAAGATGTTCC	11.1	-147		
CC_1927	Proteína hipotética	*em operon com CC_1926*			17.7	∗
CC_3213	*imuA*/Proteína indutível de mutagênese A	CATGTTCCACTTTTGTTCT	17.9	-73	16.2	
CC_2272	Proteína da família da endonuclease III	AATGTTCTTGTTATGTTCT	23.2	-26	14.7	

Tabela 2.2: continuação.

Gene	Função	Caixa SOS	P	D	ER	
CC_3424	Proteína hipotética conservada	AATGTTCCTGAATTGTTCT	20.7	-26	13.62	
CC_1330	Proteína com domínio Radical SAM	TATGTTCTTGTTATGTTCG	20.6	-33	11.7	*
CC_1054	Proteína hipotética	TTTGTTCTCGGCTTGTTCT	16.3	-3	11.3	*
CC_2040	RNA helicase dependente de ATP da família DEAD/DEAH	CATGTTCCCTTTCTGTTTC	14.0	-24	9.1	
CC_1087	*recA*/Proteína de recombinação A	CATGTTCGCAAGATGTTCC	15.5	-114	9.0	*
CC_2879	Proteína hipotética	CATGTTCTGACTATGTTCC	14.3	+56	8.0	
CC_2880	Proteína hipotética	*em operon com CC_2879*			9.74	*
CC_2881	*uvrC*/excinuclease ABC, subunidade C	*em operon com CC_2879*			1.7	
CC_3038	Proteína hipotética conservada	AATGTTCCTATAATGTTCT	21.5	-160	5.3	
CC_3037	Proteína hipotética conservada	*em operon com CC_3038*			7.7	
CC_3036	Proteína hipotética	*em operon com CC_3038*			7.3	
CC_3039	Proteína hipotética	AATGTTCCTATAATGTTCT	21.5	-159	4.4	
CC_3356	Proteína hipotética	CATGTTCTCGTATTGTTCG	18.1	-52	6.3	
CC_1531	Proteína hipotética	ATTGTTCTTGATATGTTCC	20.2	-31	5.9	
		TATGTTCCAACTTCGTTTG	11.3	-20		
CC_1983	*recN*	GATGATCCCGTTTCGTTCC	11.9	-56	5.6	*

Tabela 2.2: continuação.

Gene	Função	Caixa SOS	P	D	ER	
CC_0140	*comM*/Proteína de competência ComM	AACGTTCGTTTTTCGTTCT	15.7	-72	4.7	*
CC_0383	Proteína hipotética	TATGTTCCTGAAAAGTTCT	18.5	-14	5.0	*
CC_3238	*ruvC*	CGCGTTCATCATGTGTTCT	10.4	+2	5.1	
CC_3237	*ruvA*	*em operon com CC_3238*			5.0	
CC_3236	*ruvB*	*em operon com CC_3238*			3.4	
CC_3225	Sensory box sensor histidine kinase/response regulator	TTTGTTCGCCAGATTTTTT	12.9	+8	4.8	
CC_0382	*tag*/DNA-metiladenina glicosilase I	TATGTTCCTGAAAAGTTCT	18.5	-44	3.4	
CC_2590	*uvrA*/excinuclease ABC, subunidade A	TTTGTTCGCATCTTGTTCT	17.7	-87	3.5	*
		CTTGTTCTCGCGACGTTCG	10.6	-268		
CC_1532	Proteína hipotética conservada	ATTGTTCTTGATATGTTCC	20.2	+32	3.1	
		TATGTTCCAACTTCGTTTG	11.3	+21		
CC_3515	Proteína hipotética conservada	AGAGTTCGCATTATGTTCT	15.7	-79	3.1	
CC_1468	*ssb*/Proteína de ligação ao DNA de fita simples	TTTGTTCTCATAACGTTCT	18.6	-93	2.1	*
CC_3130	Proteína da família da sintetase de glutamina	TTTGTTCTCGAAAGGTTTC	14.4	-52	2.1	
		GTTTTTCCGGATTTGTTCT	11.7	-41		

Tabela 2.2: continuação.

Gene	Função	Caixa SOS	P	D	ER
CC_2878.1	Proteína hipotética conservada	CATGTTCTGACTATGTTCC	14.3	-55	1.4
CC_2589	Proteína hipotética	TTTGTTCGCATCTTGTTCT	17.7	-154	0.9
		CTTGTTCTCGCGACGTTCG	10.6	+27	
CC_1928	Hidrolase de nucleosídeos que prefere inosina e uridina	CATATTCCGGTTTTGTTCT	16.4	-126	0.8
		AGATTTCTTGTTTTGTTCC	11.4	-110	
		TCTGTTCACAAGATGTTCC	11.1	-144	
CC_1086	Sensory box protein	CATGTTCGCAAGATGTTCC	15.5	-114	0.8
CC_3214	Sintetase de carbamoil-fosfato/transferase de carboxil	CATGTTCCACTTTTGTTCT	17.9	-103	0.6
CC_1665	*dnaB*/DNA helicase replicativa	GATGTTCTGTGTATGTTTT	14.5	-73	0.3
CC_2433	Proteína hipotética conservada	ATTATTTTCATTATGTTTT	16.5	-105	0.1

Tabela 2.3: Modelo para os sítios de ligação de LexA em *C. crescentus* em forma de matriz posicional de probabilidade de nucleotídeos. Cada coluna corresponde a uma posição nucleotídica na caixa SOS, numerada de 1 a 19. O valor de cada célula corresponde à probabilidade de se encontrar o respectivo nucleotídeo naquela posição.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	0.37	0.61	0.07	0.07	0.04	0.01	0.01	0.10	0.10	0.22	0.34	0.28	0.55	0.07	0.01	0.01	0.01	0.01	0.01
C	0.25	0.01	0.10	0.04	0.01	0.01	0.92	0.31	0.43	0.13	0.04	0.13	0.04	0.16	0.01	0.01	0.07	0.79	0.22
G	0.10	0.07	0.01	0.89	0.04	0.01	0.04	0.22	0.07	0.52	0.07	0.10	0.07	0.04	0.95	0.01	0.01	0.01	0.13
T	0.28	0.31	0.83	0.01	0.92	0.98	0.04	0.37	0.40	0.13	0.55	0.49	0.34	0.73	0.04	0.98	0.92	0.19	0.64

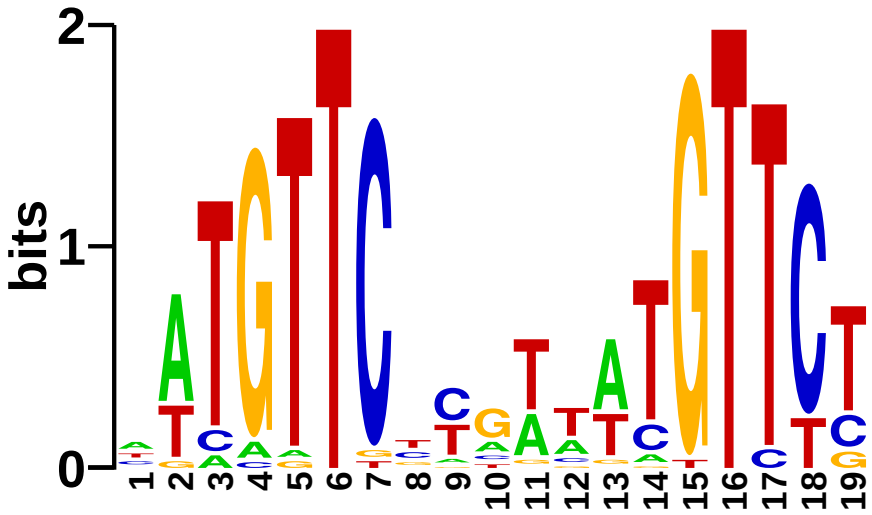

Figura 2.3: *Sequence logo* (Schneider e Stephens, 1990) do modelo usado para a busca de potenciais sítios de ligação da proteína LexA no genoma de *C. crescentus*.

Dentre os genes com expressão aumentada na cepa deficiente em LexA, a maior parte tem função biológica desconhecida, e recebe a anotação de proteína hipotética ou proteína hipotética conservada. Alguns dos genes com expressão aumentada são conhecidos como pertencentes ao regulon SOS em outras proteobactérias: *recA*, *lexA*, *uvrA* e *recN*. Os genes CC_3238, CC_3237 e CC_3236, que formam o operon *ruvCAB*, também tiveram expressão aumentada na cepa *lexA(Def)*, o que está de acordo com o estudo de Erill *et al.* (2004), que propõe que este operon faz parte do regulon SOS das proteobactérias-α. *imuA*, o primeiro gene de um operon de reparo mutagênico de DNA (Galhardo *et al.*, 2005), também mostrou níveis aumentados de expressão na cepa *lexA(Def)*. Outros genes potencialmente relacionados com metabolismo de DNA: CC_2272, CC_0382, CC_2332 e CC_1330 também estão entre os genes com expressão aumentada.

Curiosamente, *dnaB* (helicase replicativa) e CC_2433 (proteína hipotética conservada) tiveram a expressão diminuída na cepa *lexA(Def)*, o que sugere que LexA possa também funcionar como ativadora da transcrição. O fato de *dnaB*

ser reprimido como resposta SOS sugere que a célula possa exercer controle sobre a velocidade da replicação em função do nível de danos presentes no DNA.

Cabe lembrar, como autocrítica a este trabalho, que a análise estatística da expressão diferencial dos genes (Tabela 2.2) foi feita após a conclusão do estudo, de modo que foram considerados como pertencentes ao regulon SOS genes cuja variação não atingiu o nível estipulado de significância estatística (valor $p < 0.05$). No entanto, pelo fato destes genes incluírem o próprio *lexA* e *imuA*, que teve sua indução por LexA verificada por outros métodos experimentais (Galhardo *et al.*, 2005), acreditamos que a maior parte destes genes seja diferencialmente expressa entre as cepas selvagem e *lexA(Def)*. Para que a expressão diferencial destes genes seja confirmada com suporte estatístico, são necessárias medidas de expressão em mais réplicas biológicas.

Como pode ser observado na Fig. 2.3, o modelo obtido para os sítios de ligação de LexA concorda com o padrão GTTCN_7CTTC, previamente descrito como consenso para caixas SOS de proteobactérias-α (Fernandez de Henestrosa *et al.*, 1998). Este modelo também está de acordo com o modelo de matriz obtido por Erill *et al.* (2004). No entanto, o maior número de sequências usado no presente estudo nos permitiu detectar um grau parcial de conservação em posições adjacentes a cada bloco GTTC. Tais posições provavelmente contribuem na ligação DNA–LexA.

Com o objetivo de entender melhor a relação entre as caixas SOS e os níveis de alteração na expressão gênica, analisamos, para os genes da tabela 2.2, a correlação da expressão relativa entre as cepas *lexA(Def)* e selvagem com: (A) a distância da caixa SOS ao códon de início, e (B) a pontuação da caixa SOS. Como mostrado na Fig. 2.4, há uma forte correlação negativa entre a distância da caixa SOS ao códon de início e a expressão relativa. Não foi observada, entretanto, correlação entre a pontuação da caixa SOS e a expressão relativa entre as duas cepas.

Aplicamos esta análise de correlação também aos dados de Courcelle *et al.* (2001). Usando microarranjos de DNA, os autores compararam níveis de expressão gênica entre a cepa selvagem de *E. coli* e a cepa *lexA(Ind-)*, incapaz de induzir o sistema SOS, em bactérias irradiadas com luz UV. Como mostra a Fig. 2.5, há uma

correlação negativa moderada entre a distância da caixa SOS ao códon de início e a expressão relativa. Há também uma correlação positiva moderada entre a pontuação da caixa SOS e a expressão relativa.

A correlação entre as distâncias e as variações na expressão gênica sugerem que o posicionamento relativo entre a sequência operadora, e o sítio de início de tradução, e possivelmente o tamanho da região 5' não traduzida do gene, possam ter papel importante na regulação gênica. Uma análise de conservação destas distâncias em diferentes espécies pode contribuir para o melhor entendimento desta questão.

A não correlação observada entre pontuação da caixa SOS e variação na expressão gênica em *C. crescentus* pode significar que o modelo obtido não gera pontuações relacionadas com a força de ligação entre LexA e o DNA. Uma interpretação alternativa para isto pode ser que a força de ligação LexA–DNA não varie de forma gradual, mas sim abrupta, com modificações pontuais na sequência operadora. Esta questão pode ser elucidada em laboratório por meio de ensaios de mudança na mobilidade em gel de agarose para construções carregando sequências operadoras específicas.

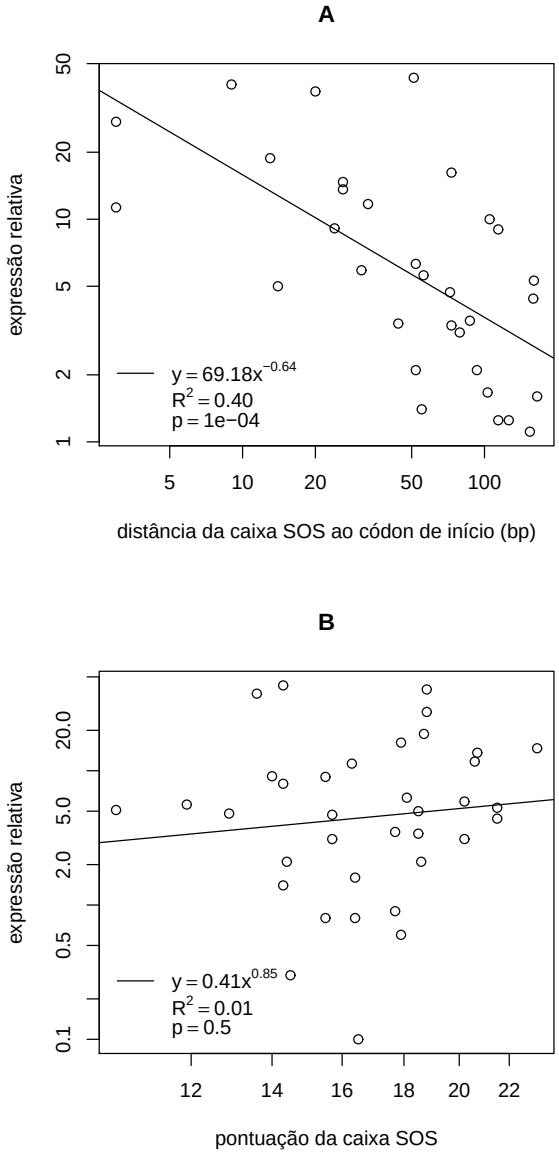

Figura 2.4: Análise de correlação da expressão relativa entre as cepas *lexA(Def)* e selvagem de *C. crescentus* com: (A) distância da caixa SOS ao códon de início; e (B) pontuação da caixa SOS, para os genes da tabela 2.2. Para os genes cotranscritos só são considerados aqueles que ocupam a primeira posição no respectivo operon. Em (A) só são considerados os genes com caixa SOS a jusante do códon de início.

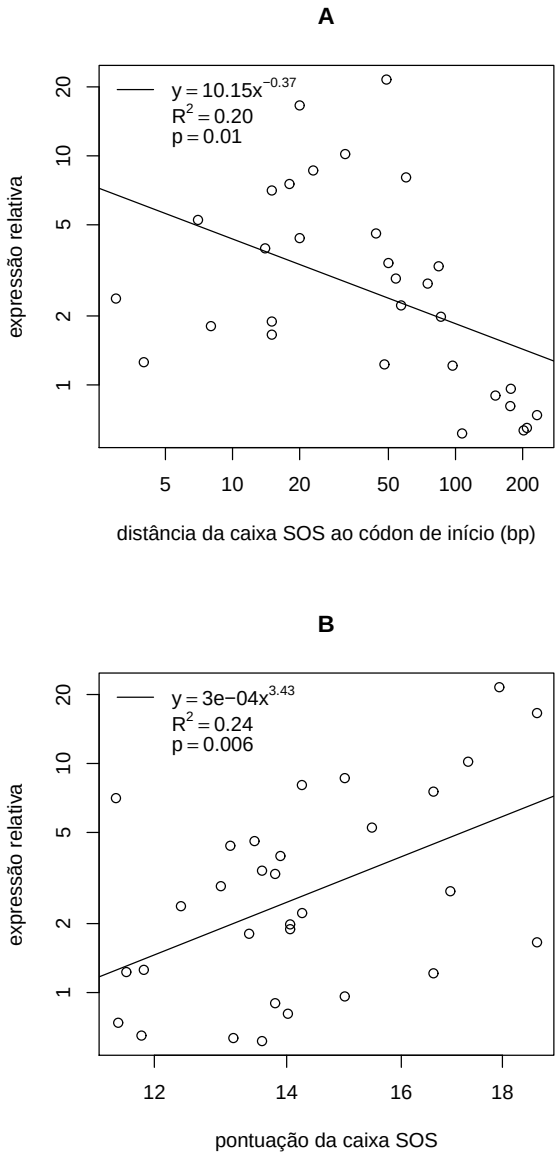

Figura 2.5: Análise de correlação da expressão relativa entre as cepas selvagem e *lexA(Ind-)* de *E. coli* após irradiação com luz UV com: (A) distância da caixa SOS ao códon de início; e (B) pontuação da caixa SOS. Dados de Courcelle *et al.* (2001).

2.5 Conclusões

Neste estudo aplicamos uma estratégia iterativa de predições *in silico* e análises de expressão gênica que permitiu a identificação de 37 genes pertencentes ao regulon SOS de *Caulobacter crescentus*, além de mostrar indícios de que proteína LexA também pode ter o papel de ativadora transcricional.

3 Um teste estatístico para a colocalização entre dois conjuntos de posições genômicas

3.1 Introdução

As regiões regulatórias na sequência de DNA de um gene, que incluem promotores e sítios de ligação de fatores de transcrição (SLFTs), especificam o programa de expressão deste gene em resposta a fatores ambientais e a sinais internos ao organismo. Compreender este programa com base na sequência de DNA é um dos objetivos da Biologia Computacional. A análise do posicionamento relativo entre promotores, SLFTs, sítios de início de transcrição e tradução, muitas vezes em conjunção com medidas de expressão gênica, é usada na elaboração de hipóteses sobre a regulação da transcrição dos genes em estudo.

Yu *et al.* (2006) desenvolveram um método computacional que faz predições de interações entre fatores de transcrição (FTs) de *Saccharomyces cerevisiae* com base em seus sítios de ligação no genoma. Um par de fatores de transcrição é anotado como de potencial interação se seus SLFTs coocorrem em um número de regiões promotoras significativamente mais alto que o esperado, caso os SLFTs fossem aleatoriamente distribuídos entre as regiões promotoras dos genes. Além disso, o método detecta distâncias preferenciais entre os pares de SLFTs, como indicador adicional da possível interação entre estes fatores de transcrição.

Bulyk *et al.* (2004) desenvolveram um método para identificar SLFTs em escala genômica, que parte do princípio de que a corregulação de genes por mais de um FT ocorre frequentemente. Tomando predições de SLFTs a partir de matrizes posicionais de probabilidade de nucleotídeos (ver seção 2.3, pág. 58) obtidas de ban-

cos de dados de SLFTs em *E. coli*, o método calcula, para cada par de FTs, uma lista de distâncias, em pares de bases, entre seus sítios de ligação. Em seguida, o método identifica, de um conjunto predefinido de faixas de distâncias, quais faixas que estão significativamente sobrerrepresentadas na lista, em comparação com o que seria esperado se os SLFTs fossem distribuídos ao acaso. Os autores então selecionaram alguns dos SLFTs cujas distâncias estão nas faixas mais sobrerrepresentadas para testes experimentais, verificando a expressão diferencial dos genes correspondentes entre a cepa selvagem e cepas que têm esses SLFTs mutados.

O trabalho de Bulyk *et al.* (2004) mostra o potencial da análise de co-ocorrência de SLFTs preditos na identificação de novos SLFTs. Seu método computacional, no entanto, baseia-se na análise de um conjunto predefinido de faixas de distâncias, o que pode implicar na perda de generalidade ou na necessidade de adaptação do algoritmo para aplicação a outros genomas.

3.2 Objetivos

O presente trabalho teve por objetivo o desenvolvimento de um método simples e de baixo custo computacional que permite testar se um conjunto Y de localizações genômicas está significativamente posicionado em relação a um conjunto de referência, X, de localizações genômicas, sem a necessidade do uso de faixas predefinidas de distâncias. Além disso, o método permite que se use uma função de densidade de probabilidade arbitrária para as posições esperadas do conjunto Y; e permite lidar com a situação em que apenas parte dos elementos do conjunto Y estejam significativamente posicionados em relação a X, identificando quais são estes elementos.

3.3 Modelo estatístico

As entradas para o teste são os conjuntos de posições genômicas X e Y, e uma função de densidade de probabilidade $f(.)$, que modela a distribuição teórica

dos elementos do conjunto Y ao longo do genoma. Os pontos de X e Y podem estar distribuídos em múltiplos cromossomos, lineares ou circulares. Porém, para maior facilidade na exposição, consideramos o caso em que há um cromossomo somente. Desta forma, os conjuntos X e Y são:

$$X = \{x_1, x_2, ..., x_m\}$$

$$Y = \{y_1, y_2, ..., y_n\}$$

onde x_i e y_i são posições em pares de bases.

A distância $d(x, y)$ entre duas posições genômicas x e y é definida simplesmente pela distância em pares de bases entre x e y. Ou seja, $d(x, y) = |x-y|$ se x e y estão em um cromossomo linear; $d(x, y) = \min(|x-y|, |x-y-1|, |x-y+1|)$ se estão em um cromossomo circular. A escolha de $d(x, y) = \infty$ para x e y em cromossomos diferentes generaliza o modelo para múltiplos cromossomos.

A distância entre a posição y e o conjunto de posições X é definida por:

$$d_X(y) = \min_{i \in \{1,2,...,m\}} d(y, x_i)$$

e a distância entre os conjuntos X e Y é definida por:

$$d_X(Y) = \max_{i \in \{1,2,...,n\}} d_X(y_i)$$

Sendo Y' uma permutação dos elementos de Y de modo que $d_X(y'_1) \leq d_X(y'_2) \leq ... \leq d_X(y'_n)$, definimos a k-ésima distância de X a Y como:

$$d_{X(k)}(Y) = d_X(y'_k) \tag{3.1}$$

Destas definições, segue que a fração do intervalo unitário composta por pontos situados a uma distância menor ou igual a d de algum ponto de X é $\int_{d_X(t) \leq d \, , \, t \in [0,1]} dt$. Com isto, podemos definir:

$$T_X(d) = \int_{d_X(t) \leq d \, , \, t \in [0,1]} f(t)dt \tag{3.2}$$

Como visto, $f(.)$ é uma função de densidade de probabilidade escolhida pelo usuário para modelar a distribuição teórica dos elementos do conjunto Y. A

escolha mais simples para $f(.)$ é distribuição uniforme. Esta função, porém, pode ser escolhida de modo que se levem em conta parâmetros biológicos que possam afetar o posicionamento dos elementos de Y, como, por exemplo, o conteúdo G+C local. A função $f(.)$ também pode ser escolhida de modo a restringir a análise a um conjunto de regiões de interesse. A figura 3.1 ilustra as definições apresentadas com um exemplo hipotético.

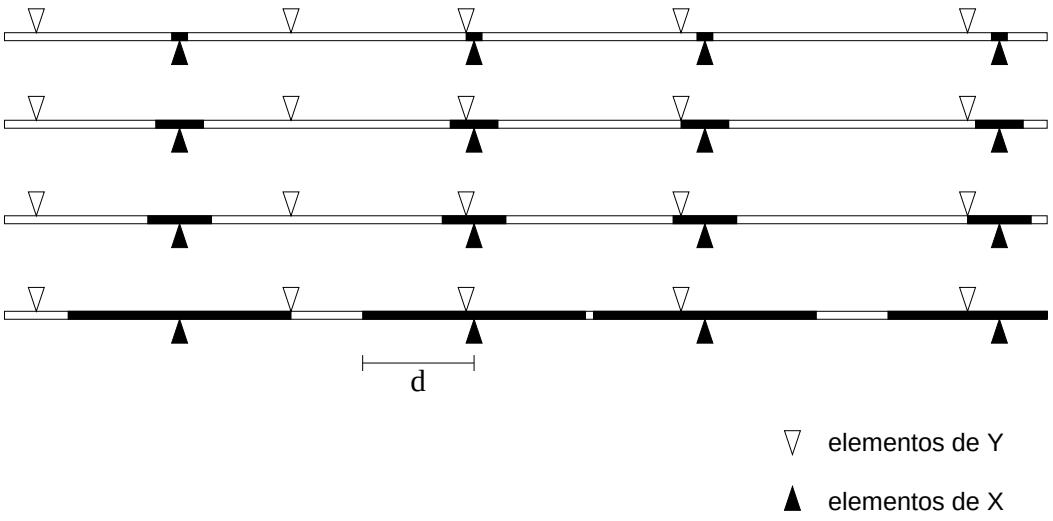

Figura 3.1 - Ilustração do modelo estatístico usando dois conjuntos hipotéticos de posições genômicas mapeadas no intervalo unitário. Regiões marcadas em preto contêm todos os pontos a uma distância menor ou igual a d de algum ponto de X. Os valores de d, de cima para baixo, são $d_{X(k)}(Y)$, para $k = 1, 2, 3, 4$, respectivamente. Para uma $f(.)$ uniforme, a fração do genoma marcada em preto é $T_X(d)$.

A proposição 1, a seguir, provê a base para a derivação da significância estatística deste teste. Esta proposição é simplesmente uma forma reescrita, na qual muda-se a ordem de integração de $f(.)$, da seguinte propriedade das funções de densidade de probabilidade: se Z é uma variável aleatória distribuída de acordo

com $f(.)$, e $F(.)$ a função de probabilidade acumulada, dada por $F(t) = \int_{-\infty}^{t} f(x)dx$, então $F(Z)$ tem distribuição uniforme no intervalo $[0, 1]$.

Proposição 1 *Se Z é uma variável aleatória distribuída de acordo com $f(.)$, então $T_X(d_X(Z)) \sim U(0, 1)$, isto é, que $T_X(d_X(Z))$ é distribuída uniformemente no intervalo $[0, 1]$.*

Prova 1 *Assuma que $Z \sim f(.)$. Das definições de $T_X(.)$ e $d_X(.)$, segue que:*

$$p(d_X(Z) \leq \delta) = T_X(\delta), \quad \forall \delta \in [0, \delta_{max}]$$

onde δ_{max} é o menor valor de distância tal que $T_X(\delta_{max}) = 1$.

Supondo-se, por hora, que $f(.)$ seja não-nula em todo o seu domínio, temos que $T_X(\delta)$ é estritamente crescente no intervalo $[0, \delta_{max}]$, o que implica que $T_X(a) \leq T_X(b)$ se e somente se $a \leq b$, para qualquer a e $b \in [0, \delta_{max}]$. Isto, por sua vez, implica:

$$p(d_X(Z) \leq \delta) = p(T_X(d_X(Z)) \leq T_X(\delta)) = T_X(\delta) \tag{3.3}$$

para todo $\delta \in [0, \delta_{max}]$. A última igualdade e o fato de que a imagem de $T_X(\delta))$, para $\delta \in [0, \delta_{max}]$, é o intervalo $[0, 1]$ implicam:

$$p(T_X(d_X(Z)) \leq \gamma) = \gamma, \quad \forall \gamma \in [0, 1]$$

que é a forma cumulativa da distribuição uniforme $U(0, 1)$ para $T_X(d_X(Z))$.

A existência de intervalos em que $f(t) = 0$ pode implicar em intervalos em que $T_X(\delta)$ é constante, o que invalidaria a afirmação de que $T_X(a) \leq T_X(b) \Leftrightarrow a \leq b, \quad \forall a, b \in [0, \delta_{max}]$. Entretanto, pelo próprio fato de que $f(t) = 0$ nestes intervalos, nenhuma realização de Z pode corresponder a um ponto $d_X(Z)$ em um destes intervalos em que $T_X(\delta)$ é constante, o que faz com que a Eq. 3.3 continue válida, completando a prova. ■

Se $\mathbf{Z} = (Z_1, Z_2, ..., Z_n)$ é uma amostra aleatória (com n valores independentes) obtida da distribuição $f(.)$, temos, pela proposição 1, que $T_X(d_{X(k)}(\mathbf{Z}))$ é a

k-ésima estatística de ordem de uma amostra aleatória de tamanho n da distribuição uniforme $U(0,1)$. Esta estatística de ordem tem, conhecidamente, uma distribuição Beta:

$$T_X(d_{X(k)}(\mathbf{Z})) \sim Beta(k, n+1-k) \tag{3.4}$$

Para conveniência de apresentação, chamamos a estatística $T_X(d_{X(k)}(Y))$, calculada a partir do conjunto de dados Y, simplesmente de λ_k. Esta estatística é uma medida do quão próximos de X estão os k elementos de Y mais próximos de X. O valor p é dado por:

$$p_k = p(T_X(d_{X(k)}(Z)) < \lambda_k) = I_{\lambda_k}(k, n+1-k) \tag{3.5}$$

onde $I_{\lambda_k}(k, n+1-k)$ é a distribuição cumulativa Beta no ponto λ_k.

Este valor p pode ser interpretado da seguinte maneira: seja Z um conjunto de pontos sorteados ao acaso, distribuídos de acordo com $f(.)$. O valor p_k é a probabilidade da k-ésima distância entre X e Z ser menor que a k-ésima distância entre X e Y.

Em algumas situações, alguns dos elementos de Y estão significativamente posicionados em relação a X enquanto que os elementos restantes distribuem-se independentemente de X ao longo do genoma. Nestes casos, pode ser de interesse biológico identificar quais desses pontos estão significativamente posicionados em relação a X. A aplicação direta da Eq. 3.5 não é adequada para isto pois, em geral, a significância estatística do k-ésimo ponto tem forte dependência dos pontos anteriores. A fim de evitar este problema, definimos uma nova estatística de teste, que considera apenas uma janela com w pontos de Y (o parâmetro w é definido pelo usuário), ignorando os pontos anteriores. A nova estatística, dada por:

$$\lambda_{k,w} = \frac{T_X(d_{X(k)}(Y)) - T_X(d_{X(k-w)}(Y))}{1 - T_X(d_{X(k-w)}(Y))} \tag{3.6}$$

é definida para $k \geq w$, assumindo-se $d_{X(0)}(Y) = 0$. O valor p (a prova disto foi omitida, mas é similar à da proposição 1) é dado por:

$$q_{k,w} = I_{\lambda_{k,w}}(w, (n-k+1)+1-w)$$

A escolha do tamanho da janela w deve levar em conta o número de pontos disponíveis e o grau de separação que se quer ter entre os valores de $q_{k,w}$: quanto maior o valor de w, maior a discriminação entre os elementos que estão significativamente posicionados próximos a X dos que não estão, porém mais suave a transição da significância $q_{k,w}$ entre os primeiros e os segundos.

3.4 Implementação

O teste foi implementado como um módulo para o ambiente estatístico R e está livremente disponível sob a licença GNU GPL versão 3, no endereço: `http://www.icb.usp.br/~mutagene/apua/co2.html`. O teste também pode ser usado via *web*, neste mesmo endereço. Estão disponíveis uma interface simplificada, que permite a análise de apenas uma molécula de DNA e usa $f(.)$ uniforme, e uma interface detalhada, que permite a análise de múltiplas moléculas de DNA e a escolha pelo usuário da função $f(.)$. Detalhes dos formatos de arquivo de entrada encontram-se na documentação, neste mesmo endereço.

As entradas para o teste são: (1) uma descrição de cada molécula de DNA do organismo em estudo, contendo: o tamanho em pares de bases e um campo binário indicando se a molécula é circular ou linear; (2) os nomes e as posições em pares de bases dos pontos de X; (3) os nomes e as posições em pares de bases dos pontos de Y; (4) a distribuição $f(.)$, especificada como uma função linear por partes; e (5) o tamanho da janela. A saída é uma tabela com as seguintes informações para cada ponto do conjunto Y: (1) o nome do ponto; (2) o nome da molécula de DNA à qual o ponto pertence; (3) a posição em pares de bases do ponto nesta molécula; (4) o nome do ponto de X mais próximo; (5) a distância até o ponto de X mais próximo; (6) o valor da estatística do teste original λ_k; (7) o valor p do teste original: p_k; (8) o valor da estatística do teste com janelas $\lambda_{k,w}$; e (9) o valor p do teste com janelas: $q_{k,w}$.

3.5 Resultados e discussão

3.5.1 Dados gerados por computador

Aplicamos o modelo, para fins de teste, em conjuntos de dados X e Y gerados artificialmente. O conjunto X tem 200 pontos distribuídos uniformemente ao longo de um genoma hipotético de 4 Mb. O conjunto Y tem 50 pontos localizados a uma distância de até 400 bp de algum ponto de X e 150 pontos distribuídos uniformemente no genoma. Este conjunto de dados é um exemplo em que apenas parte do conjunto Y tem posicionamento significativamente próximo de pontos de X. Estamos interessados, neste caso, em verificar se o modelo é capaz de reconhecer, com base na significância estatística, os 50 pontos localizados próximos de pontos de X. No gráfico A da Figura 3.2 são mostrados os valores das significâncias p_k para este conjunto de dados. No gráfico B, são mostrados os valores de p_k para o conjunto controle, em que os 200 pontos de Y são distribuídos uniformemente e de maneira independente de X. Observa-se, no gráfico A, que os valores de p_k decrescem até k próximo de 55 e depois voltam a crescer, atingindo, nos últimos pontos, valores similares aos do conjunto controle. No entanto, para $k = 100$, o valor de p é da ordem de 10^{-9}, o que indica que os 100 primeiros pontos de Y (em ordem de distância até um ponto de X) estão, como conjunto, significativamente próximos de pontos de X, embora só tenham contribuído para isto os seus 50 primeiros pontos. A fim de identificar quais os pontos de Y que realmente contribuem para o posicionamento do conjunto, aplicamos, aos mesmos dados, a estatística definida na Eq. 3.6. Para cada k, em vez de considerar o posicionamento do conjunto dos k primeiros pontos de Y, esta nova estatística considera apenas o posicionamento dos pontos de ordem $k - w + 1$ até k, ignorando os pontos anteriores e excluindo da análise as regiões do genoma com distância até $d_X(y_{k-w})$ de algum ponto de X. O tamanho da janela, w, como visto, é um parâmetro escolhido pelo usuário. Na Fig. 3.3, gráfico A, são mostrados os valores de significância $q_{k,w}$ para os mesmos dados da análise anterior e tamanho de janela $w = 15$. No gráfico B, são mostrados os valores de

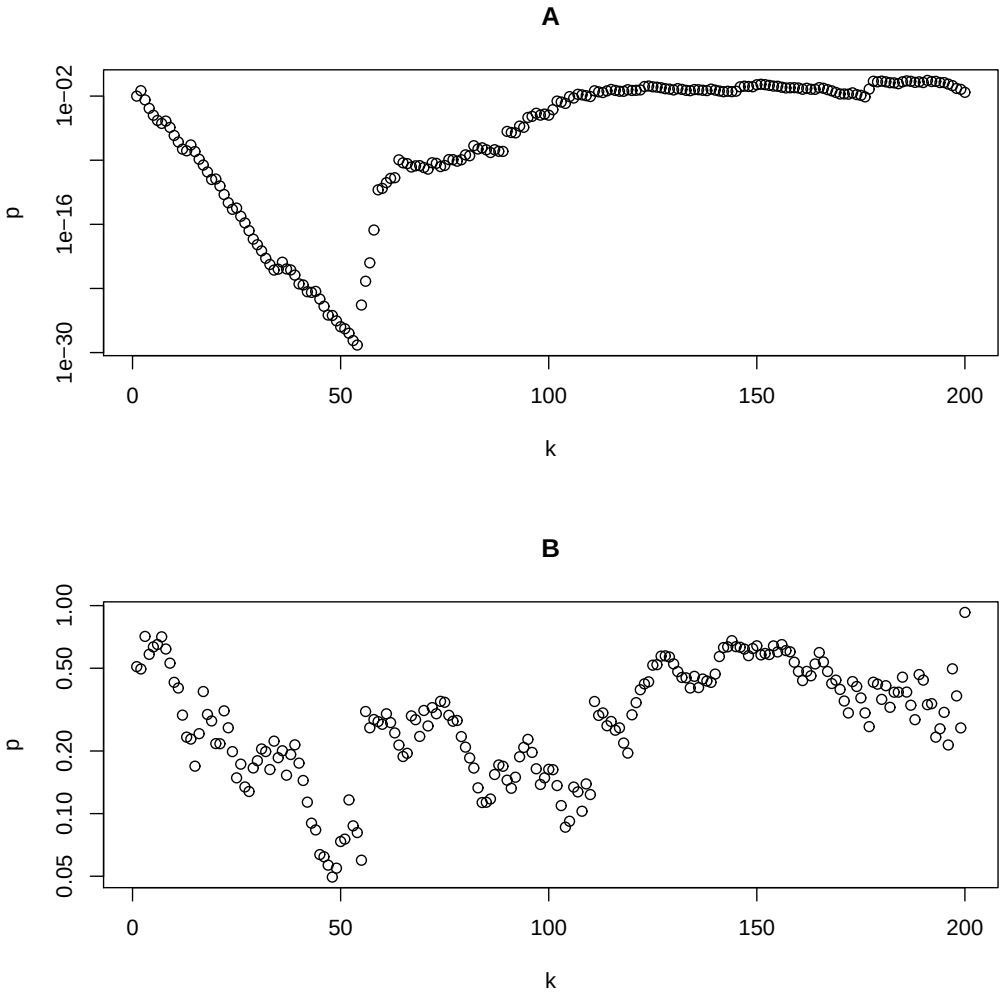

Figura 3.2 - Valores das significâncias p_k para dados gerados por computador. Em A e B, X tem 200 pontos uniformemente distribuídos em $[0, 1]$. Em A, Y tem 50 pontos a uma distância de até 10^{-4} dos primeiros 50 pontos de X e 150 pontos distribuídos uniformemente em $[0, 1]$. Em B, Y tem 200 pontos uniformemente distribuídos em $[0, 1]$. Para ambos os testes, usou-se $f(.)$ uniforme.

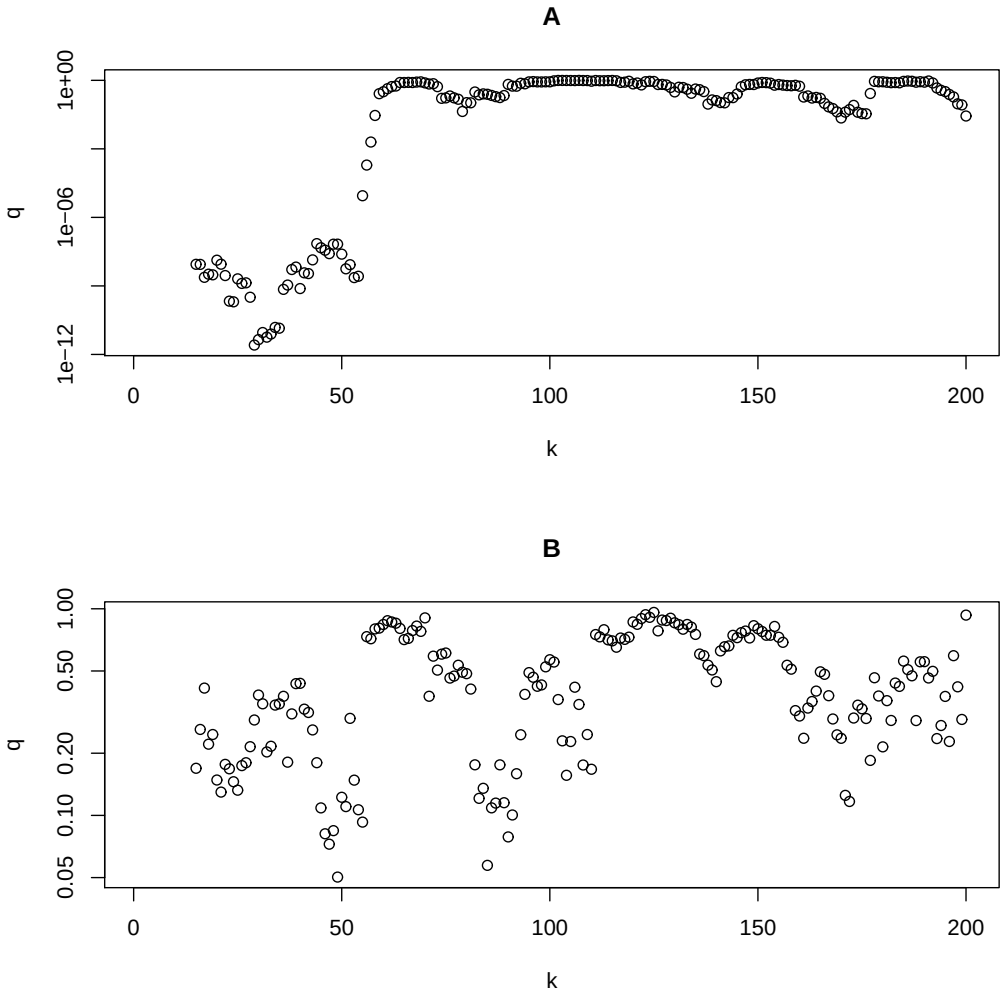

Figura 3.3 - Valores das significâncias $q_{k,w}$, com $w = 15$, para dados gerados por computador. Em A e B, X tem 200 pontos uniformemente distribuídos em $[0, 1]$. Em A, Y tem 50 pontos a uma distância de até 10^{-4} dos primeiros 50 pontos de X e 150 pontos distribuídos uniformemente em $[0, 1]$. Em B, Y tem 200 pontos uniformemente distribuídos em $[0, 1]$. Para ambos os testes, usou-se $f(.)$ uniforme.

$q_{k,w}$ para os mesmos dados do conjunto controle. Observa-se, no gráfico A, valores de $q_{k,w}$ inferiores a 10^{-6} para $k \leq 55$. Para $k > 65$, que corresponde a 50 mais o tamanho da janela, observa-se que $q_{k,w}$ atinge a mesma faixa de valores do conjunto controle, do gráfico B, mostrando uma clara separação entre duas classes de pontos: aqueles posicionados significativamente próximos de X e aqueles com posicionamento aleatório.

3.5.2 Sítios de ligação de LexA em *Caulobacter crescentus*

Aplicamos também o método aos dados da seção 2 (pág. 51), para analisar a significância estatística do posicionamento dos sítios preditos de ligação de LexA em relação aos sítios de início de tradução no genoma de *Caulobacter crescentus*. Nesta análise, o conjunto X é composto dos 3737 sítios de início de tradução deste genoma, de acordo com o banco de dados Omniome (Peterson *et al.*, 2001) e o conjunto Y é composto dos 103 sítios preditos de ligação de LexA com maior pontuação segundo o modelo na Tabela 2.3, representado também como *sequence logo* na Fig. 2.3. Este conjunto, além dos sítios verificados experimentalmente na seção 2, inclui também predições com pontuação mais baixa que o limiar efetivamente usado naquela análise. O objetivo desta escolha é inferir se estes sítios com pontuação mais baixa também estão significativamente posicionados em relação aos sítios de início de tradução, o que seria uma potencial indicação de que são capazes de se ligar com LexA.

Na Fig. 3.4, gráficos A e B, são mostrados, respectivamente, os valores das significâncias p_k e $q_{k,w}$, com tamanho de janela $w = 15$. Como pode ser observado no gráfico B, diferentemente dos dados gerados artificialmente, não houve uma transição abrupta nos valores de $q_{k,w}$ com o aumento de k, mas sim um crescimento gradual até o 76º ponto, após o qual $q_{k,w}$ atinge valores superiores a 10^{-2}, comparáveis aos do conjunto controle na análise com dados artificiais. A Fig. 3.5 mostra o valor de $q_{k,w}$ em função da distância que cada ponto tem do sítio de início de tradução mais próximo. A análise destes dados nos permite estimar, em função de um limiar de significância estatística, uma distância crítica, acima da qual não se encontram

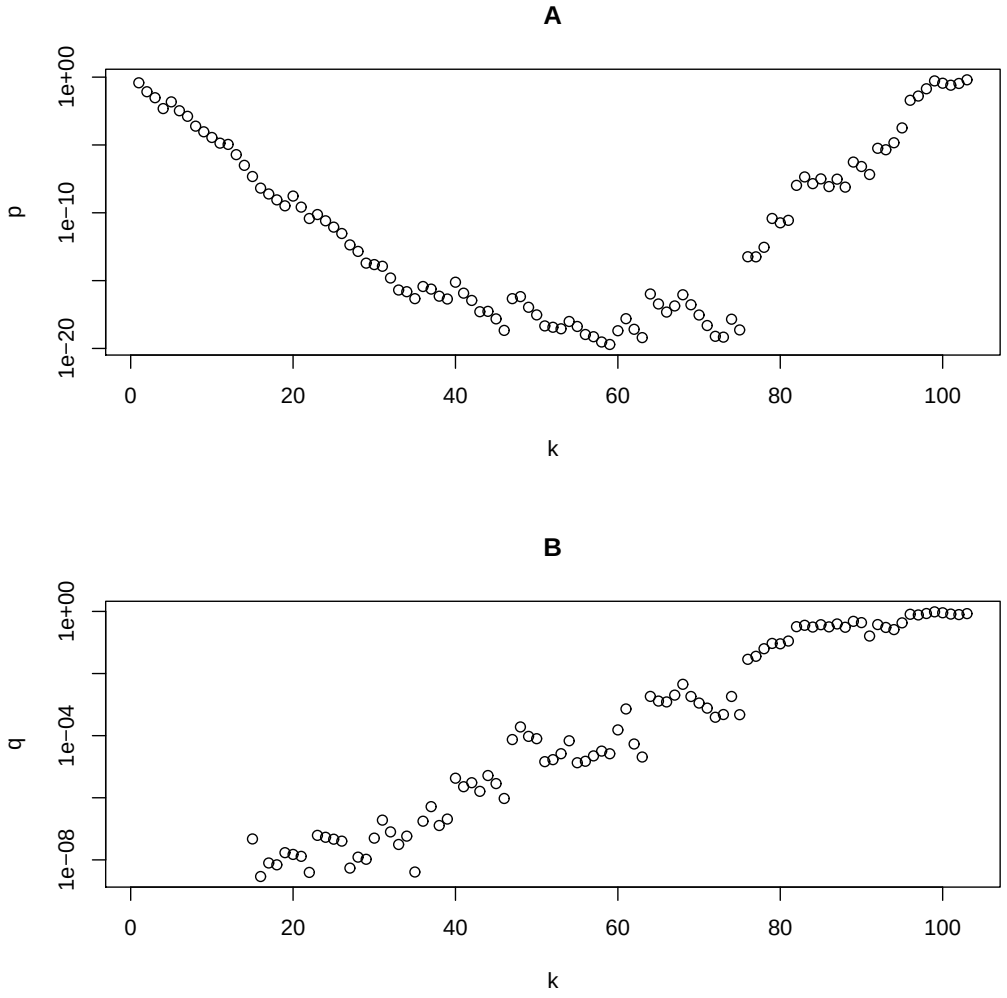

Figura 3.4 - Análise do posicionamento de sítios preditos de ligação de LexA em *Caulobacter crescentus*. O conjunto X corresponde aos sítios de início de tradução deste genoma e o conjunto Y corresponde a predições de sítios de ligação de LexA. São mostrados os valores das significâncias p_k (em A) $q_{k,w}$ (em B), para tamanho de janela $w = 15$ e $f(.)$ uniforme.

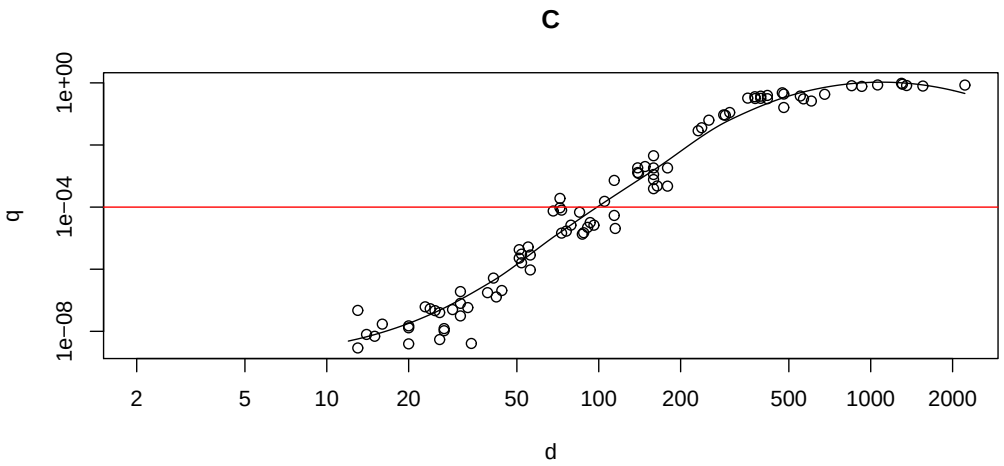

Figura 3.5 - Valores das significâncias $q_{k,w}$, para sítios preditos de ligação de LexA em *Caulobacter crescentus*, em função da distância d até o sítio de início de tradução mais próximo. Foram usados tamanho de janela $w = 15$ e $f(.)$ uniforme. A linha sólida corresponde a uma regressão local obtida com o método LOESS.

pontos do conjunto Y significativamente posicionados em relação a X. Neste caso, tomando-se o limiar 10^{-2} para $q_{k,w}$, temos, com base na curva de regressão local, uma estimativa da distância crítica em 213 bp, o que sugere que a ligação de LexA a uma distância acima desta de um sítio de início de tradução tenha pouca influência sobre a expressão do gene correspondente. Também com base nestes dados, tomando-se um limiar conservador de 10^{-4} para $q_{k,w}$, a distância crítica estimada é de 99 bp, o que indica que 59 potenciais sítios de LexA estão posicionados significativamente próximos de sítios de início de tradução. A tabela 3.1 mostra quais são estes sítios. Dos 54 genes correspondentes a estes sítios, 29 não foram estudados na seção 2, pelo fato das caixas SOS terem pontuação inferior ao valor de corte efetivamente usado naquele estudo ou, em alguns casos, por problemas experimentais. O posicionamento destes sítios sugere que a ligação de LexA possa ter também influência na regulação da expressão dos genes correspondentes.

Tabela 3.1: Sítios preditos de ligação de LexA de *Caulobacter crescentus* posicionados significativamente próximos de sítios preditos de início de tradução. P: pontuação da caixa LexA (ver seção 2.3); D: posição relativa (em nucleotídeos) da caixa LexA predita em relação o códon de início predito; q: valor da significância estatística q, com $w = 15$; T: indica se o gene teve a expressão relativa medida no estudo da seção 2.

Gene	Função	Caixa SOS	P	D	q	T
CC_3164	Transcriptional regulator, Cro/CI family	AACGGTCGTGTCATTTTCC	8.91	2	3.1e-07	Não
CC_3238	*ruvC*/Resolvase da junção de Holliday	CGCGTTCATCATGTGTTCT	10.4	2	3.1e-07	Sim
CC_1054	Proteína hipotética	TTTGTTCTCGGCTTGTTCT	16.3	-3	3.1e-07	Sim
CC_3518	Proteína hipotética conservada	GATGTTCATGTATTGTTCT	18.76	-3	3.1e-07	Sim
CC_1927	Proteína hipotética	CATATTCCGGTTTTGTTCT	16.37	-7	3.1e-07	Sim
		TCTGTTCACAAGATGTTCC	11.38	-23	6.2e-08	
		AGATTTCTTGTTTTGTTCC	11.07	11	3.1e-07	
CC_3370	Proteína hipotética conservada	CATGTTCCTGTTCCTTTTC	10.56	-7	3.1e-07	Não
CC_1543	*mreB*/rod shape-determining protein MreB	AATGTTCTCTTCCCTTTTC	9.16	8	3.1e-07	Não
CC_3225	Sensory box sensor histidine kinase/response regulator	TTTGTTCGCCAGATTTTTT	12.88	8	3.1e-07	Sim
CC_0627	Proteína hipotética	AAAGTTCGCGTTATGTTCT	18.77	-9	3.1e-07	Sim
CC_1528	*uvrD*/DNA helicase II	TTTGTGCTACATATGTTCC	11.87	-10	3.1e-07	Não
CC_0392	Proteína de membrana, putativa	GATGTTCCCCCCTTTTTTT	9.12	-13	3.1e-07	Não

Tabela 3.1: continuação.

Gene	Função	Caixa SOS	P	D	q	T
CC_0510	*phbA*/acetil-CoA acetiltransferase	ATAGCTCCCGTGACGTTCT	8.72	-13	3.1e-07	Não
CC_0575	Beta-lactamase, putativa	AATCTTCCAGCCATGTTTC	9.89	-13	3.1e-07	Não
CC_1993	Proteína hipotética conservada	TGCGTTCGCCTTATCTTTT	8.73	-13	4.7e-08	Não
CC_2332	Proteína hipotética conservada	ATCGTTCTTGATTTGTTCT	18.7	-13	2.9e-09	Sim
CC_0383	Proteína hipotética	TATGTTCCTGAAAAGTTCT	18.52	-14	7.9e-09	Sim
CC_3109	Proteína hipotética conservada	CGTGTTCCCCTCTTGTTCG	11.48	-15	6.9e-09	Não
CC_3620	*tktA*/Transketolase I	TTCCGTCTTTATATGTTTT	10.56	-16	1.7e-08	Não
CC_1531	Proteína hipotética	ATTGTTCTTGATATGTTCC	20.22	-31	8.0e-08	Sim
		TATGTTCCAACTTCGTTTG	11.34	-20	1.5e-08	
CC_2469	Proteína hipotética	TTTGTTCTTGTTTCGTCCA	11.4	-20	1.3e-08	Não
CC_3467	Proteína hipotética conservada	CATGTTCCAGCTTTGTTCG	13.65	-20	4.0e-09	Sim
CC_2040	RNA helicase dependente de ATP da família DEAD/DEAH	CATGTTCCCTTTCTGTTTC	14.02	-24	5.3e-08	Sim
CC_3380	*kduD*/2-deoxy-D-gluconate 3-dehydrogenase	ATTGTTCTTGAAACGCTTG	10.91	-25	4.6e-08	Não
CC_2272	Proteína da família da endonuclease III	AATGTTCTTGTTATGTTCT	23.17	-26	4.0e-08	Sim
CC_3424	Proteína hipotética conservada	AATGTTCCTGAATTGTTCT	20.74	-26	5.4e-09	Sim
CC_2111	Proteína hipotética conservada	AACGTTCTCTGGTTCTTCT	8.96	27	1.2e-08	Não

Tabela 3.1: continuação.

Gene	Função	Caixa SOS	P	D	q	T
CC_2589	Proteína hipotética	TTTGTTCGCATCTTGTTCT	17.67	-154	1.3e-05	Sim
		CTTGTTCTCGCGACGTTCG	10.61	27	1.0e-08	
CC_0041	Proteína hipotética	AATGATCCCGTTTACTTCC	9.2	29	5.0e-08	Não
CC_0692	Proteína hipotética	TTTGTTGAGAAATTGTTGT	9.26	-31	1.9e-07	Não
CC_2498	Proteína hipotética	AATTTTCGTGATTTTTTCG	11.66	-31	3.1e-08	Não
CC_1330	Proteína com domínio Radical SAM	TATGTTCTTGTTATGTTCG	20.59	-33	5.8e-08	Sim
CC_0782	Regulador transcricional da família LuxR, putativo	AGCGTTCGCATAAGGTTCT	11.27	34	4.1e-09	Não
CC_2017	Proteína hipotética	TTTCGTGCTTTTATGTTCT	10.6	39	1.8e-07	Não
CC_3130	Proteína da família da sintetase de glutamina	TTTGTTCTCGAAAGGTTTC	14.39	-52	3.1e-06	Sim
		GTTTTTCCGGATTTGTTCT	11.75	-41	5.2e-07	
CC_2025	Proteína hipotética	AATATTGACTTTATATTTT	10.23	-42	1.3e-07	Não
CC_0420	Proteína hipotética	AATAGTCCAGAAATTTTCG	9.19	44	2.1e-07	Não
CC_1836	Proteína hipotética conservada	AATATTTGAGTCATGTTCC	12.27	-51	4.3e-06	Não
CC_1902	*lexA*	AATGTTCTCCTGGTGTTCC	14.31	-51	2.3e-06	Sim
CC_3356	Proteína hipotética	CATGTTCTCGTATTGTTCG	18.08	-52	1.6e-06	Sim
CC_0686	*groES*/Chaperonina, 10 kDa	CTTGGTCTTGGTTTCTTCT	9.82	56	2.9e-06	Não

Tabela 3.1: continuação.

Gene	Função	Caixa SOS	P	D	q	T
CC_1983	*recN*	GATGATCCCGTTTCGTTCC	11.89	-56	9.5e-07	Sim
CC_0085	*pgm*/phosphoglucomutase	ATAGTTCGTGAAATTGTCT	9.84	-68	7.5e-05	Não
CC_0140	ComM protein	AACGTTCGTTTTTCGTTCT	15.71	-72	1.9e-04	Sim
CC_2630	*hfaD*/hfaD protein, authentic frameshift	TTCGTTCCCAGATTGTTGT	9.56	-72	9.5e-05	Não
CC_1665	*dnaB*/DNA helicase replicativa	GATGTTCTGTGTATGTTTT	14.52	-73	8.0e-05	Sim
CC_3213	Proteína hipotética	CATGTTCCACTTTTGTTCT	17.89	-73	1.5e-05	Sim
CC_0550	Proteína hipotética	ATTATTCCAGACACATTCT	9.01	-76	1.7e-05	Não
CC_3515	Proteína hipotética conservada	AGAGTTCGCATTATGTTCT	15.74	-79	2.6e-05	Sim
CC_3284	*purE*/phosphoribosylaminoimidazole carboxylase, catalytic subunit	AACGCTCTTGTGTCGTTTT	9.43	-85	6.8e-05	Não
CC_2590	*uvrA*/excinuclease ABC, subunidade A	TTTGTTCGCATCTTGTTCT	17.67	-87	1.3e-05	Sim
CC_0975	Proteína hipotética	AATGTGCCAGTTGCGTTCT	10.61	88	1.5e-05	Não
CC_0781	Proteína hipotética	TTTGTTCCCCTAGAGTCCT	10.28	-91	2.2e-05	Não
CC_1468	*ssb*/Proteína de ligação ao DNA de fita simples	TTTGTTCTCATAACGTTCT	18.6	-93	3.2e-05	Sim
CC_3514	Proteína hipotética	TTTGTTCCCGGTCCGTTCG	10.07	-96	2.6e-05	Não

3.6 Conclusões

Neste trabalho, desenvolvemos um método simples e de baixo custo computacional para avaliar se as posições genômicas de um conjunto Y estão significativamente posicionadas em relação às posições genômicas de um conjunto X. Com base no exposto, podemos chegar às seguintes conclusões:

- o método não depende de intervalos pré-fixados de distâncias.
- uma função arbitrária (na implementação, uma função linear por partes) pode ser usada para modelar a distribuição esperada dos pontos do conjunto Y.
- a aplicação do método em dados gerados por computador obteve a separação entre os elementos de Y posicionados artificialmente próximos de X e aqueles posicionados de maneira aleatória.
- a aplicação do método aos dados de caixas SOS da seção 2 indica que pelo menos 29 genes que não foram analisados naquele estudo têm potenciais sítios de LexA em posições significativamente próximas aos seus sítios de início de tradução, o que sugere estes genes possam ser regulados por LexA.

Referências Bibliográficas

ALTSCHUL, S. F.; MADDEN, T. L.; SCHÄFFER, A. A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Res.**, v.25, n.17, p.3389–402, 1997.

ARIFUZZAMAN, M.; MAEDA, M.; ITOH, A.; NISHIKATA, K.; TAKITA, C.; SAITO, R.; ARA, T.; NAKAHIGASHI, K.; HUANG, H.-C.; HIRAI, A.; TSUZUKI, K.; NAKAMURA, S.; ALTAF-UL-AMIN, M.; OSHIMA, T.; BABA, T.; YAMAMOTO, N.; KAWAMURA, T.; IOKA-NAKAMICHI, T.; KITAGAWA, M.; TOMITA, M.; KANAYA, S.; WADA, C.; MORI, H. Large-scale identification of protein-protein interaction of Escherichia coli K-12. **Genome Res.**, v.16, n.5, p.686–91, 2006.

BEIKO, R. G.; HARLOW, T. J.; RAGAN, M. A. Highways of gene sharing in prokaryotes. **Proc. Natl. Acad. Sci. U.S.A.**, v.102, n.40, p.14332–7, 2005.

BERG, O. G.; VON HIPPEL, P. H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. **J Mol Biol**, v.193, n.4, p.723–50, 1987.

BROWN, P. J. B.; HARDY, G. G.; TRIMBLE, M. J.; BRUN, Y. V. Complex regulatory pathways coordinate cell-cycle progression and development in Caulobacter crescentus. **Adv. Microb. Physiol.**, v.54, p.1–101, 2009.

BULYK, M. L.; MCGUIRE, A. M.; MASUDA, N.; CHURCH, G. M. A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in Escherichia coli. **Genome Res**, v.14, n.2, p.201–8, 2004.

BUTLAND, G.; PEREGRÍN-ALVAREZ, J. M.; LI, J.; YANG, W.; YANG, X.; CANADIEN, V.; STAROSTINE, A.; RICHARDS, D.; BEATTIE, B.; KROGAN, N.; DAVEY, M.; PARKINSON, J.; GREENBLATT, J.; EMILI, A. Interaction network containing conserved and essential protein complexes in Escherichia coli. **Nature**, v.433, n.7025, p.531–7, 2005.

CHEN, I.; DUBNAU, D. DNA uptake during bacterial transformation. **Nat. Rev. Microbiol.**, v.2, n.3, p.241–9, 2004.

CHOI, I.-G.; KIM, S.-H. Global extent of horizontal gene transfer. **Proc. Natl. Acad. Sci. U.S.A.**, v.104, n.11, p.4489–94, 2007.

CLAVERYS, J.-P.; PRUDHOMME, M.; MARTIN, B. Induction of competence regulons as a general response to stress in gram-positive bacteria. **Annu. Rev. Microbiol.**, v.60, p.451–75, 2006.

COOK, R. D. Influential Observations in Linear Regression. **Journal of the American Statistical Association**, v.74, p.169–174, 1979.

COURCELLE, J.; KHODURSKY, A.; PETER, B.; BROWN, P. O.; HANAWALT, P. C. Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient Escherichia coli. **Genetics**, v.158, n.1, p.41–64, 2001.

DELSUC, F.; BRINKMANN, H.; PHILIPPE, H. Phylogenomics and the reconstruction of the tree of life. **Nat. Rev. Genet.**, v.6, n.5, p.361–75, 2005.

DUBNAU, D. DNA uptake in bacteria. **Annu. Rev. Microbiol.**, v.53, p.217–44, 1999.

ERILL, I.; JARA, M.; SALVADOR, N.; ESCRIBANO, M.; CAMPOY, S.; BARBE, J. Differences in LexA regulon structure among Proteobacteria through in vivo assisted comparative genomics. **Nucleic Acids Res**, v.32, n.22, p.6617–26, 2004.

FELSENSTEIN, J.; CHURCHILL, G. A. A Hidden Markov Model approach to variation among sites in rate of evolution. **Mol. Biol. Evol.**, v.13, n.1, p.93–104, 1996.

FOX, G. E.; STACKEBRANDT, E.; HESPELL, R. B.; GIBSON, J.; MANILOFF, J.; DYER, T. A.; WOLFE, R. S.; BALCH, W. E.; TANNER, R. S.; MAGRUM, L. J.; ZABLEN, L. B.; BLAKEMORE, R.; GUPTA, R.; BONEN, L.; LEWIS, B. J.; STAHL, D. A.; LUEHRSEN, K. R.; CHEN, K. N.; WOESE, C. R. The phylogeny of prokaryotes. **Science**, v.209, n.4455, p.457–63, 1980.

FRIEDBERG, E. C.; , W. G. C.; , S. W.; , W. R. F.; , S. R. A.; , E. T. **DNA Repair And Mutagenesis, 2nd edition**. ASM Press, Washington, DC, 2006.

GAL-MOR, O.; FINLAY, B. B. Pathogenicity islands: a molecular toolbox for bacterial virulence. **Cell. Microbiol.**, v.8, n.11, p.1707–19, 2006.

GALHARDO, R. S.; ROCHA, R. P.; MARQUES, M. V.; MENCK, C. F. An SOS-regulated operon involved in damage-inducible mutagenesis in Caulobacter crescentus. **Nucleic Acids Res**, v.33, n.8, p.2603–14, 2005.

GEORGI, B.; SCHLIEP, A. Context-specific independence mixture modeling for positional weight matrices. **Bioinformatics**, v.22, n.14, p.e166–73, 2006.

GRIFFITHS, A. J. F.; WESSLER, S. R.; LEWONTIN, R. C.; GELBART, W. M.; SUZUKI, D. T.; MILLER, J. H. **An Introduction to Genetic Analysis**. 8 ed. New York: W. H. Freeman, 2004.

HACKER, J.; KAPER, J. B. Pathogenicity islands and the evolution of microbes. **Annu. Rev. Microbiol.**, v.54, p.641–79, 2000.

HAYES, F. Toxins-antitoxins: plasmid maintenance, programmed cell death, and cell cycle arrest. **Science**, v.301, n.5639, p.1496–9, 2003.

FERNANDEZ DE HENESTROSA, A. R.; OGI, T.; AOYAGI, S.; CHAFIN, D.; HAYES, J. J.; OHMORI, H.; WOODGATE, R. Identification of additional genes

belonging to the LexA regulon in Escherichia coli. **Mol Microbiol**, v.35, n.6, p.1560–72, 2000.

FERNANDEZ DE HENESTROSA, A. R.; RIVERA, E.; TAPIAS, A.; BARBE, J. Identification of the Rhodobacter sphaeroides SOS box. **Mol Microbiol**, v.28, n.5, p.991–1003, 1998.

HOPE, A. C. A. A Simplified Monte Carlo Significance Test Procedure. **Journal of the Royal Statistical Society. Series B (Methodological)**, v.30, n.3, p.582–598, 1968.

HSIAO, W. W. L.; UNG, K.; AESCHLIMAN, D.; BRYAN, J.; FINLAY, B. B.; BRINKMAN, F. S. L. Evidence of a large novel gene pool associated with prokaryotic genomic islands. **PLoS Genet.**, v.1, n.5, p.e62, 2005.

JAIN, R.; RIVERA, M. C.; LAKE, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. **Proc. Natl. Acad. Sci. U.S.A.**, v.96, n.7, p.3801–6, 1999.

JAIN, R.; RIVERA, M. C.; MOORE, J. E.; LAKE, J. A. Horizontal gene transfer in microbial genome evolution. **Theoretical population biology**, v.61, n.4, p.489–95, 2002.

JAIN, R.; RIVERA, M. C.; MOORE, J. E.; LAKE, J. A. Horizontal gene transfer accelerates genome innovation and evolution. **Mol. Biol. Evol.**, v.20, n.10, p.1598–602, 2003.

KANHERE, A.; VINGRON, M. Horizontal Gene Transfers in prokaryotes show differential preferences for metabolic and translational genes. **BMC Evol. Biol.**, v.9, n.1, p.9, 2009.

KARLIN, S. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. **Trends Microbiol.**, v.9, n.7, p.335–43, 2001.

KARLIN, S.; MRÁZEK, J.; CAMPBELL, A. M. Codon usages in different gene classes of the Escherichia coli genome. **Mol. Microbiol.**, v.29, n.6, p.1341–55, 1998.

KENYON, C. J.; WALKER, G. C. DNA-damaging agents stimulate gene expression at specific loci in Escherichia coli. **Proc. Natl. Acad. Sci. U.S.A.**, v.77, n.5, p.2819–23, 1980.

KOBAYASHI, I. Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. **Nucleic Acids Res.**, v.29, n.18, p.3742–56, 2001.

KOONIN, E. V.; WOLF, Y. I. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. **Nucleic Acids Res.**, v.36, n.21, p.6688–719, 2008.

LASSMANN, T.; FRINGS, O.; SONNHAMMER, E. L. L. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. **Nucleic Acids Res.**, v.37, n.3, p.858–65, 2009.

LAWRENCE, J. G. Gene transfer in bacteria: speciation without species? **Theoretical population biology**, v.61, n.4, p.449–60, 2002.

LAWRENCE, J. G.; OCHMAN, H. Amelioration of bacterial genomes: rates of change and exchange. **J. Mol. Evol.**, v.44, n.4, p.383–97, 1997.

LAWRENCE, J. G.; RETCHLESS, A. C. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. **Methods Mol. Biol.**, v.532, p.29–53, 2009.

LEWIS, L. K.; HARLOW, G. R.; GREGG-JOLLY, L. A.; MOUNT, D. W. Identification of high affinity binding sites for LexA which define new DNA damage-inducible genes in Escherichia coli. **J. Mol. Biol.**, v.241, n.4, p.507–23, 1994.

LIMA, W. C.; MENCK, C. F. M. Replacement of the arginine biosynthesis operon in Xanthomonadales by lateral gene transfer. **J. Mol. Evol.**, v.66, n.3, p.266–75, 2008.

LIMA, W. C.; PAQUOLA, A. C. M.; VARANI, A. M.; VAN SLUYS, M.-A.; MENCK, C. F. M. Laterally transferred genomic islands in Xanthomonadales related to pathogenicity and primary metabolism. **FEMS Microbiol. Lett.**, v.281, n.1, p.87–97, 2008.

LIMA, W. C.; VAN SLUYS, M. A.; MENCK, C. F. Non-gamma-proteobacteria gene islands contribute to the Xanthomonas genome. **OMICS**, v.9, n.2, p.160–72, 2005.

LIMA, W. C.; VARANI, A. M.; MENCK, C. F. M. NAD biosynthesis evolution in bacteria: lateral gene transfer of kynurenine pathway in Xanthomonadales and Flavobacteriales. **Mol. Biol. Evol.**, v.26, n.2, p.399–406, 2009.

MANN, H. B.; WHITNEY, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. **Annals of Mathematical Statistics**, v.18, n.1, p.50–60, 1947.

MARTINS-PINHEIRO, M.; MARQUES, R. C. P.; MENCK, C. F. M. Genome analysis of DNA repair genes in the alpha proteobacterium Caulobacter crescentus. **BMC Microbiol.**, v.7, p.17, 2007.

MCCUTCHEON, J. P.; MCDONALD, B. R.; MORAN, N. A. Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. **PLoS Genet.**, v.5, n.7, p.e1000565, 2009.

MERKL, R. SIGI: score-based identification of genomic islands. **BMC Bioinformatics**, v.5, p.22, 2004.

NAKAMURA, Y.; ITOH, T.; MATSUDA, H.; GOJOBORI, T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. **Nat. Genet.**, v.36, n.7, p.760–6, 2004.

NIERMAN, W. C.; FELDBLYUM, T. V.; LAUB, M. T.; PAULSEN, I. T.; NELSON, K. E.; EISEN, J. A.; HEIDELBERG, J. F.; ALLEY, M. R.; OHTA, N.; MADDOCK, J. R.; POTOCKA, I.; NELSON, W. C.; NEWTON, A.; STEPHENS, C.; PHADKE, N. D.; ELY, B.; DEBOY, R. T.; DODSON, R. J.; DURKIN, A. S.; GWINN, M. L.; HAFT, D. H.; KOLONAY, J. F.; SMIT, J.; CRAVEN, M. B.; KHOURI, H.; SHETTY, J.; BERRY, K.; UTTERBACK, T.; TRAN, K.; WOLF, A.; VAMATHEVAN, J.; ERMOLAEVA, M.; WHITE, O.; SALZBERG, S. L.; VENTER, J. C.; SHAPIRO, L.; FRASER, C. M.; EISEN, J. Complete genome sequence of Caulobacter crescentus. **Proc. Natl. Acad. Sci. U.S.A.**, v.98, n.7, p.4136–41, 2001.

PETERSON, J. D.; UMAYAM, L. A.; DICKINSON, T.; HICKEY, E. K.; WHITE, O. The Comprehensive Microbial Resource. **Nucleic Acids Res.**, v.29, n.1, p.123–5, 2001.

PODELL, S.; GAASTERLAND, T. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. **Genome Biol.**, v.8, n.2, p.R16, 2007.

PÁL, C.; PAPP, B.; LERCHER, M. J. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. **Nat. Genet.**, v.37, n.12, p.1372–5, 2005.

QUARDOKUS, E. M.; BRUN, Y. V. Cell cycle timing and developmental checkpoints in Caulobacter crescentus. **Curr. Opin. Microbiol.**, v.6, n.6, p.541–9, 2003.

RAIN, J. C.; SELIG, L.; DE REUSE, H.; BATTAGLIA, V.; REVERDY, C.; SIMON, S.; LENZEN, G.; PETEL, F.; WOJCIK, J.; SCHÄCHTER, V.; CHEMAMA, Y.; LABIGNE, A.; LEGRAIN, P. The protein-protein interaction map of Helicobacter pylori. **Nature**, v.409, n.6817, p.211–5, 2001.

REVA, O. N.; TÜMMLER, B. Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. **BMC Bioinformatics**, v.6, p.251, 2005.

RIVERA, M. C.; JAIN, R.; MOORE, J. E.; LAKE, J. A. Genomic evidence for two functionally distinct gene classes. **Proc. Natl. Acad. Sci. U.S.A.**, v.95, n.11, p.6239–44, 1998.

DA ROCHA, R. P.; PAQUOLA, A. C. D. M.; MARQUES, M. D. V.; MENCK, C. F. M.; GALHARDO, R. S. Characterization of the SOS regulon of Caulobacter crescentus. **J. Bacteriol.**, v.190, n.4, p.1209–18, 2008.

SCHNEIDER, T. D.; STEPHENS, R. M. Sequence logos: a new way to display consensus sequences. **Nucleic Acids Res.**, v.18, n.20, p.6097–100, 1990.

SCHNEIKER, S.; PERLOVA, O.; KAISER, O.; GERTH, K.; ALICI, A.; ALTMEYER, M. O.; BARTELS, D.; BEKEL, T.; BEYER, S.; BODE, E.; BODE, H. B.; BOLTEN, C. J.; CHOUDHURI, J. V.; DOSS, S.; ELNAKADY, Y. A.; FRANK, B.; GAIGALAT, L.; GOESMANN, A.; GROEGER, C.; GROSS, F.; JELSBAK, L.; JELSBAK, L.; KALINOWSKI, J.; KEGLER, C.; KNAUBER, T.; KONIETZNY, S.; KOPP, M.; KRAUSE, L.; KRUG, D.; LINKE, B.; MAHMUD, T.; MARTINEZ-ARIAS, R.; MCHARDY, A. C.; MERAI, M.; MEYER, F.; MORMANN, S.; MUÑOZ-DORADO, J.; PEREZ, J.; PRADELLA, S.; RACHID, S.; RADDATZ, G.; ROSENAU, F.; RÜCKERT, C.; SASSE, F.; SCHARFE, M.; SCHUSTER, S. C.; SUEN, G.; TREUNER-LANGE, A.; VELICER, G. J.; VORHÖLTER, F.-J.; WEISSMAN, K. J.; WELCH, R. D.; WENZEL, S. C.; WHITWORTH, D. E.; WILHELM, S.; WITTMANN, C.; BLÖCKER, H.; PÜHLER, A.; MÜLLER, R. Complete genome sequence of the myxobacterium Sorangium cellulosum. **Nat. Biotechnol.**, v.25, n.11, p.1281–9, 2007.

SHAPIRO, S. S.; WILK, M. B. An Analysis of Variance Test for Normality (Complete Samples). **Biometrika**, v.52, n.3/4, p.591–611, 1965.

SKERKER, J. M.; LAUB, M. T. Cell-cycle progression and the generation of asymmetry in Caulobacter crescentus. **Nat Rev Microbiol**, v.2, n.4, p.325–37, 2004.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. **J. Mol. Biol.**, v.147, n.1, p.195–7, 1981.

SOLOMON, J. M.; GROSSMAN, A. D. Who's competent and when: regulation of natural genetic competence in bacteria. **Trends Genet.**, v.12, n.4, p.150–5, 1996.

STANTON, T. B. Prophage-like gene transfer agents-novel mechanisms of gene exchange for Methanococcus, Desulfovibrio, Brachyspira, and Rhodobacter species. **Anaerobe**, v.13, n.2, p.43–9, 2007.

STOREY, J. D.; TIBSHIRANI, R. Statistical significance for genomewide studies. **Proc. Natl. Acad. Sci. U.S.A.**, v.100, n.16, p.9440–5, 2003.

STORMO, G. D. DNA binding sites: representation and discovery. **Bioinformatics**, v.16, n.1, p.16–23, 2000.

TATUSOV, R. L.; FEDOROVA, N. D.; JACKSON, J. D.; JACOBS, A. R.; KIRYUTIN, B.; KOONIN, E. V.; KRYLOV, D. M.; MAZUMDER, R.; MEKHEDOV, S. L.; NIKOLSKAYA, A. N.; RAO, B. S.; SMIRNOV, S.; SVERDLOV, A. V.; VASUDEVAN, S.; WOLF, Y. I.; YIN, J. J.; NATALE, D. A. The COG database: an updated version includes eukaryotes. **BMC Bioinformatics**, v.4, p.41, 2003.

THOMAS, C. M.; NIELSEN, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. **Nat. Rev. Microbiol.**, v.3, n.9, p.711–21, 2005.

THOMPSON, W.; ROUCHKA, E. C.; LAWRENCE, C. E. Gibbs Recursive Sampler: finding transcription factor binding sites. **Nucleic Acids Res**, v.31, n.13, p.3580–5, 2003.

TSIRIGOS, A.; RIGOUTSOS, I. A new computational method for the detection of horizontal gene transfer events. **Nucleic Acids Res.**, v.33, n.3, p.922–33, 2005.

VERNIKOS, G. S.; PARKHILL, J. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. **Bioinformatics**, v.22, n.18, p.2196–203, 2006.

WAACK, S.; KELLER, O.; ASPER, R.; BRODAG, T.; DAMM, C.; FRICKE, W. F.; SUROVCIK, K.; MEINICKE, P.; MERKL, R. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. **BMC Bioinformatics**, v.7, p.142, 2006.

WELLNER, A.; LURIE, M. N.; GOPHNA, U. Complexity, connectivity, and duplicability as barriers to lateral gene transfer. **Genome Biol.**, v.8, n.8, p.R156, 2007.

WOESE, C. R.; FOX, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. **Proc. Natl. Acad. Sci. U.S.A.**, v.74, n.11, p.5088–90, 1977.

YU, X.; LIN, J.; MASUDA, T.; ESUMI, N.; ZACK, D. J.; QIAN, J. Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae. **Nucleic Acids Res**, v.34, n.3, p.917–27, 2006.