RODRIGO LUIS ALVES CARDOSO

ANÁLISE *IN SILICO* DE PROTEÍNAS TRANSPORTADORAS PRESENTES NO GENOMA DE *HERBASPIRILLUM SEROPEDICAE*

CURITIBA

2007

RODRIGO LUIS ALVES CARDOSO

ANÁLISE *IN SILICO* DE PROTEÍNAS TRANSPORTADORAS PRESENTES NO GENOMA DE *HERBASPIRILLUM SEROPEDICAE*

Monografia apresentada à disciplina de Estágio em Bioquímica como requisito à conclusão do curso de Bacharel em Ciências Biológicas, Setor de Ciências Biológicas, Departamento de Bioquímica, Universidade Federal do Paraná

Orientador: Prof. Dr. Leonardo M. Cruz

CURITIBA

2007

SUMÁRIO

1. ÍNDICE DE FIGURAS E TABELAS....3
2. LISTA DE ABREVIATURAS....5
3. INTRODUÇÃO....7
3.1. Sistemas de transporte....7
3.1.1. Proteínas canal....8
3.1.2. Transportadores secundários....10
3.1.3. Transportadores primários....11
3.1.4. Transportadores incompletamente caracterizados....12
3.2. Famílias de proteínas transportadoras....12
3.2.1. Família ABC....12
3.2.2. Sistema PTS....15
3.2.3. Proteínas transportadoras relacionadas à patogenicidade: o Sistema de Secreção do Tipo III (TTSS)....16
3.3. O sistema TC de classificação de proteínas transportadoras....19
3.4. Herbaspirillum seropedicae....21
3.5. Projeto GENOPAR....23
4. MATERIAL E MÉTODOS....24
4.1. Fonte de dados – projeto GENOPAR....24
4.2. Identificação de proteínas transportadoras no genoma de H. seropedicae....24
4.2.1. Banco de dados de proteínas transportadoras de genomas publicados – TransportDB (REN et al., 2007)....24
4.2.2. KAAS – KEGG Automatic Anotation Service (MORIYA et al., 2007)....26
4.2.3. Busca de informações na anotação do genoma de H. seropedicae....27
4.2.4. BLAST das ORF anotadas contra o banco de dados TCDB (SAYER et al., 2006)....27
4.3. Uso de rede neuronal para validação das ORF encontradas como possíveis proteínas transportadoras....28
4.4. Classificação das proteínas transportadoras....34
4.5. Similaridade das proteínas transportadoras de H. seropedicae e H. rubrisubalbicans....34
4.6. Análise de preferência de uso de códon....36
4.7. Análise de transportadores da família ABC....37
4.7.1. Análise de domínios transmembrana....38
4.7.2. Similaridade e conservação da “vizinhança” entre os genes de transportadores ABC em H. seropedicae e bactérias relacionadas....40
4.7.3. Análise filogenética....42
4.8. Automação das etapas de análise....43
5. RESULTADOS E DISCUSSÃO....44
5.1. Identificação das proteínas transportadoras....44
5.2. Uso de rede neuronal para validação das ORF encontradas como possíveis proteínas transportadoras....49
5.3. Comparação de proteínas transportadoras nos genomas de H. seropedicae e de outras bactérias....52
5.4. Classificação das proteínas transportadoras....57

5.5. Comparação entre possíveis proteínas transportadoras de H. seropedicae e H. rubrisubalbicans....68
5.6. Análise de preferência de códon....72
5.7. Análise dos transportadores ABC....80
5.7.1. Análise de domínios transmembrana....82
5.7.2. Análise de similaridade com outros organismos....87
5.7.3. Análise de "motivo C"....88
5.7.4. Similaridade e conservação da "vizinhança" entre os genes de transportadores ABC em H. seropedicae e bactérias relacionadas....93
5.7.5. Análise filogenética....99
6. CONCLUSÕES....101
7. REFERÊNCIAS BIBLIOGRÁFICAS....103
8. APÊNDICES....110
8.1. Apêndice A – Programa BLAST (Basic Local Alignment Search Tool - ALTSCHUL et al.; 1997)....110
8.2. Apêndice B - Scripts....112
8.2.1. Contigextract.sh....112
8.2.2. Baseextract.sh....114
8.2.3. Seqextractor.sh....115
8.2.4. Getsequences.sh....116
8.2.5. Separa_ORF....117
8.2.6. Blastparser.pl....118
8.2.7. Comparaorf.sh....121

1. ÍNDICE DE FIGURAS E TABELAS

Figura 1 – Interface gráfica de treinamento do software EasyFan.....31
Figura 2 – Gráficos que representam o aprendizado da FAN.....32
Figura 3 – Esquema representando o processo de “Coaprendizado”.....33
Figura 4 – Gráfico gerado pelo programa TMHMM.....39
Figura 5 – Exemplos de resultados produzidos pelo programa STRING.....41
Figura 6 – Resultados para busca de similaridade entre as ORF de *H. seropedicae* contra o banco de dados de proteínas transportadoras TransportDB.....46
Figura 7 – Validação das ORF de possíveis proteínas transportadoras segundo a anotação do GENOPAR.....47
Tabela 1 – A. Número de ORF identificadas como possíveis proteínas transportadoras através do programa BLASTX contra os bancos de dados TransportDB e TCDB e com o programa KAAS.....48
B. Número de ORF identificadas pelas ferramentas em conjunto.....48
Figura 8 – Classificação dos alinhamentos obtidos através de pesquisa BLAST contra o banco de dados TransportDB com uso de rede neuronal FAN.....50
Tabela 2 – Exemplo de classificação de alinhamentos produzidos pelo programa BLAST, realizada com rede neuronal FAN em coaprendizado com usuário.....51
Tabela 3 – Comparação do número de proteínas transportadoras em diferentes organismos.....53
Tabela 4 – Comparação entre proteínas transportadoras presentes no genoma de algumas Proteobacteria relacionadas a *H. seropedicae*.....54
Figura 9 – Proporção de ORF totais e número de ORF de proteínas transportadoras em relação ao tamanho dos genomas (Mb).....55
Tabela 5 – Classificação geral das possíveis proteínas transportadoras presentes no genoma de *H. seropedicae*.....56
Tabela 6 – Número de ORF encontradas distribuídas em suas respectivas famílias, conforme classificação encontrada no site TransportDB.....58
Figura 10 – Mapas de famílias de proteínas transportadoras construídos pelo KAAS:
A - sistema de secreção do tipo III.....60
B - sistema de secreção do tipo II. 61
C - proteínas de excreção. 62
D - montagem de flagelo. 63
E – PTS. 64
F – família ABC. 65
Tabela 7 – Subunidades constituintes do Sistema de Secreção do Tipo III ausentes em outras bactérias.....67
Figura 11 – Similaridade entre as ORF de possíveis transportadores em *H. seropedicae* e *H. rubrisubalbicans* através de pesquisa BLAST seguida de classificação através de rede neuronal FAN.....69
Tabela 8 – Possíveis proteínas transportadoras de *H. seropedicae* com indícios em *H. ribrisubalbicans*.....71
Figura 12 – Gráficos representando a distribuição do uso de códons pelo organismo *H. seropedicea:*
A: Uso de códons corresponde ao resultado obtido para todas as ORF de *H.*

seropedicae. 73
B: Uso de códons nas 880 proteínas transportadoras encontradas por pesquisa BLAST em relação ao banco de dados TransportDB.....74
Figura 13 – Comparação entre índices de tendência no uso de códons para ORF anotadas de *H. seropedicae* e para ORF de possíveis proteínas transportadoras:
A – Nc x CAI.....77
B – CAI x GC3s.....78
C – Nc x GC3s.....79
Tabela 9 – Sistemas de transporte da família ABC completos identificados no genoma de *H. seropedicae*.....83
Tabela 10 – Número de possíveis *operons* inteira ou parcialmente completos, e o número de hélices transmembrana encontrados na subunidade transmembrana (permease).....86
Tabela 11 – Proximidade taxonômica entre as subunidades para ligação de ATP em transportadores ABC de *H. seropedicae* e proteínas do mesmo tipo em outros organismos.....89
Figura 14 – Rede associativa para as proteínas da família ABC, realizada com o programa STRING.....95
Figura 15 – Um exemplo de *neighborhood*.....98
Figura 16 – Árvore filogenética das unidades ligadoras de ATP de transportadores ABC.....100

2. LISTA DE ABREVIATURAS

aa – aminoácidos
ABC – do inglês *ATP-binding cassete*
ATP - adenosina tri-fosfato (do inglês *Adenosine Triphosphate*)
ATPase – adenosina tri-fosfatase
BASH – do inglês *Bourne Again SHell*
BLAST – do inglês *Basic Local Alignment Search Tool*
CAI – do inglês *Codon Adaptation Index*
CAP3 – do inglês Contig Assembly Program 3
CIC – do inglês *Chloride Channell*
DMT – do inglês *Drug/Metabolite Transporter*
DNA – Ácido desoxirribonucléico (do inglês *desoxirribonucleic acid)*
Dr. - doutor
EC – do inglês *Enzyme Commission*
Embrapa – CNPAB - Centro Nacional de Agrobiologia da Embrapa
Embrapa – CNPSo - Centro Nacional de Pesquisa de Soja da Embrapa
EI - enzima I
ENc, ou Nc – do inglês *Effective Number of Codons*
ex. - exemplo
FAN – do inglês *Free Associative Neurons*
FeoB – do inglês *Ferrous Iron Uptake*
GCUA – do inglês *Graphical Codon Usage Analyser*
GENOPAR – Genoma do Paraná
HSP – do inglês *High Scoring pairs*
IAPAR - Instituto Agronômico do Paraná
KAAS – do inglês *KEGG Automatic Anotation Service*
kb - kilobase
KEGG – do inglês *Kyoto Encyclopedia of Genes and Genomes*
Mb - megabase
MerTP – do inglês *Mercuric Ion (Hg^{2+}) Permease*
MFS – do inglês *Major Facilitator Superfamily*
MIT - *CorA Metal Ion Transporter*
Mpb – *Mega pares de bases*
MscS – do inglês *Small Conductance Mechanosensitive Ion Channel*
MSD - Domínio transmembrana (do inglês *membrane-spanning domains*)
NBD – Domínio de ligação a nucleotídeo (do inglês *nucleotide-binding domains*)
NC-IUBMB - Comitê de Nomenclatura da União Internacional de Bioquímica e Biologia Molecular (do inglês *Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*)
nr - não redundante
OBS – Observação
ORF – do inglês *Open Reading Frames*
PAM – *do inglês Point Accepted Mutation*
PEP - do inglês *phosphoenolpyruvate*

PEP:PTS – do inglês *phosphoenolpyruvate (PEP)-dependent phosphotransferase system(PTS)*
PERL – do inglês *Practical Extraction and Report Language*
PPi – pirofosfato
Prof. - professor
PTS - sistema fosfotransferase (do inglês *phosphotransferase system*)
PUC-PR - Pontifícia Universidade Católica do Paraná
RNA – Ácido ribonucléico (do inglês *ribonucleic acid)*
RND – do inglês *Resistance-Nodulation-Cell Division*
STRING – do inglês *Search Tool for the Retrieval of Interacting Proteins*
TC - Classificação de transporte (do inglês *Transport Classification*)
TCDB – do inglês *Transport Classification Database*
TMHMM – do inglês *transmembrane hidden Markov model*
TRAP-T – do inglês*Tripartite ATP-independent Periplasmic Transporter*
TTSS - Sistema de secreção do tipo III (do inglês *Type III Secretion System)*
UEL - Universidade Estadual de Londrina
UEM - Universidade Estadual de Maringá
UFPR - Universidade Federal do Paraná
UFRGS - Universidade Federal do Rio Grande do Sul
UFRJ - Universidade Federal do Rio de Janeiro
UFSC - Universidade Federal de Santa Catarina
UNIOESTE - Universidade Estadual do Oeste do Paraná
UNIPAR - Universidade Paranaense
URL – do inglês *Uniform Resource Locator*
valor E - *expect*
www – do inglês *World Wide Web*

3. INTRODUÇÃO

3.1. Sistemas de transporte

A membrana plasmática é uma barreira seletiva que separa a célula do ambiente extracelular (KONINGS, 2006). Por desempenhar essa função, cabe à membrana plasmática uma variedade de papéis, dependendo da célula, tais como a eliminação de compostos tóxicos, manutenção e regulação da pressão de turgor (no caso de células vegetais), recebimento e processamento das informações do meio, transdução de energia, motilidade celular, crescimento, diferenciação, importação de nutrientes, exportação de metabólitos e blocos para construção da parede celular (LENGELER et al., 1998).

Devido a essa variedade de papéis, a membrana plasmática dá suporte para diversas proteínas que a ajudam no desempenho de todos eles. Estima-se que proteínas de membrana compreendam cerca de 10-40% de todas as proteínas em bactérias (SIEBOLD et al., 2001), com destaque para os eventos de transporte (LENGELER et al., 1998), visto que são inúmeras as situações em que a capacidade de transporte através da membrana plasmática é o fator que determina a sobrevivência desses organismos (KONINGS, 2006).

Devido às características hidrofóbicas da bicamada lipídica, essa serve como uma barreira para a passagem da maioria das moléculas polares e íons (KONINGS, 2006). Bactérias, entretanto, necessitam transportar solutos em altas taxas através da membrana plasmática para seu crescimento e metabolismo (LENGELER et al., 1998).

A passagem de íons e da maioria das moléculas biológicas através dessa membrana requer a atividade de proteínas específicas (KONINGS, 2006). Essas proteínas são chamadas de transportadores, sistemas de transporte, carreadores, ou permeases (LENGELER et al., 1998).

Os sistemas de transporte podem ser divididos em três classes principais: proteínas canal, sistemas de transporte primário e sistemas de transporte secundário; enquanto os processos de transporte podem ser classificados de acordo com diferentes aspectos, tais como estrutura do transportador, soluto transportado, entre outros, podendo ser divididos em quatro classes: difusão; transporte secundário; transporte primário; translocação de grupo (LENGELER et al., 1998).

Cada transportador tem um ou mais sítios de ligação ao substrato, e transportam esse substrato através de mudanças conformacionais reversíveis na sua estrutura. Comparações de peptídios revelam grande similaridade molecular entre transportadores, tanto de transporte ativo quanto passivo, o que sugere uma relação evolutiva entre as proteínas da superfamília de transportadores. A maior parte dos transportadores possui uma estrutura comum, indicando que eles possuem além de uma função geral comum, provavelmente uma origem evolutiva comum (SAYER, 1994 e 2000).

3.1.1. Proteínas canal

As proteínas canal transportam água ou íons específicos para onde sua concentração ou potencial elétrico são mais baixos, num evento energeticamente favorável. Elas permitem a esses íons ou moléculas fluir rapidamente pelas membranas, realizando desse modo transporte passivo ou difusão facilitada

(LENGELER et al., 1998).

Essas proteínas formam poros aquosos que se estendem através da bicamada lipídica, que ao serem abertos, permitem a solutos específicos (geralmente água ou íons inorgânicos de carga e tamanho apropriados) passar através deles e cruzar a membrana (TAKATA et al., 2004; GRANGEIRO et al., 2004). Não surpreendentemente, transportes através de canais ocorrem numa taxa muito mais rápida em relação aos mediados por carreadores (LENGELER et al., 1998).

Um grupo de proteínas canal é o das porinas, as quais permitem a livre passagem de íons ou moléculas polares através da membrana externa de bactérias. A membrana plasmática de muitas células também possui proteínas canal de água (aquaporinas). Através delas, moléculas de água são capazes de atravessar a membrana muito mais rapidamente que por difusão (TAKATA et al., 2004).

Esses tipos de sistemas de transporte, usados para aumentar a velocidade com a qual moléculas lipossolúveis (hidrofóbicas) ou água (que já possuem permeabilidade à membrana) atravessam a membrana, são chamados de facilitadores (LENGELER et al., 1998).

Dentre as proteínas canal pode-se ainda citar os membros das famílias CIC (*Chloride Channell*), responsáveis pela regulação do volume celular, regulação do pH intracelular, e excitabilidade da membrana (FOSKETT, 1998); MscS (*Small Conductance Mechanosensitive Ion Channel*)**,** canais desse tipo convertem forças mecânicas na bicamada lipídica da membrana em sinais elétricos, importantes, por exemplo, para o controle da forma e do volume celular (HURST et al. 2007); MIT (CorA *Metal Ion Transporter*), responsável pelo transporte de íons metálicos, com

destaque para íons magnésio, sendo o sistema primário de importação desse íon em procariotos (LUNIN et al. 2006).

3.1.2. Transportadores secundários

Os transportadores secundários, ao contrário dos canais, ligam-se somente a uma molécula de substrato de cada vez. Esses transportadores realizam transporte dirigido por diferença de concentração ou potencial elétrico (SAYER, 2000). No transporte uniporte, a permeação de um simples soluto é facilitada, e o transporte é direcionado simplesmente pela diferença de concentração do soluto através da membrana (LENGELER et al., 1998).

Esses transportadores, também chamados de cotransportadores ou carreadores acoplados (*coupled carriers)*, possuem ainda duas classes: os transportadores simporte, os quais usam o fluxo de um soluto para direcionar o fluxo de um outro soluto, simultaneamente e na mesma direção, através da membrana; e os transportadores antiporte, que acoplam o fluxo de entrada de um substrato ao fluxo de saída de outro substrato, em direções opostas, através da membrana (SAYER, 2000).

Devido ao fato de transportadores simporte e antiporte realizarem movimento contra o gradiente de concentração de algumas moléculas, eles são freqüentemente chamados de transportadores ativos, mas diferente das bombas, não fazem hidrólise de ATP durante o transporte. Um termo usado para designar esses transportadores é cotransportadores, devido à sua capacidade de transportar dois solutos diferentes simultaneamente (MARKOVICH & MURER, 2004).

Entre os transportadores secundários destacam-se os membros das famílias

MFS (*Major Facilitator Superfamily*), os quais transportam uma grande quantidade de compostos, tais como açúcares simples, oligossacarídeos, inositol, drogas, aminoácidos, nucleosídeos, metabólitos do ciclo de Krebs, e uma grande variedade de ânions e cátions (PAO et al. 1998). RND (*Resistance-Nodulation-Cell Division*), DMT (*Drug/Metabolite Transporter* – JACK et al., 2001), TRAP-T (*Tripartite ATP-independent Periplasmic Transporter*), entre outras, também são numerosas e diversificadas.

A família RND está envolvida, entre outros processos, nas exportações de metais pesados (DINH et al., 1994), drogas, e oligossacarídios para nodulação com finalidade de fixação de nitrogênio simbiótica (SAYER, 2000; SAYER et al., 2006). Membros da família TRAP-T, podem entre outras funções, estar envolvidas na importação de derivados de carboxilato (KELLY & THOMAS, 2001).

3.1.3. Transportadores primários

Os sistemas de transporte primários acoplam ao transporte de um soluto uma reação química ou fotoquímica (LENGELER et al., 1998). ATPases são bombas que usam a energia da hidrólise de ATP para movimentar íons ou pequenas moléculas através da membrana contra um gradiente de concentração ou potencial elétrico. Esse processo, chamado transporte ativo, é um exemplo de uma reação química acoplada (SAYER, 2000).

Os transportadores primários, tais como as famílias ABC e P-ATPase, acoplam ao transporte energia química, elétrica ou solar. Os membros da família P-ATPase geralmente estão envolvidos na importação e efluxo de íons; entre outras funções, em *Listeria monocytogenes* conferem resistência a cádmio (WU et al.,

2006).

3.1.4. Transportadores incompletamente caracterizados

Ainda existem transportadores cuja função é conhecida, porém os mecanismos de transporte desconhecidos, tais como as famílias MerTP (*Mercuric Ion (Hg^{2+}) Permease*) que confere resistência a mercúrio (QIAN et al., 1998) e FeoB (*Ferrous Iron Uptake*) responsável pela importação de ferro (KAMMLER et al., 1993).

3.2. Famílias de proteínas transportadoras

3.2.1. Família ABC

Dentre os transportadores existentes, destacam-se os transportadores pertencentes à superfamília ABC (ou *ATP-binding cassete*). Estes transportadores acoplam a energia de hidrólise do ATP à translocação de uma grande variedade de substâncias para dentro ou fora das células e organelas (ANNILO et al., 2006).

Todo transportador ABC aparentemente é composto por quatro domínios protéicos ou subunidades: dois domínios transmembrana hidrofóbicos (MSD, do inglês *membrane-spanning domains,* ou TMD, do inglês *transmembrane domain*) que se presume constituir a via de translocação ou canal através da membrana; e dois domínios hidrofílicos de ligação a nucleotídeos (NBD, do inglês *nucleotide-binding domains*) também conhecidos como subunidades de ligação a ATP, os quais interagem na superfície citoplasmática para fornecer energia ao transporte ativo, fazendo o transportador funcionar através de ligação e hidrólise do ATP (DAVIDSON & CHEN, 2004; BIEMANS-OLDEHINKEL et al., 2006; ANNILO et al., 2006).

Presumivelmente, a ligação ao ATP e/ou sua hidrólise estão acopladas a mudanças conformacionais no MSD que é mediador do bombeamento unidirecional de substratos através da membrana. (DAVIDSON & CHEN, 2004).

Mesmo havendo baixa homologia entre os domínios MSD em diferentes subfamílias, um grau maior de homologia é mantido entre toda a superfamília em relação aos NBD (25% a 30% de identidade), sugerindo um mecanismo similar empregado para o acoplamento do transporte à hidrólise de ATP (DAVIDSON & CHEN, 2004).

Análises de seqüências mostram que membros da superfamília ABC podem ser organizados dentro de subfamílias e sugere-se que tenham divergido de uma forma ancestral comum. Sistemas de transporte ABC podem ser encontrados em procariotos, arqueobactérias e eucariotos. A maioria deles é mediador na importação ativa ou efluxo de moléculas específicas através de membranas biológicas. Eles manipulam uma grande variedade de compostos, os quais diferem em natureza e tamanho (FICHANT et al. 2006).

Os domínios de ligação ao substrato possuem duas funções, sendo responsáveis pelo transporte de alta afinidade, característico desses transportadores, e pela estimulação da ATPase (DAVIDSON & CHEN, 2004).

Os transportadores ABC também funcionam no efluxo de substâncias em bactérias, os quais incluem componentes da superfície da célula (tais como polissacarídeos capsulares, lipopolissacarídeos e ácido teicóico); proteínas envolvidas na patogênese bacteriana (como hemolisina, proteína de ligação heme, e protease alcalina); antibióticos peptídeos, heme, drogas e sideróforos (DAVIDSON &

CHEN, 2004; BIEMANS-OLDEHINKEL et al., 2006).

A respeito da grande diversidade de substratos transportados, as seqüências dos componentes ABC são conservadas entre todos os transportadores ABC. Vários motivos de seqüência conservados, tais como os motivos “Walker A” e "Walker B” que são encontrados em várias ATPases, podem ser identificados, e mutações nestas regiões freqüentemente reduzem severamente ou eliminam o transporte e a atividade da ATPase (BIEMANS-OLDEHINKEL et al., 2006).

A estrutura de um monômero NBD pode ser dividida em dois subdomínios: um subdomínio semelhante à *RecA* consistindo de duas folhas beta e seis alfa hélices e um subdomínio helicoidal menor formado por três ou quatro alfa hélices. O subdomínio helicoidal é específico para os transportadores ABC e não ocorre em outras ATPases. O motivo sinal, também conhecido como motivo LSGGQ (*liker peptide*) ou motivo C, é usado como uma “assinatura” para identificar transportadores ABC e é o único principal motivo conservado que não está em contato com o nucleotídeo na estrutura do monômero (BIEMANS-OLDEHINKEL et al., 2006; ANNILO et al., 2006).

Todos os transportadores ABC parecem ter dois domínios NBD e a hidrólise de ATP é altamente cooperativa. Várias evidências indicam que a associação e dissociação dos NBD é uma característica chave dos transportadores ABC. Alguns autores sugerem que somente uma das duas ligações a ATP é hidrolisada em cada evento de transporte, e que os dois sítios alternam a catálise. É difícil determinar se somente um ou ambos os ATP são hidrolisados a cada evento de transporte. Medida de crescimento *in vivo* em bactéria sugere que somente um ATP é

necessário para transportar um substrato para dentro da célula, entretanto uma recente descrição usando transportador *OpuA* purificado e reconstituído sugere que dois ATP sejam necessários. Ambos os modelos foram propostos (DAVIDSON & CHEN, 2004).

3.2.2. Sistema PTS

As bactérias utilizam diferentes mecanismos de transporte para a captação de solutos: difusão facilitada, transporte ativo movido a ATP ou gradiente iônico, e translocação de grupo. Translocação de grupo de carboidratos é mediado pelo sistema PEP:PTS (*phosphoenolpyruvate dependent phosphotransferase system*) (KUNDIG et al., 1964; MITCHELL et al., 2007). O PTS catalisa a translocação com concomitante fosforilação de açúcares e hexitóis e também regula o metabolismo em resposta à disponibilidade de carboidratos (POSTMA et al., 1993 e POSTMA et al., 1996).

O PTS consiste de duas proteínas citoplasmáticas, enzima I (EI) e Hpr, e um número variável de complexos transportadores de açúcar (enzimas $II^{açúcar}$). A EI transfere grupos fosforil do PEP para a proteína carreadora de fosforil Hpr. A Hpr então transfere os grupos fosforil para os diferentes complexos de transporte (SIEBOLD et al., 2001).

Os PTS ocorrem em bactérias, mas não ocorrem em arqueobactérias e eucariotos (SIEBOLD et al., 2001). As seqüências de aminoácidos dos componentes EI e Hpr são altamente conservadas em todas as bactérias. O número e estrutura de transportadores PTS varia entre as espécies. Eles podem ser agrupados por comparação de seqüência em quatro famílias estruturalmente diferentes (PAULSEN

et al., 2000).

Escherichia coli codifica para 38 proteínas PTS em 22 transportadores diferentes. *Mycoplasma genitalium* contém somente um gene para EI e Hpr e dois genes para transportadores de açúcar (enzima II). *Treponema pallidum*, *Chlamydia trachomatis* e *Xylella fastidiosa*, contém proteínas similares a EI e Hpr, mas nenhum transportador de açúcar (enzima II). *Mycobacterium tuberculosis* também não apresenta nenhum PTS completo (SIEBOLD et al., 2001).

Os transportadores de açúcar (enzima II, ou EII) consistem geralmente de três unidades funcionais, IIA, IIB e IIC, as quais ocorrem como subunidades protéicas em um complexo ou como domínios de uma única cadeia polipeptídica. As unidades IIA e IIB transferem seqüencialmente grupos fosforil do Hpr para o açúcar transportado. A unidade IIC contém o sítio de ligação ao açúcar. As unidades EI, Hpr e IIA são fosforiladas em uma histidina, a unidade IIB é fosforilada em uma cisteína ou histidina, dependendo do transportador (SIEBOLD et al., 2001).

3.2.3. Proteínas transportadoras relacionadas à patogenicidade: o Sistema de Secreção do Tipo III (TTSS)

Bactérias patogênicas gram-negativas desenvolveram mecanismos sofisticados para infectar e colonizar seus hospedeiros. Alguns destes mecanismos requerem a montagem de “organelas” multicomponentes na superfície bacteriana. Anteriormente à sua montagem, cada subunidade deve antes ser exportada até seu ponto de incorporação na estrutura nascente (GRANGEIRO et al., 2004).

Devido ao envelope celular das bactérias gram-negativas apresentar uma barreira ao movimento dos componentes da organela, as bactérias desenvolveram

um mecanismo de secreção/transporte protéico para facilitar a montagem organelar superficial (GRANGEIRO et al., 2004).

A via de secreção do tipo III, a qual participa da montagem do flagelo e organelas associadas à virulência, secreta proteínas através das duas membranas, independentemente da via de secreção, sem a necessidade de um intermediário periplasmático ou processamento proteolítico (KIMBROUGH & MILLER, 2002).

O sistema de secreção do tipo III (TTSS, do inglês *Type III Secretion System*) é usado para transportar fatores de virulência (efetores) do patógeno até a célula hospedeira e só é ativado quando a bactéria entra em contato com seu hospedeiro (GALAN & COLLMER, 1999).

As subunidades protéicas do TTSS são muito similares àquelas encontradas na biossíntese de flagelo (KOMORIYA et al., 1999). Entretanto, enquanto as subunidades flagelares formam uma estrutura em anel para permitir a secreção da flagelina e é uma parte integral do próprio flagelo, as sub-unidades do tipo III na membrana externa translocam proteínas secretadas através de uma estrutura de canal. As proteínas do flagelo também compartilham similaridade, provavelmente devido a evolução do TTSS a partir da via biossintética do flagelo (GRANGEIRO et al., 2004).

O TTSS associado a virulência são organelas especializadas que translocam proteínas de virulência bacteriana (efetores) do citoplasma bacteriano diretamente para o interior do citoplasma das células hospedeiras. Estes efetores translocados alteram funções celulares básicas do hospedeiro, como transdução de sinal, arquitetura citoesquelética, tráfego de membrana e expressão gênica (GRANGEIRO et al., 2004).

Em *Salmonella typhimurium* os genes para o TTSS estão localizados em uma região de 40Kb do cromossomo. Estes genes são divididos em:

a) componentes do aparato de exportação – constitui o núcleo do aparato de exportação, sendo em sua maioria proteínas integrais de membrana (inclui SpaOPQRS, InvAC, OrgB);

b) componentes estruturais do complexo da agulha – é composto pelos seguintes componentes: PrgHIJK e InvG;

c) *translocons* – proteínas SspBCD (ou SipBCD) que promovem o movimento das proteínas efetoras através da membrana eucariótica; acredita-se que formem um poro na membrana eucariótica; na ausência de um destes componentes, as proteínas efetoras são incapazes de cruzar a membrana eucariótica;

d) reguladores – restringem a expressão do TTSS a locais específicos no hospedeiro e coordenam a montagem do aparato de secreção; são codificados dentro do SPI1 (InvF, HilA, HilD, SirC, SprB) ou em outros locais no genoma (PhoP/PhoQ e SirA/BarA);

e) efetores – alguns são codificados dentro do SPI1 (SspA/SipA, SptP, AvrA) ou em outros locais no genoma (SopABDEE2, SspH1, SlrP);

f) chaperonas – proteínas pequenas, acídicas, formadas principalmente por alfa-hélices, que facilitam a secreção e translocação de proteínas efetoras específicas (ex., SicA, InvB e SicP) (KIMBROUGH & MILLER, 2002).

3.3. O sistema TC de classificação de proteínas transportadoras

O Sistema de Classificação de Transporte (*Transport Classification (TC) System*) é um sistema de classificação aprovado pelo NC-IUBMB (*Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*) análogo ao sistema de classificação de enzimas EC (*Enzyme Commission*), mas que incorpora informação filogenética (BUSCH & SAYER, 2002).

O sistema TC emprega uma etiqueta de cinco dígitos, onde:

1º. dígito é um número e designa a CLASSE das proteínas transportadoras;

2º. dígito é uma letra e designa a SUBCLASSE, referindo-se ao mecanismo de translocação e/ou a fonte de energia usada para o processo;

3º. dígito é um número e especifica a FAMÍLIA da proteína transportadora;

4º. dígito representa a SUBFAMÍLIA;

Estes níveis hierárquicos são definidos e diferenciados com base na sua estrutura primária.

5º. dígito indica uma proteína transportadora em particular.

A classificação dos transportadores, segundo o sistema TC, em classe e subclasse é a seguinte:

1. Poros e canais

1.A. canais alfa-hélice

1.B. porinas folha-beta

1.C. toxinas formadoras de poros

OBS.: proteínas/peptídios sintetizados por uma células e excretados para inserção na membrana de outra célula, onde irão formar poros transmembrana.

1.D. canais não ribossômicos

1.E. Holinas

OBS.: compreendem cerca de 40 famílias distintas de proteínas que exibem características estruturais e funcionais comuns, mas sem similaridade significativa entre as seqüências entre membros de famílias distintas; estão envolvidas na lise celular.

2. Transportadores movidos por potencial eletroquímico

2.A . transportadores ou carreadores (uniporte, simporte e antiporte)

2.B. transportadores não ribossômicos

OBS.: substâncias não peptídicas que ligam cátions em seu interior hidrofílico e transportam o complexo através da membrana expondo seu exterior hidrofóbico.

3. Transportadores ativos primários

3.A . transportadores movidos por hidrólise de ligação PPi

3.B. transportadores movidos por descarboxilação

OBS.: restrito a procariotos.

3.C. transportadores movidos por transferência de grupo metil

OBS.: uma única família de proteína foi caracterizada (*Na^+-transporting methyltetrahydromethanopterin:coenzyme M methyltransferase*); restrito a arqueobactérias.

3.D. transportadores movidos por oxidoredução

3.E. transportadores movidos pela luz

4. Translocadores

5. Carreadores de elétrons transmembrana

8. Fatores acessórios envolvidos no transporte

9. Sistemas de transporte incompletamente caracterizados

3.4. Herbaspirillum seropedicae

Devido às diferenças de finalidades para os solutos, para o transporte desses, bem como do tipo de célula e membrana que realiza o transporte, existe uma grande diversidade de sistemas de transporte, principalmente em bactérias, visto que esses organismos estão mais sujeitos a mudanças do meio e também porque eles usam os sistemas de transporte para outras finalidades (sensorial e mobilidade, por exemplo) (LENGELER et al., 1998).

Nesses organismos, os transportadores estão envolvidos na importação em larga escala de moléculas, em mecanismos de virulência e resistência a antibióticos (FICHANT et al., 2000).

Devido a essa importância, as proteínas dessa superfamília foram estudadas na anotação do genoma da bactéria *Herbaspirillum seropedicae* (BALDANI et al. 1986). Essa bactéria é diazotrófica, ou seja, capaz de fixar nitrogênio, elemento constituinte dos aminoácidos e nucleotídeos, num processo chamado de fixação biológica de nitrogênio (RONCATO-MACCARI et al., 2003).

O ar atmosférico é rico em nitrogênio na forma de um gás inerte. A molécula desse gás é formada por dois átomos de nitrogênio ligados por uma tripla ligação química, muito estável (N_2), e que não pode ser captado e utilizado pelos seres vivos nessa forma. O gás nitrogênio sofre então a ação de uma enzima denominada nitrogenase, a qual é capaz de quebrar a tripla ligação química desse gás, convertendo-o em amônia. Essa molécula pode ser assimilada por outros

organismos, e com isso o nitrogênio é fixado biologicamente (BARNEY et al., 2006).

Devido à capacidade de fixar nitrogênio, *H. seropedicae* e os outros organismos diazotróficos desempenham um importante papel ecológico, necessário para a manutenção do equilíbrio de diversos ecossistemas (BALDANI et al., 1997).

H. seropedicae também é uma bactéria endofítica e associativa, ou seja, é capaz de colonizar o interior de tecidos de plantas sem causar algum dano aparente ao hospedeiro (BALDANI & BALDANI, 2004), encontrada nas raízes, folhas e caules de plantas, principalmente gramíneas economicamente importantes, tais como arroz e cana-de-açúcar (RONCATO-MACCARI et al., 2003).

Atualmente o gênero *Herbaspirillum* possui 9 espécies, incluindo a espécie *Herbaspirillum seropedicae,* e está taxonomicamente posicionado na família Oxalobacteraceae, ordem Burkholderiales, classe Betaproteobacteria, filo Proteobacteria (BALDANI et al., 1986; BALDANI et al., 1996; KIRCHHOF et al., 2001; VALVERDE et al., 2003; DING & YOKOTA, 2004; IM et al., 2004; ROTHBALLER et al., 2006).

3.5. Projeto GENOPAR

O seqüenciamento genômico da bactéria endofítica fixadora de nitrogênio *H. seropedicae* está sendo realizado pelo projeto GENOPAR – Genoma do Paraná, coordenado pelo Prof. Dr. Fábio de Oliveira Pedrosa, com participação de diversas instituições dentro e fora do Estado do Paraná, tais como a Universidade Federal do Paraná (UFPR, sede do projeto), Pontifícia Universidade Católica do Paraná (PUC-PR), Instituto Agronômico do Paraná (IAPAR), Universidade Estadual de Londrina (UEL), Centro Nacional de Pesquisa de Soja da Embrapa (Embrapa – CNPSo), Universidade Estadual de Maringá (UEM), Universidade Paranaense (UNIPAR), Universidade Estadual do Oeste do Paraná (UNIOESTE), Universidade Federal de Santa Catarina (UFSC), Universidade Federal do Rio de Janeiro (UFRJ), Centro Nacional de Agrobiologia da Embrapa (Embrapa – CNPAB) e Universidade Federal do Rio Grande do Sul (UFRGS) (GENOPAR – www.genopar.org).

Duas fases podem ser bem definidas num seqüenciamento genômico: a fase experimental, que tem como objetivo extrair, fragmentar, clonar e seqüenciar DNA; e a fase de análise computacional, que tem por objetivos montar e anotar o genoma através dos resultados obtidos com a fase experimental, retornando a ela se necessárias novas análises (GENOPAR – www.genopar.org).

Atualmente o genoma de *H. seropedicae* encontra-se em fase de análise computacional (preenchimento de gaps/falhas, montagem e anotação), possuindo em torno de 5,7 Mb, 287 *contigs* e 5.100 ORF (GENOPAR – www.genopar.org).

4. MATERIAL E MÉTODOS

4.1. Fonte de dados – projeto GENOPAR

Para o desenvolvimento deste trabalho foram usados os dados de seqüenciamento e anotação do genoma da bactéria *Herbaspirillum seropedicae*, obtidos do projeto GENOPAR (www.genopar.org) e gentilmente cedidos pelo Prof. Dr. Fábio Pedrosa, coordenador do projeto.

4.2. Identificação de proteínas transportadoras no genoma de H. seropedicae

Para a identificação das ORF (Fase de Leitura Aberta – do inglês *Open Reading Frames*) de *H. seropedicae* que transcrevem para proteínas transportadoras, foi utilizado o programa BLAST – *Basic Local Alignment Search Tool* (ALTSCHUL et al., 1997) (Apêndice A) e um banco de dados constituído de proteínas transportadoras encontradas em genomas publicados (TransportDB – REN et al., 2004; REN et al., 2007). Para complementar a análise também foram utilizados o programa KAAS – KEGG *Automatic Anotation Service* (MORIYA et al., 2007), a anotação do genoma de *H. seropedicae* (GENOPAR – www.genopar.org), e pesquisa de similaridade BLAST contra o banco de dados TCDB – *Transport Classification Database* (SAYER et al., 2006).

4.2.1. Banco de dados de proteínas transportadoras de genomas publicados –TransportDB (REN et al., 2007)

Um banco de dados foi criado utilizando o conteúdo do *site* TransportDB

(REN et al., 2007), o qual disponibiliza seqüências de proteínas transportadoras encontradas em genomas seqüenciados. Este banco de dados contém proteínas transportadoras identificadas a partir de seqüências genômicas publicamente disponíveis. O banco de dados contém cerca de 37.000 seqüências (REN et al., 2007).

Visto que o conteúdo do *site* não estava disponível para *download* em sua integridade, foi utilizado na criação do banco de dados, um *script* em SHELL BASH, o qual extrai esse conteúdo através da busca de URLs de todas as famílias de proteínas presentes, conforme a organização do *site* (Apêndice B – ver Getsequences.sh). Extraídos esses dados o banco foi criado para atender ao programa BLAST, utilizando-se um programa dentro do próprio BLAST chamado "*formatdb"* (BEDELL et al., 2003).

A busca foi realizada com o programa BLASTX (converte seqüências de nucleotídeos em proteínas para busca em um banco de dados contendo seqüências de proteínas) utilizando-se as 5.100 ORF da anotação do genoma de *H. seropedicae* contra o banco de dados de proteínas transportadoras extraído do *site* TransportDB.

Os parâmetros do programa foram ajustados para exibir somente a seqüência de maior similaridade no banco de dados (opções "-b1" e "-v1") e com um "valor E" (EXPECT) arbitrário menor ou igual a 5×10^{-5}.

Os resultados foram obtidos segundo a formatação padrão de saída do programa BLAST, e as informações extraídas através de um *script* em PERL utilizando bibliotecas do BIOPERL, versão 1.4 (www.bioperl.org): biblioteca *GenericHit* e biblioteca *GenericHSP.* O *script* foi usado para extrair informações

sobre os *Hits* (as seqüências que produziram alinhamento) e HSP (*High Scoring pairs;* alinhamentos para cada *Hit* encontrado) (Apêndice B - ver Blastparser.pl).

Dentre os valores extraídos encontram-se a porcentagem de similaridade do alinhamento, recalculada para os tamanhos totais das *queries* (seqüências submetidas à análise) e *subjects* (seqüências presentes no banco de dados), bem como a proporção entre os tamanhos das *queries* e *subjects*.

Os valores obtidos na análise, bem como aqueles recalculados foram usados para avaliar e validar as ORF identificadas como possíveis proteínas transportadoras usando uma rede neuronal (ver *Uso de rede neuronal para validação das ORF encontradas como possíveis proteínas transportadoras* em Material e Métodos).

4.2.2. KAAS – *KEGG Automatic Anotation Service* (MORIYA et al., 2007)

Todas as ORF, com as seqüências convertidas em proteínas, da anotação do genoma de *H. seropedicae,* também foram submetidas à análise pelo programa KAAS, visando obter informações sobre em quais processos metabólicos essas proteínas estariam participando, quais suas subunidades no caso de serem poliméricas, e encontrar detalhes sobre as proteínas transportadoras já identificadas com o programa BLAST, usando essas informações para análises mais específicas.

A ferramenta KAAS, disponível no *site* KEGG (*Kyoto Encyclopedia of Genes and Genomes* – KANEHISA, 2002), utiliza a estratégia de BLAST bidirecional. Nesta estratégia, todas as seqüências são usadas como *query* e *subject* e a determinação da homologia é feita quando duas seqüências obtém seus melhores *hits* uma em relação à outra. O método é duas vezes mais lento porém duas vezes mais preciso que uma pesquisa BLAST comum. Nesta pesquisa, o banco de dados KEGG foi

usado (MORIYA et al., 2007).

Essa pesquisa usando o programa KAAS mostrou um resultado positivo para 2.652 ORF das 5.100 ORF anotadas no projeto GENOPAR. Somente aquelas identificadas como possíveis proteínas transportadoras foram usadas neste trabalho.

4.2.3. Busca de informações na anotação do genoma de *H. seropedicae*

De posse das ORF de *H. seropedicae* já anotadas pelo projeto GENOPAR, as informações referentes a essas ORF foram utilizadas como suporte para os resultados obtidos nas etapas anteriores, tais como as validações das ORF, presença de *frameshift* (mudança de fase de leitura), e provável proteína produzida pela ORF para fins de comparação com os outros resultados.

4.2.4. BLAST das ORF anotadas contra o banco de dados TCDB (SAYER et al., 2006)

Com finalidade de comparação dos resultados, também foi realizada uma pesquisa BLAST contra o banco de dados *Transport Classification Database* – TCDB (SAYER et al., 2006), cujas proteínas que compõe o banco estão disponíveis para *download*. Essa pesquisa BLAST seguiu as mesmas especificações daquela realizada em relação ao conteúdo do *site* TransportDB.

TCDB é um banco de dados "curado", ou seja, as seqüências só são adicionadas ao banco se houver referências de trabalhos publicados nos quais foram usadas as proteínas relativas a essas seqüências. Assim, os dados do banco foram avaliados com base em cerca de 10.000 referências. O banco de dados possui cerca de 3.000 proteínas classificadas em mais de 550 famílias de

transportadores de acordo com o sistema de classificação *TC system* (SAYER et al., 2006).

4.3. Uso de rede neuronal para validação das ORF encontradas como possíveis proteínas transportadoras

Para resolver eventuais problemas de classificação e validação das ORF de proteínas transportadoras, foi utilizado uma rede neuronal para analisar e classificar alinhamentos produzidos pelo programa BLAST. A rede neuronal utilizada foi a FAN (*Free Associative Neurons*).

As redes FAN são uma abordagem de aprendizado *neuro-fuzzy*. O método é baseado no desenvolvimento de uma estratégia de reconhecimento de padrões que garanta boa performance no aprendizado aliado às vantagens computacionais da clareza na representação dos padrões, e portabilidade das unidades de representação, que são chamadas neurônios ou FAN (RAITTZ, 2002).

Atualmente conta-se com programas desenvolvidos para facilitar o uso da abordagem FAN. O programa EasyFan (GARRETT et al., 2006), desenvolvido na UFPR e de código aberto, é um ambiente para treinamento de redes (figura 1) que tem capacidade de comunicar-se com planilhas eletrônicas tipo *Excel*, o que populariza e facilita seu uso.

Para analisar os dados com o *software* EasyFan, as características dos alinhamentos encontradas pelo programa BLAST foram extraídas conforme mencionado anteriormente (ver *Identificação de Proteínas Transportadoras* em Material e Métodos), e classificados em três classes (alta similaridade, baixa similaridade e média similaridade). A classificação foi feita em relação ao

alinhamento, não ao fato da proteína ser ou não um transportador. Porém, o fato de ter-se uma ORF de alinhamento válido contra um banco de dados de proteínas transportadoras, foi considerado um forte indício de que a proteína realmente pertença a esse grupo.

Através das características extraídas de cada alinhamento, o usuário fez uma primeira classificação, de forma subjetiva, dividindo os alinhamentos nas três classes utilizadas (alta, baixa e média similaridade), formando um padrão de classificação.

A rede neuronal FAN foi treinada para o reconhecimento do padrão dessa classificação, e reclassificou os alinhamentos. Alguns gráficos gerados pelo programa EasyFan a respeito do treinamento são apresentados na figura 2.

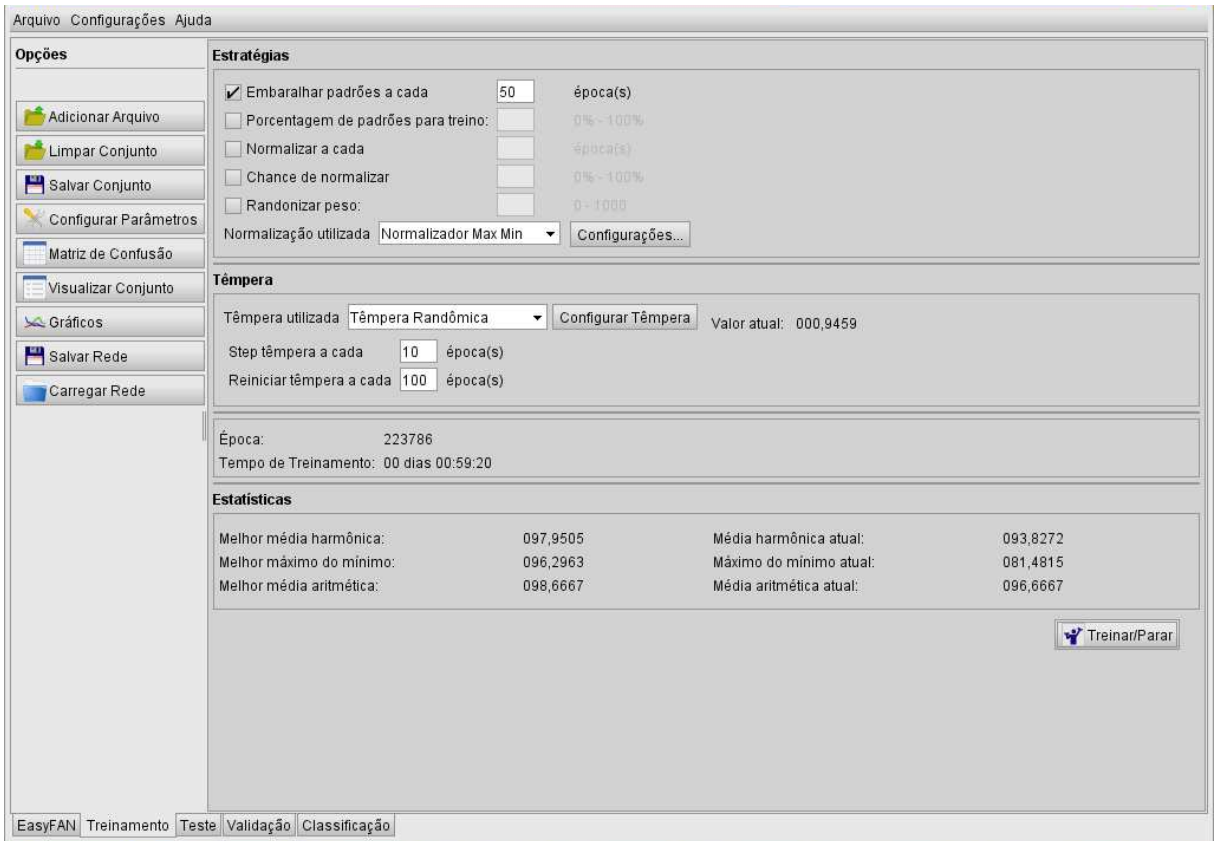
Verificou-se através desse processo, que a classificação subjetiva realizada pelo usuário apresentou incoerências (o padrão apresentado no início da classificação foi sutilmente diferente do padrão ao término dessa).

Como a rede neuronal FAN faz um uso mais coerente do padrão encontrado, ao comparar-se as classificações dos dois, rede e usuário, tornaram-se visíveis para esse último suas incoerências no padrão produzido. Assim, foi possível que esse melhorasse sua classificação, através da correção de suas incoerências (melhora do padrão), levando em consideração a opinião/classificação da rede.

Novamente a rede neuronal FAN foi treinada, agora com a classificação do usuário melhorada, e o processo foi refeito. Visando melhorar ao máximo a classificação, todo esse ciclo repetiu-se várias vezes, num processo chamado “coaprendizado”, fazendo com que o usuário e a rede aprendessem juntos a classificar os dados, até os dois concordarem o máximo possível (figura 3).

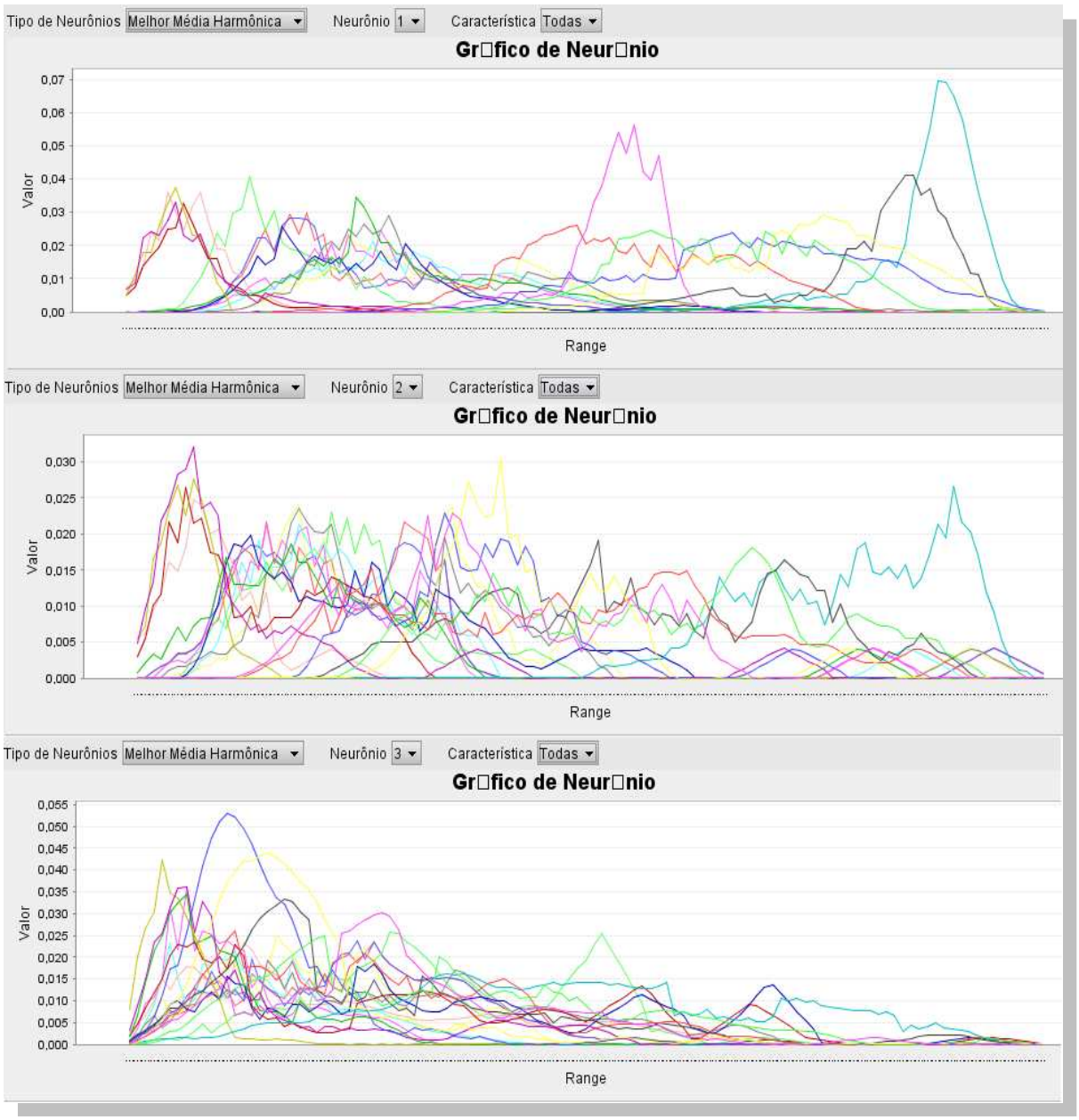
Esse ciclo foi repetido 17 vezes, até o ponto em que o usuário e a rede concordaram com porcentagem em torno de 98,6% de média harmônica. O ciclo foi parado no momento em que as alterações numa das classificações não mais surtiu efeito na reclassificação, ou seja, quando as modificações feitas pelo usuário não foram mais aceitas pela rede neuronal e vice-versa, de modo que a classificação não pôde mais ser melhorada.

Figura 1 – Interface gráfica de treinamento do *software* EasyFan



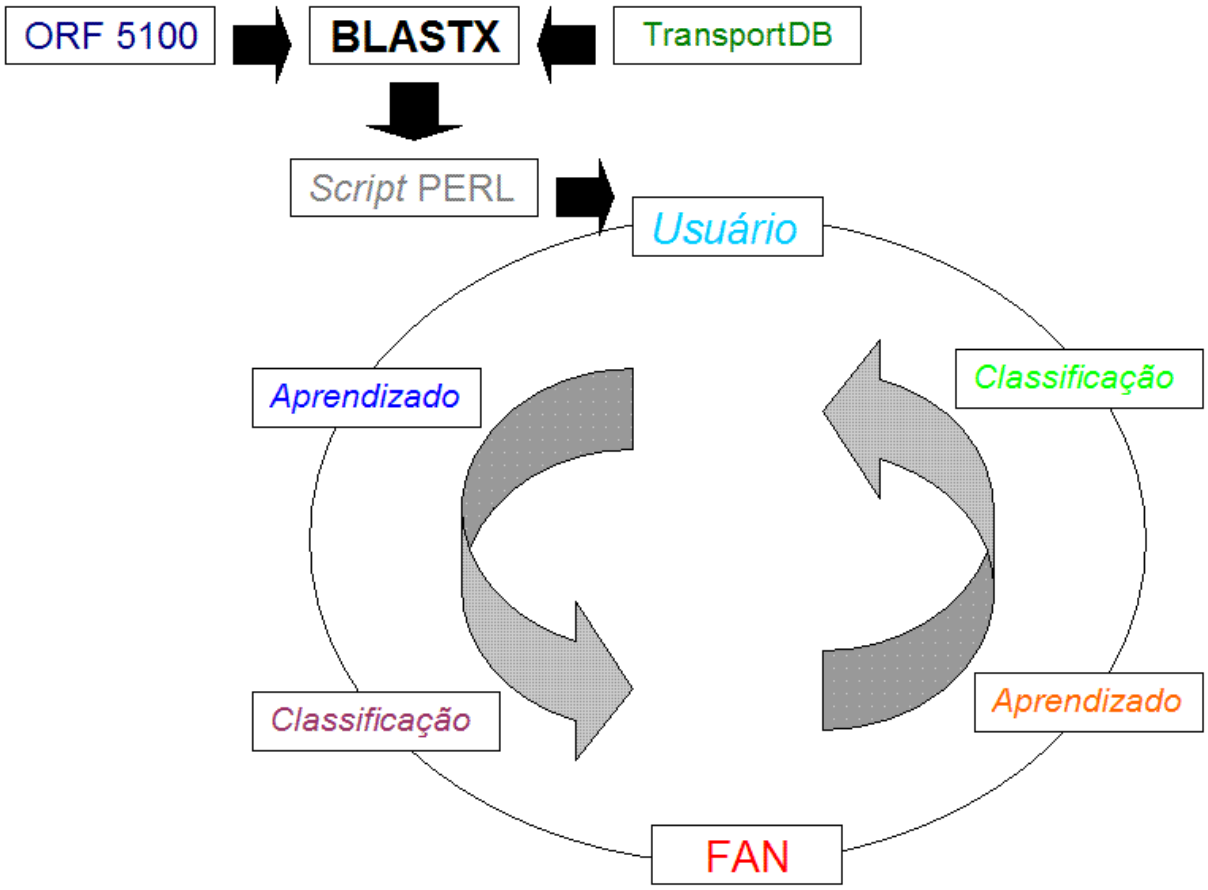
Na imagem, o menu à esquerda permite, entre outras coisas, que sejam importados arquivo para análise; os campos "Estratégias" e "Têmpera" permitem que alguns parâmetros da rede neuronal FAN sejam editados; o campo "Época" indica o número de vezes que a rede neuronal FAN leu o arquivo de treinamento; "Tempo de Treinamento" marca o tempo decorrido desde o início do treinamento; no campo "Estatísticas", os valores da direita são relativos ao padrão encontrado na "Época" atual, e os valores da esquerda são relativos ao melhor padrão encontrado pela rede.

Figura 2 – Gráficos que representam o aprendizado da FAN



Cada gráfico (neurônios) representa uma das classes: alta, média e baixa similaridade respectivamente. Cada uma das linhas coloridas representa uma característica do alinhamento, distribuídas do seu menor ao seu maior valor ("Range", eixo "x"; menores valores à esquerda, maiores à direita), e sua contribuição para a classe ("Valor", observar picos; maior pico, maior contribuição). Assim, para a FAN dizer se um alinhamento possui alta-similaridade (neurônio 1 ou primeiro gráfico), ela está usando valores altos das características "azul-clara" e "preta" (observar os picos dessas características na parte direita do "Range"), e valor intermediário da característica "rosa". Já para dizer se um alinhamento possui baixa-similaridade (neurônio 3 ou terceiro gráfico), a rede FAN usa valores baixos de diversas características, como "azul-escuro" e "amarelo" (observar picos à esquerda do "Range", que corresponde a valores baixos desse).

Figura 3 – Esquema representando o processo de “Coaprendizado”



O processo de “Coaprendizado” esquematizado na figura funciona da seguinte forma:

Após a realização de busca de similaridade com uso do programa “BLAST”, os dados são extraídos pelo “Usuário”. Com esses dados o “Usuário” realiza um “Aprendizado”, no qual aprende a distinguir as diferentes classes de alinhamentos (alta, baixa, média similaridade), permitindo-lhe fazer uma “Classificação”. Essa é enviada para a rede neuronal “FAN”, que aprende o padrão de “Classificação” do “Usuário” através de um novo “Aprendizado”, e faz uma nova “Classificação”.

O “Usuário” agora compara a sua “Classificação” com a “Classificação” da rede. Através de discrepâncias entre as duas, ele percebe que seu padrão de “Classificação” esteve incoerente em alguns casos, realizando então um novo “Aprendizado” com essa comparação. Dessa forma, sua “Classificação” é melhorada e reenviada para a “FAN”, que novamente passa por um processo de “Aprendizado” gerando uma nova “Classificação”.

Esse ciclo é repetido várias vezes, até o usuário e a rede concordarem (não houver discrepância de classificação), ou até a nova classificação realizada pelo usuário não acrescentar nada ao aprendizado da rede e vice-versa.

4.4. Classificação das proteínas transportadoras

Todas as ORF do genoma de *H. seropedicae* que apresentaram similaridade com o banco de dados TransportDB através de pesquisa BLAST, foram classificadas quanto à sua família e componente do sistema de transporte, quando pertinente (ex., grande parte dos transportadores do tipo ABC são formados por duas unidades de ligação a ATP, duas unidades transmembrana e uma unidade periplasmática de ligação ao substrato).

A classificação foi feita a partir da similaridade de seqüências de aminoácidos obtidas com o programa BLAST de acordo com a classificação apresentada no banco de dados TransportDB, TCDB e KEGG (através do programa KAAS).

O programa KAAS permitiu a identificação e classificação das ORF segundo o tipo de transportador (ex., transportador do tipo ABC para açúcar simples) e subunidade de transporte (ex., subunidade de ligação ao ATP).

4.5. Similaridade das proteínas transportadoras de H. seropedicae e H. rubrisubalbicans

As proteínas transportadoras encontradas em *H. seropedicae* foram submetidas à pesquisa BLAST contra *contigs* de *H. rubrisubalbicans.* Inicialmente foi realizado uma montagem das seqüências de *H. rubrisubalbicans,* gentilmente cedidas pelo Prof. Dr. Emanuel M. Souza, utilizando os programas PHRED (EWING & GREEN, 1998) para análise dos cromatogramas, CROSS_MATCH (http://www.phrap.org/phredphrap/general.html) para filtro de vetor e seqüências do operon rRNA, e o programa CAP3 (*Contig Assembly Program*) (HUANG & MADAN,

1999) para montagem de seqüências contíguas (*contigs*).

O programa CAP3 necessita de arquivos de seqüências no formato FASTA e qualidades correspondentes. O programa funciona em três etapas: na primeira, as pontas de baixa qualidade de cada seqüência são identificadas e retiradas, em seguida as sobreposições são identificadas e aquelas que são consideradas falsas são removidas; na segunda etapa, as seqüências são agrupadas em *contigs*; na terceira etapa é construído um alinhamento múltiplo das seqüências, gerando-se um consenso para os *contigs* assim como um valor para a qualidade destes *contigs* (HUANG & MADAN, 1999).

4.6. Análise de preferência de uso de códon

Todas as ORF anotadas pelo projeto GENOPAR foram submetidas à análise de códons, com auxílio dos softwares GCUA (*Graphical Codon Usage Analyser* – MCINERNEY, 1998) e CODONW (PENDEN, 1999). A análise permitiu estabelecer a preferência de uso de códon das ORF de *H. seropedicae* e sugerir uma classificação baseada no nível de expressão e tendência de uso de códon a partir do cálculo de diversos índices: CAI, ENc (ou Nc), etc. Uma análise comparativa foi feita usando-se as ORF identificadas como proteínas transportadoras.

O índice CAI (*Codon Adaptation Index*), é uma medida que relaciona o uso de códons por um gene ao uso de códons por genes altamente expressos, obtida através do uso relativo de um determinado códon em relação ao códon mais usado para um determinado aminoácido (SHARP & LI, 1987). Esse índice é derivado de estatísticas de preferência de códons, normalizadas para cada aminoácido (CAI Calculator – http://www.evolvingcode.net/codon/cai/cai.php), e refere-se à capacidade de expressão da proteína através dos códons por ela usados, variando seu valor de 0 a 1, onde quanto maior o valor, mais expressa é a proteína (SHARP & LI, 1987).

O índice ENc, ou Nc (*Effective Number of Codons*), é análogo ao número efetivo de alelos usado em genética de populações (WRIGHT, 1990) trabalhando com a probabilidade de dois códons escolhidos ao acaso serem idênticos (POWELL & MORIYAMA, 1997), e refere-se ao quão aleatório é o uso de códons sinônimos pelo gene, sendo que seu valor varia de 20 a 60, e quanto mais alto, mais aleatório é o uso de códons sinônimos pelo gene (WRIGHT, 1990).

4.7. Análise de transportadores da família ABC

Os transportadores da família ABC estão envolvidos na translocação de uma grande variedade de substratos, tais como íons, açúcares, aminoácidos, vitaminas, lipídios, antibióticos, drogas, oligossacarídeos, oligopeptídeos e até proteínas (BIEMANS-OLDEHINKEL et al., 2006).

A superfamília ABC é a mais abundante família encontrada em *Bdellovibrio bacteriovorus* (BARABOTE et al., 2007) e também é o principal sistema de transporte encontrado em *Chromobacterium violaceum* (GRANGEIRO et al., 2004).

Devido à sua importância, as ORF identificadas como relativas à subunidades de transportadores ABC foram agrupadas conforme a "via de transporte" (*pathway*), nas quais foram classificadas segundo a montagem de vias realizada pelo programa KAAS.

As ORF somente foram agrupadas quando verificou-se proximidade entre elas num mesmo *contig.* Com isso, ORF de subunidades que compõe uma mesma via de transporte ABC, e próximas no genoma de *H. seropedicae,* foram consideradas indícios de formação de *operons*.

A idéia de que os transportadores ABC formem *operons*, provem do fato de que o sistema de transporte, dependente de proteína ligadora periplásmica, é uma subfamília da superfamília dos transportadores ABC, que podem ser subdivididos em 8 grupos relacionados filogeneticamente (SAURIN & DASSA, 1994). Alguns autores sugerem que a evolução destes grupos se deu antes da divergência entre os grupos de bactérias e que, devido à semelhança na filogenia dos componentes protéicos, é provável que todo o sistema tenha co-evoluído (SAURIN & DASSA,

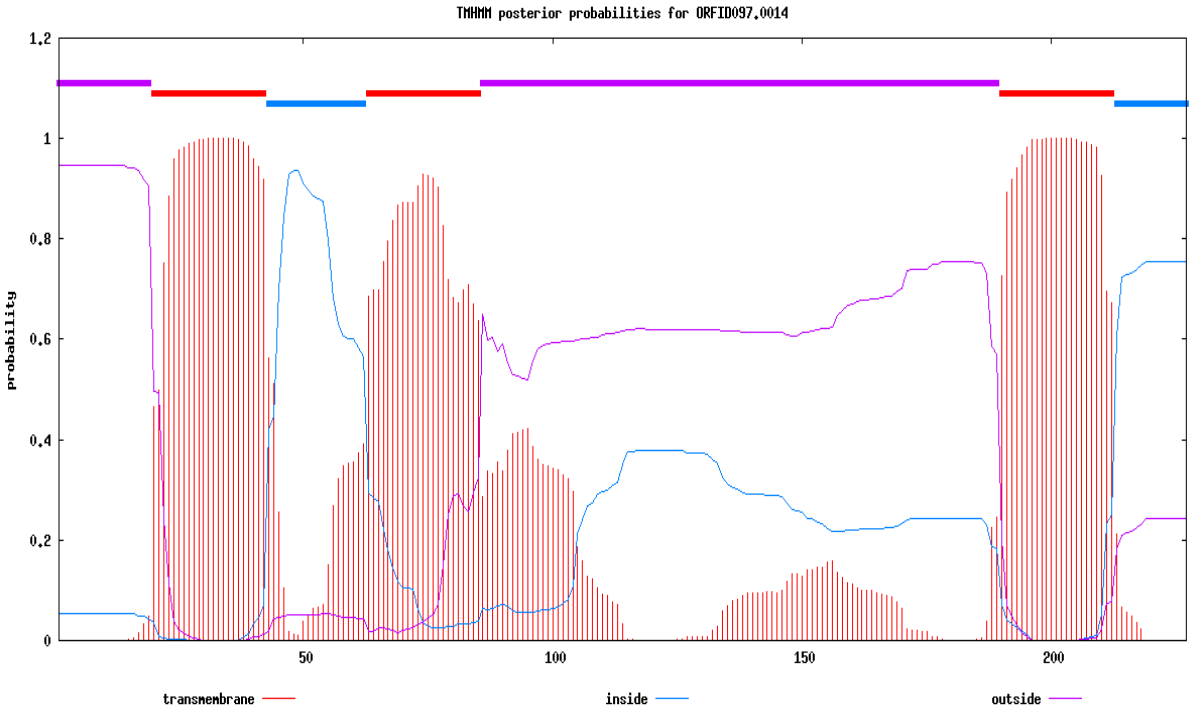
1994).

4.7.1. Análise de domínios transmembrana

Com o objetivo de verificar se as proteínas da superfamília ABC identificadas possuem hélices transmembrana, foi utilizada o programa TMHMM (*transmembrane hidden Markov model* - KROGH et al., 2001), o qual identifica possíveis regiões formadoras dessas hélices a partir da seqüência de aminoácidos de uma proteína, através da análise de hidrofobicidade da seqüência, polarização de carga, tamanho das hélices, e restrições num modelo com estimativas e predições já existentes (KROGH et al., 2001).

Um exemplo de gráfico gerado pelo programa mostrando as regiões transmembrana pode ser visto na figura 4.

Figura 4 – Gráfico gerado pelo programa TMHMM



A figura mostra a probabilidade de uma determinada região (no eixo "y") ao longo da proteína (eixo "x") ser hélice transmembranar. As linhas verticais vermelhas indicam possíveis hélices transmembrana, as quais são validadas pela presença de uma barra horizontal vermelha na parte superior do gráfico. A linha azul, bem como as barras azuis na parte superior do gráfico indicam regiões voltadas para o lado citoplasmático; as linhas e barras roxas indicam regiões que estão voltadas para o lado externo da célula.

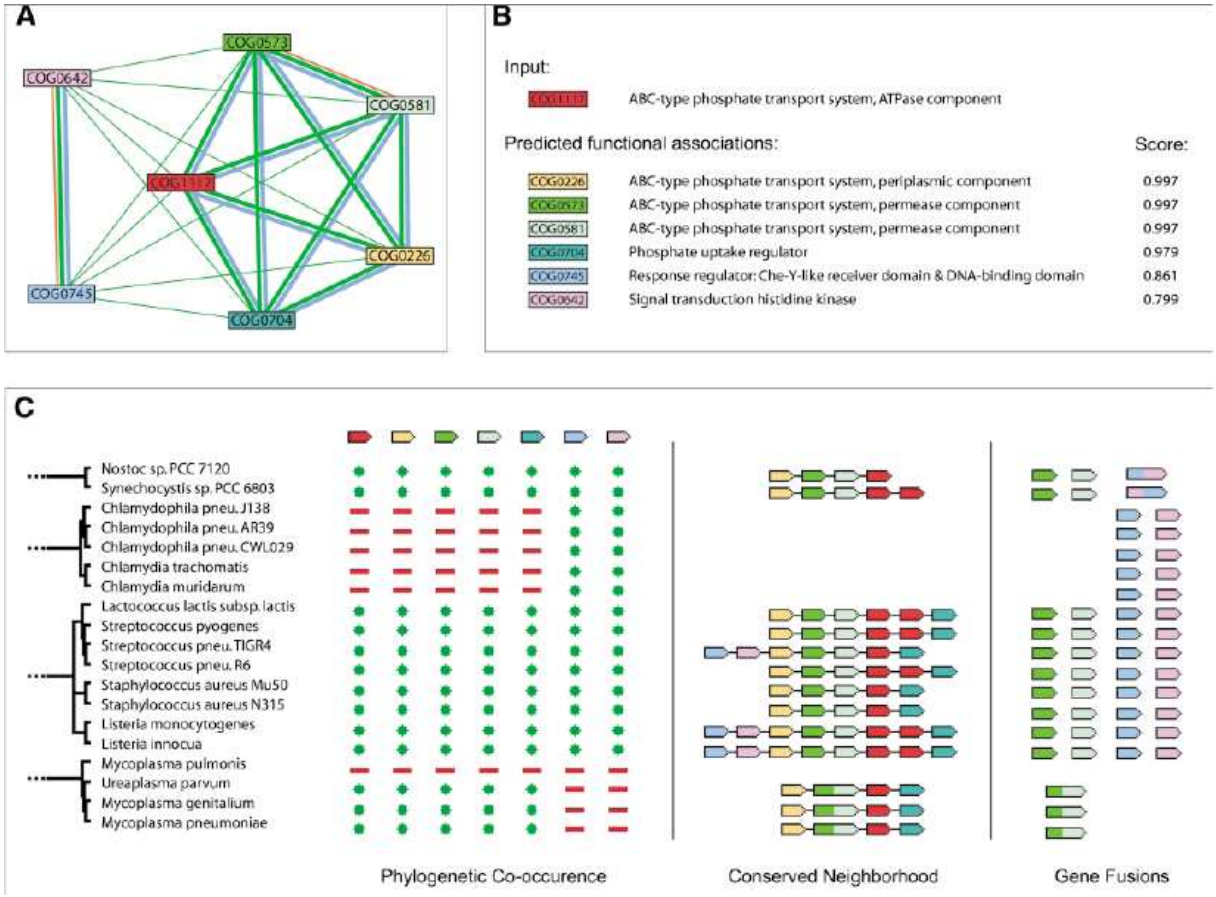
4.7.2. Similaridade e conservação da "vizinhança" entre os genes de transportadores ABC em *H. seropedicae* e bactérias relacionadas

As ORF identificadas como subunidades da superfamília ABC e que foram agrupadas em possíveis *operons*, a partir da sua localização nos *contigs* do genoma de *H. seropedicea,* e segundo as vias de transporte montadas pelo programa KAAS, foram submetidas à análise pelo programa STRING – *Search Tool for the Retrieval of Interacting Proteins* (VON MERING et al., 2007).

Essa análise foi realizada visando obter-se maiores informações sobre a relação dos genes relativos a essas subunidades em outros organismos, tais como coocorrência, "vizinhança" (*neighborhood*), fusão gênica e coexpressão, fortalecendo os indícios de formação de *operons*.

O programa STRING verifica a existência desses indícios entre um grupo de sequências de proteínas submetidas, e os mostra de maneira gráfica (VON MERING et al., 2003) (figura 5).

Figura 5 – Exemplos de resultados produzidos pelo programa STRING (VON MERING et al., 2003)



A figura mostra alguns resultados produzidos pelo programa String para uma proteína da família ABC que participa do transporte de fosfato (a proteína é um exemplo e não é nenhuma ORF de H. seropedicae). Em "A" está representada uma rede de associação entre a proteína submetida (vermelho) e as que estão relacionadas a ela em alguns genomas disponíveis; os diferentes tipos de relações são mostrados pelas diferentes cores das linhas: verde indica vizinhança entre as proteínas em alguns genomas disponíveis, azul indica coocorrência filogenética dessas proteínas, e vermelho indica que os genes relativos a essas proteínas podem estar fundidos em alguns genomas. Pode-se notar uma relação maior em outros genomas entre as cinco proteínas da direita, que são componentes estruturais, e entre as duas da esquerda, que são reguladores. Em "B" está representado um sumário dos scores obtidos; os maiores scores de associação ocorrem entre as proteínas estruturais. Em "C" são mostradas em maior detalhe as relações entre as proteínas; primeiro a coocorrência filogenética dessas proteínas em genomas disponíveis (indicada pelos pontos verdes), depois a vizinhança dos genes correspondentes a essas proteínas, e por último as evidências de fusão gênica (VON MERING et al., 2003).

4.7.3. Análise filogenética

As ORF de transportadores ABC identificadas como subunidade de ligação a ATP, por serem as subunidades mais conservadas (TOMMI & KANEHISA, 1998), foram submetidas à análise filogenética.

Esta análise foi realizada utilizando-se 60 ORF traduzidas para proteínas correspondentes às subunidades, bem como as proteínas mais similares a essas, identificadas através de pesquisa de similaridade com o programa BLASTX *on line,* presente no banco de dados do *site* NCBI – *National Center for Biotechnology Information* (http://www.ncbi.nlm.nih.gov/). A pesquisa BLAST foi editada para mostrar somente os dez melhores alinhamentos, através dos quais as proteínas similares foram identificadas e extraídas. Um total de 655 seqüências de proteínas foram utilizadas.

A análise filogenética foi realizada com uso do programa MEGA4 (TAMURA et al., 2007). Esse programa permitiu às seqüências das ORF serem alinhadas entre si e com suas similares, utilizando-se o programa ClustalW (THOMPSON et al., 1994) presente nele, bem como às distâncias genéticas, para montagem da árvore filogenética, serem calculadas com a matriz de substituição PAM (*Point Accepted Mutation* – Dayhoff et al., 1978), e à árvore ser obtida pelo método de *Neighbour-Joining* (SAITOU & NEI, 1987).

4.8. Automação das etapas de análise

Cada uma das etapas, por envolver um grande número de seqüências e análises, foram realizadas com o auxílio de programas específicos, os quais exigem uma formatação própria dos dados. Os resultados gerados por eles também necessitam de formatação adequada para facilitar a análise.

Scripts foram desenvolvidos para que essas análises fossem conduzidas de forma contínua e com o mínimo de interferência humana possível, permitindo uma integração entre a execução dos diversos programas usados. Alguns *scripts* podem ser vistos no Apêndice B.

Estes *scripts* foram desenvolvidos em linguagem de programação BASH (*Bourne Again SHell*) e PERL (*Practical Extraction and Report Language*) para plataforma de sistemas baseados em UNIX.

A programação BASH é nativa de sistemas LINUX e permite executar tarefas e programas automaticamente através do Sistema Operacional. Essa linguagem possui ainda a vantagem de ser portável a qualquer sistema UNIX, sem a necessidade de instalação ou adaptação do sistema e programas, exceto aqueles específicos para Bioinformática.

A linguagem PERL também foi usada por sua facilidade e versatilidade na análise de arquivos e padrões de texto. Por ser também uma linguagem historicamente usada em Bioinformática, muitos *scripts* para análises de seqüências foram desenvolvidos em código aberto, permitindo que fossem livremente adquiridos e modificados.

5. RESULTADOS E DISCUSSÃO

5.1. Identificação das proteínas transportadoras

A identificação das ORF feita através do programa BLASTX teve seus resultados analisados e visualmente inspecionados. Uma análise da qualidade dos alinhamentos obtidos contra o banco de dados TransportDB é apresentada na figura 6, mostrando a cobertura (proporção de tamanho) obtida pelo comprimento total da *query* (seqüência para pesquisa) em relação ao *subject* (seqüências do banco de dados).

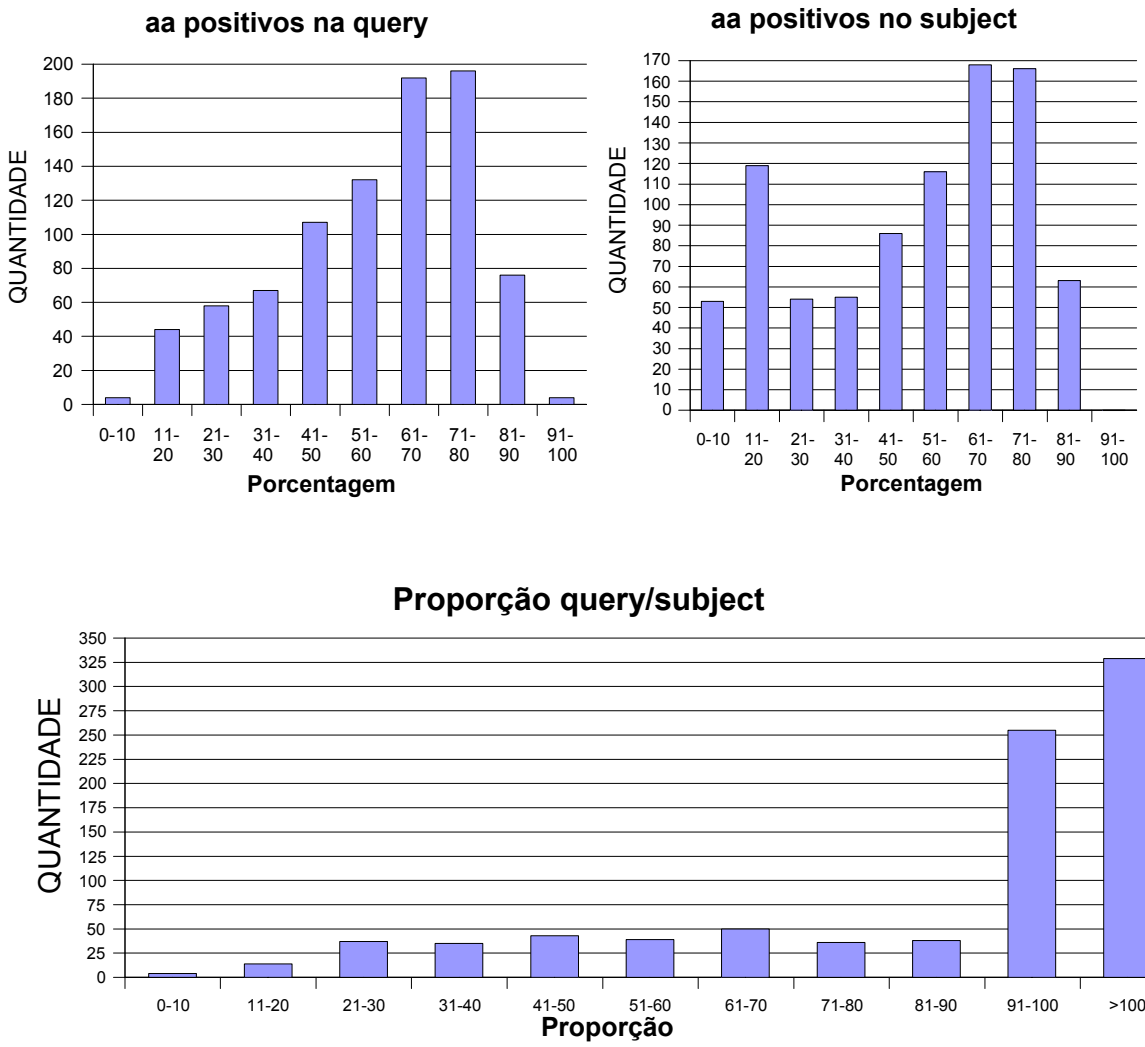
Ao final dessa análise a maior parte dos alinhamentos (em torno de 575 ORF de 880) produziram cobertura superior a 90% do comprimento total da *query*, em relação ao comprimento total do *subject*. Isso indica similaridade entre toda a extensão das proteínas e não somente entre domínios conservados, servindo como um ponto positivo para validar de modo geral a pesquisa BLAST realizada.

Um número de 880 ORF no genoma de *Herbaspirillum seropedicae* apresentou algum alinhamento na análise realizada com o programa BLASTX contra o banco de dados TransportDB. Deste total, 195 ORF foram classificadas como "pendentes" e 154 constam como *frameshift* (mudança na fase de leitura; 144 pendentes e *frameshift*), segundo a anotação do GENOPAR (figura 7).

A identificação de ORF para possíveis proteínas transportadoras foi variável nas análises para o BLASTX contra o banco de dados TCDB e também quando foi usado o programa KAAS (tabela 1). As ORF identificadas pela análise do BLASTX contra o banco de dados TransportDB foram usadas como padrão para outras

análises, bem como a classificação da rede neuronal FAN para esse conjunto de ORF, como será visto adiante.

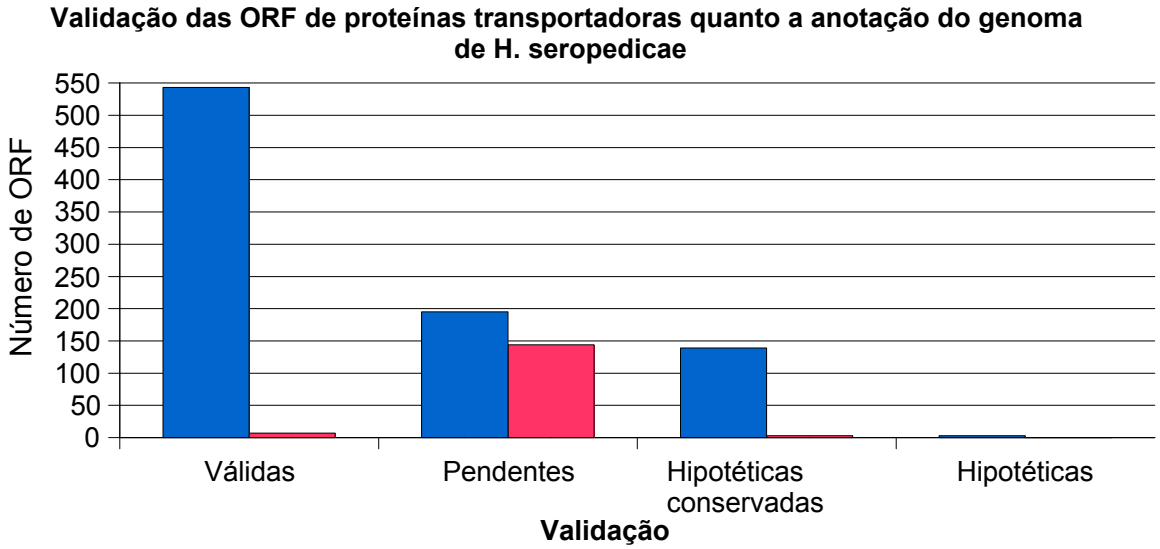
Figura 6 – Resultados para busca de similaridade entre as ORF de *H. seropedicae* contra o banco de dados de proteínas transportadoras TransportDB



aa – aminoácidos;
Porcentagem – porcentagem de aminoácidos positivos em relação ao comprimento total da "query" ou "subject";
Proporção – proporção entre os comprimentos das "query" em relação aos comprimentos dos "subject";
Quantidade – quantidade de ORF (alinhamentos).

Os dois primeiros gráficos mostram os números de aminoácidos positivos do alinhamento em relação aos tamanhos totais da "query" e do "subject". Pode-se notar que a maior parte dos alinhamentos possuem uma porcentagem entre 50 e 80% de aminoácidos positivos, tanto em relação ao tamanho total da "query" quanto do "subject", um valor de similaridade considerado alto.
O terceiro gráfico mostra a proporção de tamanhos entre a "query" e o "subject". O maior número de alinhamentos possui uma proporção de tamanho das duas seqüências superior a 90%, sendo grande parte das "queries" maiores que os "subjects", indicando grande similaridade entre as proteínas submetidas e as proteínas presentes no banco de dados, em toda a sua extensão e não somente em domínios conservados.

Figura 7 – Validação das ORF de possíveis proteínas transportadoras segundo a anotação do GENOPAR



As barras azuis indicam o número de ORF identificadas com pesquisa BLAST (880) distribuídas em cada classe de validação segundo o projeto GENOPAR; as barras vermelhas indicam a quantidade de ORF com frameshift em cada uma das classes. Pode-se notar uma proximidade numérica entre o número de ORF válidas segundo o projeto GENOPAR para essas 880 ORF, que é de 543, em relação ao número de ORF identificadas com uso dos três bancos de dados em conjunto: TransportDB, TCDB e KEGG (através da ferramenta KAAS), que é de 537 (tabela 1A). Aos dois conjuntos, são comuns 424 ORF (tabela 1B).

Tabela 1 – A: Número de ORF identificadas como possíveis proteínas transportadoras através do programa BLASTX contra os bancos de dados TransportDB e TCDB e com o programa KAAS

A	TransportDB	TCDB	KAAS	TCDB + KAAS
TransportDB	880	721	590	537[1]
TCDB		1220	837	
KAAS			2652[2]	

[1] O número mostrado corresponde às ORF identificadas pelas três análises.
[2] O número mostrado corresponde a todas as ORF identificadas pela ferramenta KAAS, não somente aquelas para possíveis proteínas transportadoras.

B: Número de ORF identificadas pelas ferramentas em conjunto

B	TransportDB	+ TCDB	+ KAAS	+ GENOPAR*	+ FAN**
TransportDB	880	721	537	424	370

**Considerando apenas as ORF anotadas como "válidas"*
***O sinal de " + " indica a ferramenta de análise em adição às ferramentas anteriores.*

A tabela "A" mostra diferença no número de proteínas transportadoras identificadas nas três pesquisas BLAST realizadas (considerando o KAAS, que também funciona através de pesquisa BLAST). O número encontrado pelo KAAS é o mais alto, porém deve-se lembrar que o banco de dados utilizado por ele, possui diversos tipos de proteínas e não somente proteínas transportadoras. Na última coluna está o número de proteínas encontradas nas três análises (537), número próximo ao dessas ORF que estão anotadas como válidas segundo o projeto GENOPAR (543).
A tabela "B" mostra que 424 das 537 ORF identificadas pelas três pesquisas BLAST estão anotadas como válidas segundo o projeto GENOPAR.

5.2. Uso de rede neuronal para validação das ORF encontradas como possíveis proteínas transportadoras

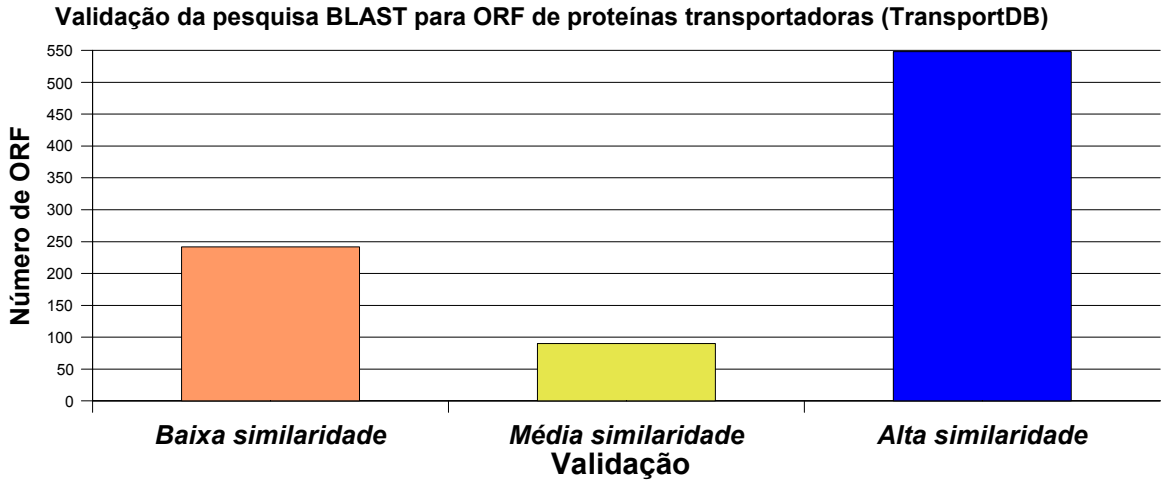
As 880 ORF identificadas como possíveis proteínas transportadoras foram classificadas quanto à "qualidade" de seus alinhamentos, produzidos pelo programa BLASTX contra o banco de dados TransportDB. A classificação foi realizada através do treinamento da rede neuronal FAN, sendo os alinhamentos agrupados nas categorias: "alta similaridade", "média similaridade" ou "baixa similaridade".

O método usado foi de "coaprendizado" (ver *Uso de rede neuronal para validação das ORF encontradas como possíveis proteínas transportadoras* em Material e Métodos), e a classificação gerada pode ser visualizada na figura 8. A tabela 2 mostra exemplos da classificação realizada.

Do total de 880 ORF de possíveis proteínas transportadoras identificadas, 548 (62,3%) ORF foram classificadas como "alta similaridade" em relação às características do alinhamento. Esse número aproxima-se do número de ORF válidas segundo anotação do projeto GENOPAR, que é de 543 ORF, e das 537 ORF identificadas no conjunto de três bancos de dados usados: TransportDB, TCDB, KEGG (por análise com o programa KAAS).

Das 880 ORF de possíveis proteínas transportadoras encontradas em *H. seropedicae,* 370 foram validadas por todas as ferramentas usadas, cujos resultados podem ser visualizados na tabela 1B.

Figura 8 – Classificação dos alinhamentos obtidos através de pesquisa BLAST contra o banco de dados TransportDB com uso de rede neuronal FAN



O gráfico mostra 548 de 880 ORF possuindo alta similaridade com proteínas presentes num banco de dados de proteínas transportadoras (TransportDB). Esse número está próximo às 543 das 880 ORF que são válidas segundo o projeto GENOPAR, e ao de ORF encontradas nos três bancos de dados usados, TransportDB, TCDB e KEGG (pelo programa KAAS).

Tabela 2 – Exemplo de classificação de alinhamentos produzidos pelo programa BLAST, realizada com rede neuronal FAN em coaprendizado com usuário

QUERY ID	TAMANHO QUERY (aa)	TAMANHO SUBJECT (aa)	TAMANHO ALINHAMENTO	ALINHAMENTO/ TAMANHO QUERY	ALINHAMENTO/ TAMANHO SUBJECT	POSITIVOS/ TAMANHO QUERY	POSITIVOS/ TAMANHO SUBJECT	PROPORÇÃO	CLASSE
ORFID089.0004	49	1049	42	0,86	0,04	0,78	0,04	0,05	Baixa-similaridade
ORFID121.0015	211	1245	158	0,75	0,13	0,46	0,08	0,17	Baixa-similaridade
ORFID122.0009	206	303	180	0,87	0,59	0,72	0,49	0,68	Média-similaridade
ORFID122.0012	86	303	85	0,99	0,28	0,93	0,26	0,28	Média-similaridade
ORFID122.0024	507	500	502	0,99	1	0,8	0,82	1,01	Alta-similaridade
ORFID129.0007	242	244	240	0,99	0,98	0,84	0,84	0,99	Alta-similaridade
ORFID129.0023	405	378	380	0,94	1,01	0,62	0,67	1,07	Alta-similaridade
ORFID130.0003	74	654	62	0,84	0,09	0,64	0,07	0,11	Baixa-similaridade
ORFID130.0011	476	656	392	0,82	0,6	0,55	0,4	0,73	Média-similaridade

Na tabela, alinhamentos considerados como "baixa-similaridade" possuem valores baixos, sobretudo em relação ao "subject", enquanto os considerados como possuindo "alta-similaridade" são o oposto. A característica "proporção" indica a proporção entre os tamanhos da "query" e "subject".

5.3. Comparação de proteínas transportadoras nos genomas de H. seropedicae e de outras bactérias

A porcentagem de ORF de proteínas transportadoras (17,1%, considerando 880 ORF de proteínas transportadoras do total de 5100 ORF) encontrados no genoma de *H. seropedicae,* através da análise com o programa BLASTX contra o banco de dados TransportDB, é alto quando comparado a outras bactérias (tabelas 3 e 4). Já no resultado da análise utilizando a rede neuronal FAN, 548 ORF foram classificadas como de "alta similaridade". Esse número corresponde a aproximadamente 10,74% do total de ORF do genoma de *H. seropedicae*.

Informações sobre o genoma e proteínas transportadoras de organismos seqüenciados são apresentadas nas tabelas 3 e 4. Há uma grande variação no número dessas proteínas presentes no genoma de bactérias, mas que está, até certo ponto, correlacionado com o tamanho do genoma (figura 9). Essa relação permite observar que o número de possíveis proteínas transportadoras encontrados em *H. seropedicae* é alto, considerando o número de proteínas transportadoras por Mb do genoma (96,14 – para um tamanho de genoma estimado em 5,7Mb).

Em *Chromobacterium violaceum* foi encontrado um total de 489 ORF para possíveis proteínas transportadoras, o que corresponde a 11,1% do total de ORF (GRANGEIRO et al., 2004). A distribuição dessas proteínas também diferiu bastante nestes organismos, como mostra a tabela 5.

Tabela 3 – Comparação do número de proteínas transportadoras em diferentes organismos

	Média	SD[1]	Máximo	Mínimo
Bacteria (150 genomas)				
Genoma (Mpb)	3,19	1,89	9,11	0,58
Proteínas transportadoras	172,91	120,43	548	9
P.Transportadoras/Mb [2]	51,69	16,62	100,67	9,68
Arqueobactérias (19 genomas)				
Genoma (Mpb)	2,2	1,14	5,75	0,5
Proteínas transportadoras	98,26	44,65	215	16
P.Transportadoras/Mb	46,09	14,13	73,55	21,3
Eucariotos (12 genomas)				
Genoma (Mpb)	303,62	897,45	3150	2,5
Proteínas transportadoras	404,75	295,12	855	42
P.Transportadoras/Mb	10,52	7,03	24,92	0,26

[1] SD – Desvio padrão
[2] Transportadores/Mb – razão entre o número de proteínas transportadoras encontradas e o tamanho do genoma em megabases

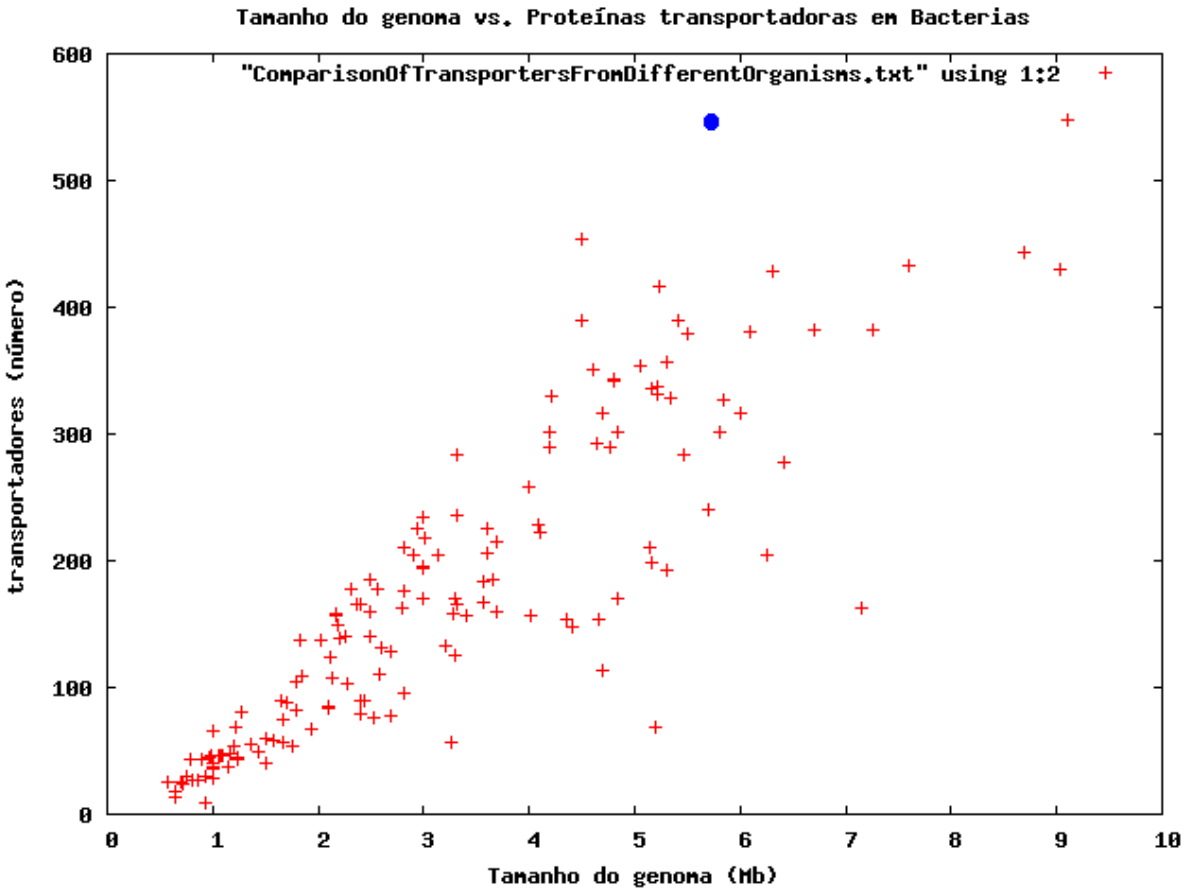
A tabela mostra que a média de proteínas transportadoras (ou genes relativos a essas proteínas) em bactéria, fica em torno de 173, sendo o desvio padrão bastante elevado. O número máximo de proteínas transportadoras é 548, número igual ao de ORF de proteínas transportadoras de H. seropedicae classificadas como "alta-similaridade" pela rede neuronal FAN.

Tabela 4 – Comparação entre proteínas transportadoras presentes no genoma de algumas Proteobacteria relacionadas a *H. seropedicae*

Organismo	Genoma (Mpb)	No. de genes	Proteínas Transportadoras	% dos genes
Herbaspirillum seropedicae	~5,7	5.100	548	10,74
Burkholderia mallei	5,8	5.025	327	6,50
Burkholderia pseudomallei	7,2	5.855	382	6,52
Ralstonia solanacearum	5,8	5.120	301	5,88

Pode-se notar que o número de possíveis proteínas transportadoras encontradas em H. seropedicea é elevado, maior até mesmo do que o encontrado em Burkholderia pseudomallei, que possui um genoma bem maior.

Figura 9 – Proporção de ORF totais e número de ORF de proteínas transportadoras em relação ao tamanho dos genomas (Mb)



Pode-se observar uma relação entre o tamanho do genoma e o número de proteínas transportadoras encontradas, na qual, quanto maior o tamanho do genoma, maior é o número de proteínas transportadoras encontradas. H. seropedicae corresponde ao ponto azul no alto do mapa; pode-se notar que esse ponto está deslocado em relação aos demais, ou seja, o número de proteínas encontradas em H. seropedicae é alto em relação ao tamanho do seu genoma.

Tabela 5 – Classificação geral das possíveis proteínas transportadoras presentes no genoma de *H. seropedicae*

CLASSE DE PROTEÍNA TRANSPORTADORA	NÚMERO DE ORF	
	H. seropedicae	*C. violaceum*
Poros e canais	16 (1,8%, 0,3%)	62 (12,7%, 1,4%)
Transportadores movidos por potencial eletroquímico	263 (29,9%, 5,1%)	154 (31,5%, 3,5%)
Transporte ativo primário	553 (62,8%, 10,8%)	212 (43,4%, 4,8%)
Translocador de grupo	0	
Carreadores de elétrons transmembrana	0	
Fatores acessórios envolvidos no transporte	0	
Sistemas de transporte incompletamente caracterizados	8 (0,9%, 0,1%)	
Não classificados	40 (4,0%, 0,8%)	61 (12,4%, 1,4%)[1]
TOTAL	880 (100%, 17,1%)	489 (100%, 11,1%)

[1]Esse número inclui os outros sistemas de transporte que não são canais/poros, transporte movido por potencial eletroquímico ou transporte primário.

Na tabela, os números de transportadores primários e secundários (movidos por potencial eletroquímicos) encontrados para as duas bactérias, mantêm uma proporção relativa à proporção entre o número total de proteínas transportadoras para essas duas bactérias. Os números de canais/poros não segue essa proporção.

5.4. Classificação das proteínas transportadoras

As 880 ORF de possíveis proteínas transportadoras, foram classificadas em famílias, segundo a classificação encontrada no *site* TransportDB, e em classes segundo o Sistema de Classificação de Transportadores (*Transport Classification (TC) System*) aprovado pelo NC-IUBMB. A quantidade de ORF correspondentes às famílias e classes encontradas são apresentadas nas tabelas 6 e 5 respectivamente.

Os mapas obtidos pelo programa KAAS permitiram identificar vários sistemas de transporte específicos e suas sub-unidades protéicas presentes no genoma de *H. seropedicae*. Os mapas construídos para os sistemas de secreção do tipo II e III, proteínas de excreção, sistema fosfotransferase e transportadores ABC podem ser visualizados na figura 10. O mapa para montagem de flagelo também é mostrado, devido à alta homologia que suas proteínas apresentam em relação aos Sistemas de Secreção do Tipo III (GRANGEIRO et al., 2004).

Somente 2 ORF para sistema PTS geral (GPTS) e 3 ORF para sistema PTS específico para açúcar (SSPTS) foram encontrados em *H. seropedicae* (tabela 6). A análise com o programa KAAS identificou a proteína transportadora de fosfato PtsH, um componente chave do sistema PTS de procariotos (figura 10E) e ainda duas subunidades da enzima II, as proteínas SgaA, que participa do transporte de L-ascorbato, e PtsN, a qual participa de processo envolvendo nitrogênio (figura 10E).

A análise dos mapas para o sistema PTS das bactérias presentes no banco de dados KEGG indica a presença de um grande número de componentes do sistema PTS em alguns grupos de bactérias, como nas enterobactérias (ex., *Escherichia coli*, *Salmonella enterica*, etc.), mas poucos componentes em outros

organismos, relacionados a *H. seropedicae*, como *Xylella fastidiosa*, *Xanthomonas campestris*, *Neisseria meningitidis*, *Ralstonia solanacearum*, *Burkholderia pseudomallei*, etc.

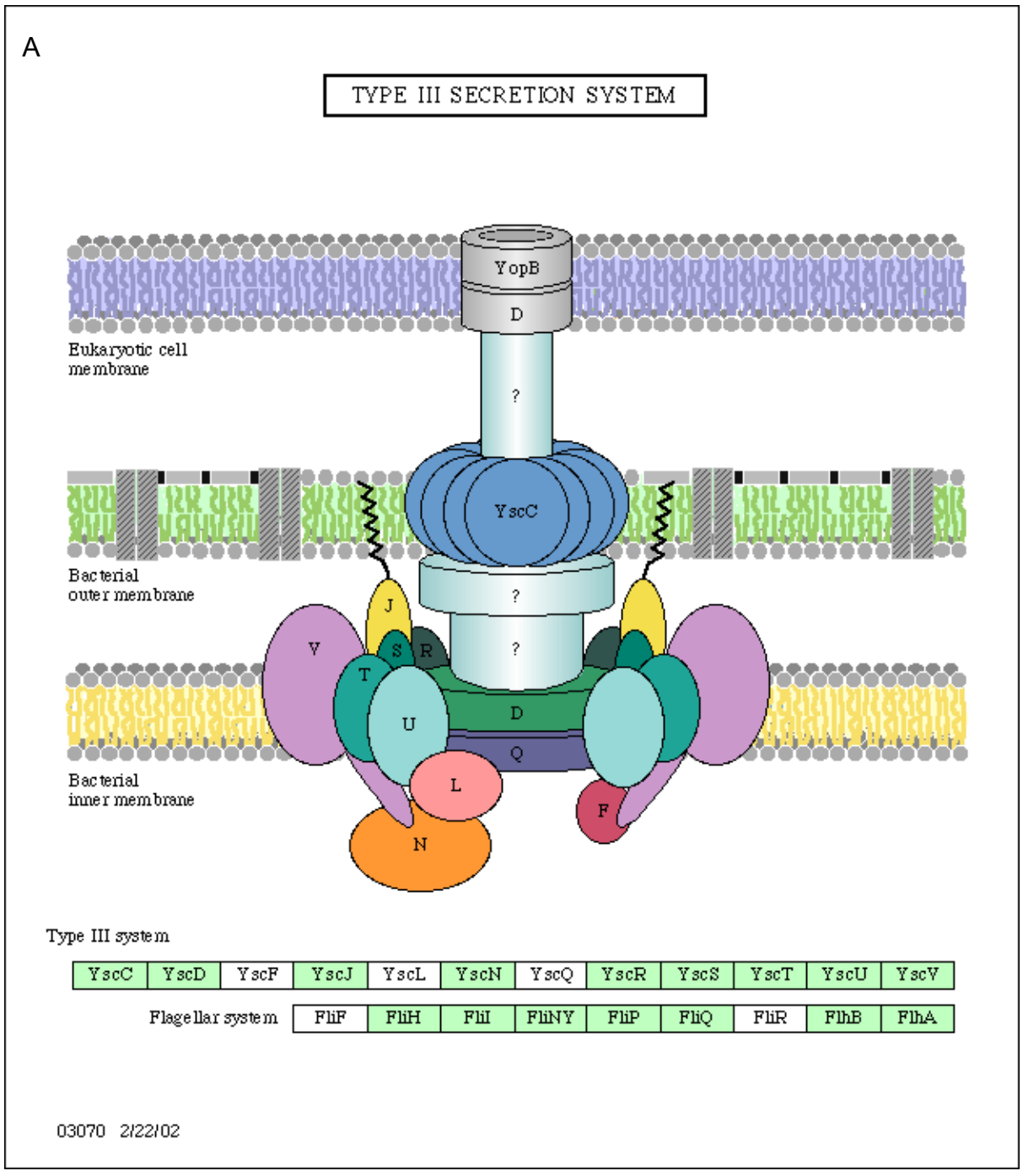
Tabela 6 – Número de ORF encontradas distribuídas em suas respectivas famílias, conforme classificação encontrada no *site* TransportDB

FAMÍLIA	NÚMERO DE ORF ENCONTRADAS	FAMÍLIA	NÚMERO DE ORF ENCONTRADAS	FAMÍLIA	NÚMERO DE ORF ENCONTRADAS
ABC	528 (60%; 10%)[1]	**ArsB**	1 (0,1%; 0,02%)	**F-ATPase**	11 (1,2%; 0,2%)
MFS	105 (12%; 2%)	**H+-PPase**	1 (0,1%; 0,02%)	**Tat**	5 (0,6%; 0,1%)
MscS	4 (0,4%; 0,1%)	**GPTS**	2 (0,2%; 0,04%)	**APC**	4 (0,4%; 0,1%)
CIC	8 (0,9%; 0,2%)	**DASS**	1 (0,1%; 0,02%)	**GPH**	1 (0,1%; 0,02%)
BenE	1 (0,1%; 0,02%)	**PnuC**	1 (0,1%; 0,02%)	**MOP**	6 (0,7%; 0,1%)
P-ATPase	13 (1,5%; 0,3%)	**MIT**	4 (0,4%; 0,1%)	**Amt**	19 (2,1%; 0,4%)
CHR	2 (0,2%; 0,04%)	**BASS**	3 (0,3%; 0,06%)	**CPA**	10 (1,1%; 0,2%)
NCS2	2 (0,2%; 0,04%)	**GntP**	3 (0,3%; 0,06%)	**SSPTS**	3 (0,3%; 0,06%)
SSS	2 (0,2%; 0,04%)	**FeoB**	5 (0,6%; 0,1%)	**LIV-E**	1 (0,1%; 0,02%)
AEC	2 (0,2%; 0,04%)	**NhaA**	2 (0,2%; 0,04%)	**PiT**	2 (0,2%; 0,04%)
TTT	3 (0,3%; 0,06%)	**RhtB**	5 (0,6%; 0,1%)	**KUP**	1 (0,1%; 0,02%)
DAACS	3 (0,3%; 0,06%)	**NCS1**	1 (0,1%; 0,02%)	**AAA**	4 (0,4%; 0,1%)
DMT	20 (2,3%; 0,4%)	**SulP**	2 (0,2%; 0,04%)	**CitMHS**	1 (0,1%; 0,02%)
RND	31 (3,5%; 0,6%)	**Nramp**	1 (0,1%; 0,02%)	**OFeT**	1 (0,1%; 0,02%)
TRAP-T	14 (1,6%; 0,3%)	**MerTP**	1 (0,1%; 0,02%)	**CDF**	3 (0,3%; 0,06%)
Oxa1	1 (0,1%; 0,02%)	**OPT**	1 (0,1%; 0,02%)	**Outras(proteínas de membrana)**	35 (4%; 0,7%)
				TOTAL	**880**

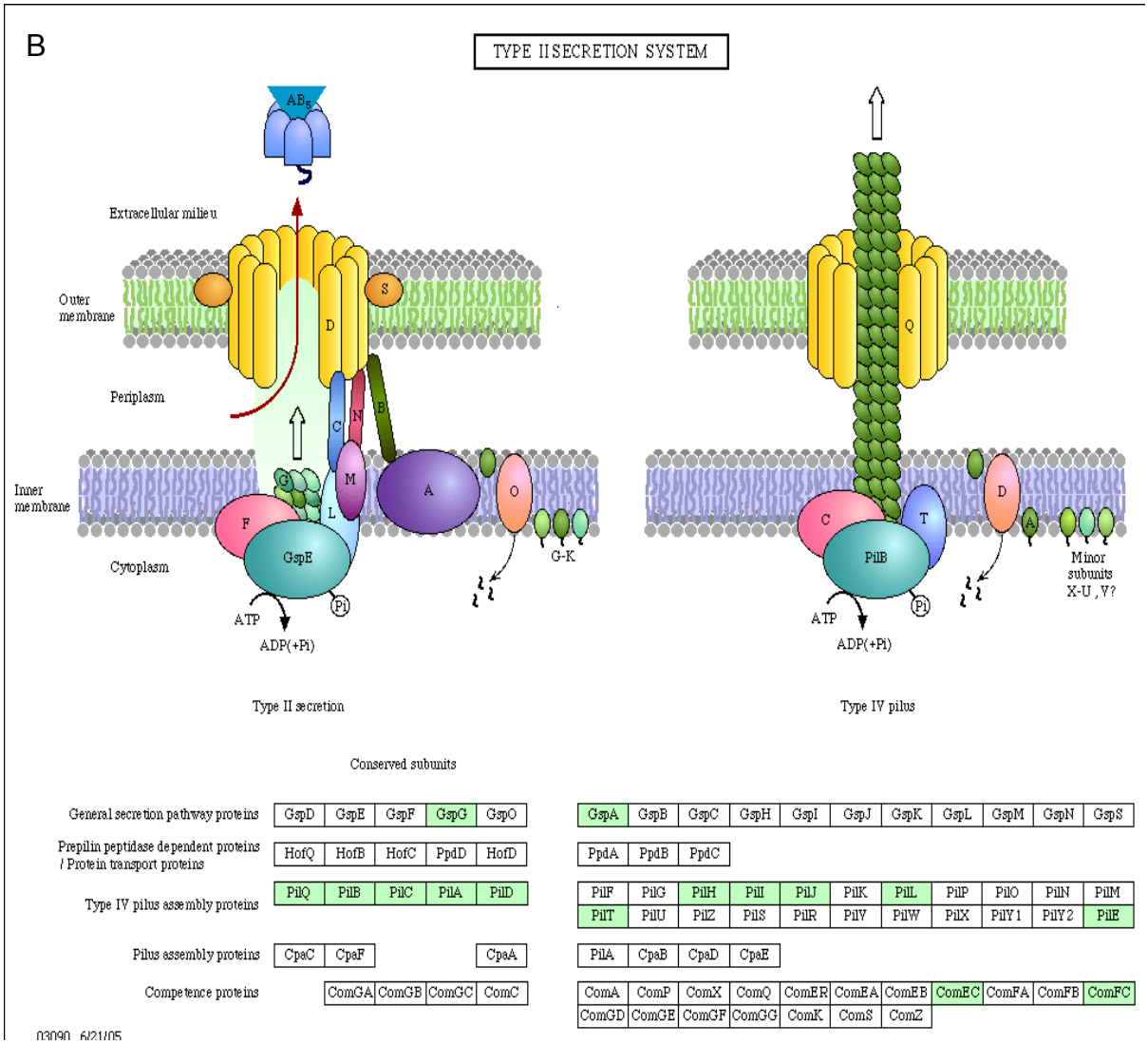
[1] Número de ORF encontradas; entre parênteses: porcentagem em relação ao total de 880 ORF de possíveis transportadores e porcentagem em relação ao total de 5.100 ORF do genoma de H. seropedicae, respectivamente.

Na tabela, a família com o maior número de transportadores em H. seropedicae é a família ABC, também encontrada em grande quantidade em outros organismos. Outras também destacam-se pelo número, como as famílias MFS, DMT e RND, que são transportadores secundários e transportam, entre outros substratos, drogas e/ou metabólitos tóxicos.

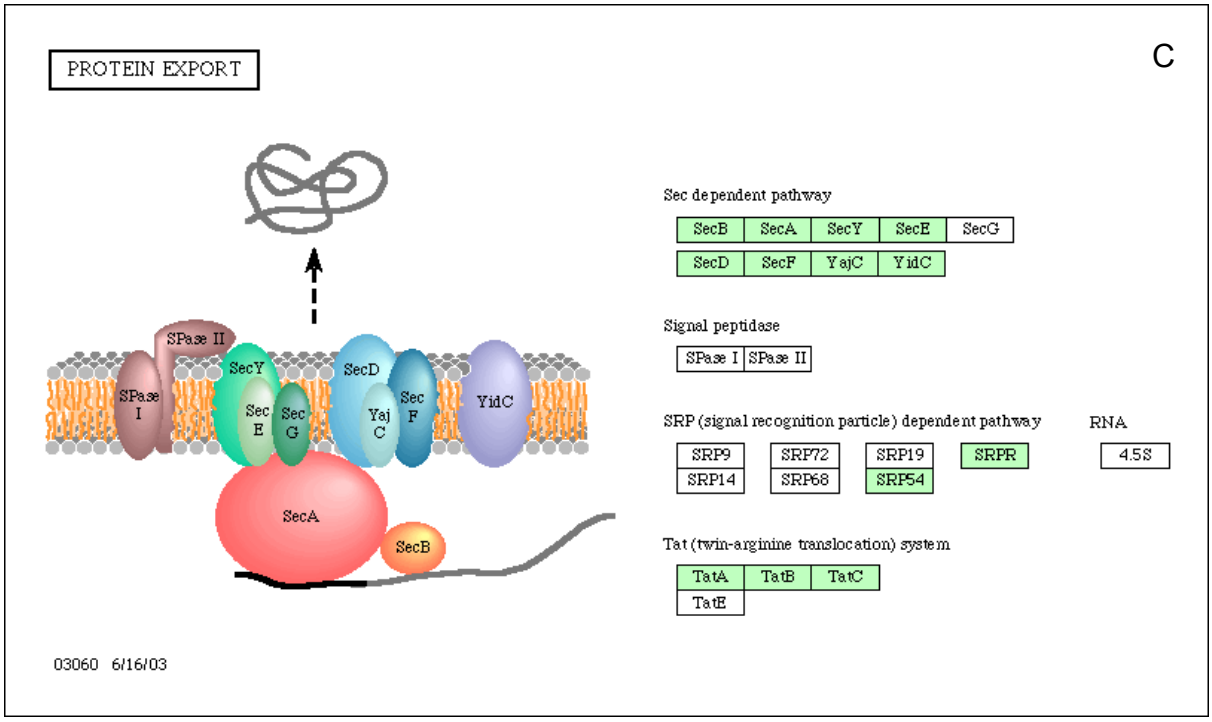
Figura 10 – Mapas de famílias de proteínas transportadoras construídos pelo KAAS



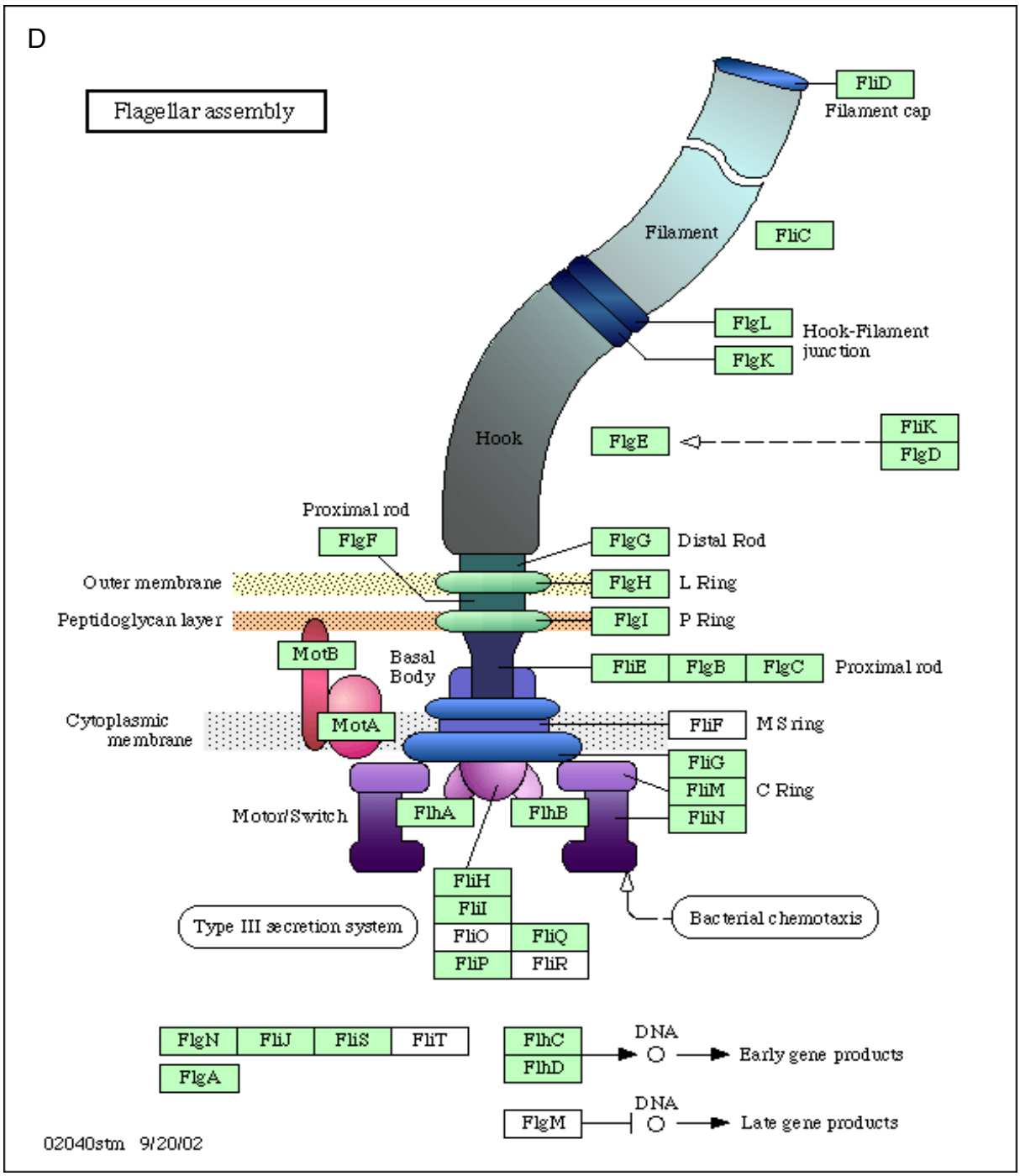
A – sistema de secreção do tipo III. As proteínas em verde são aquelas encontradas entre as ORF da anotação do genoma de H. seropedicae. Pode-se notar que o sistema está praticamente completo, com ausência das subunidades F, L e Q. As proteínas de montagem do flagelo também são mostradas devido à alta homologia com esse sistema.



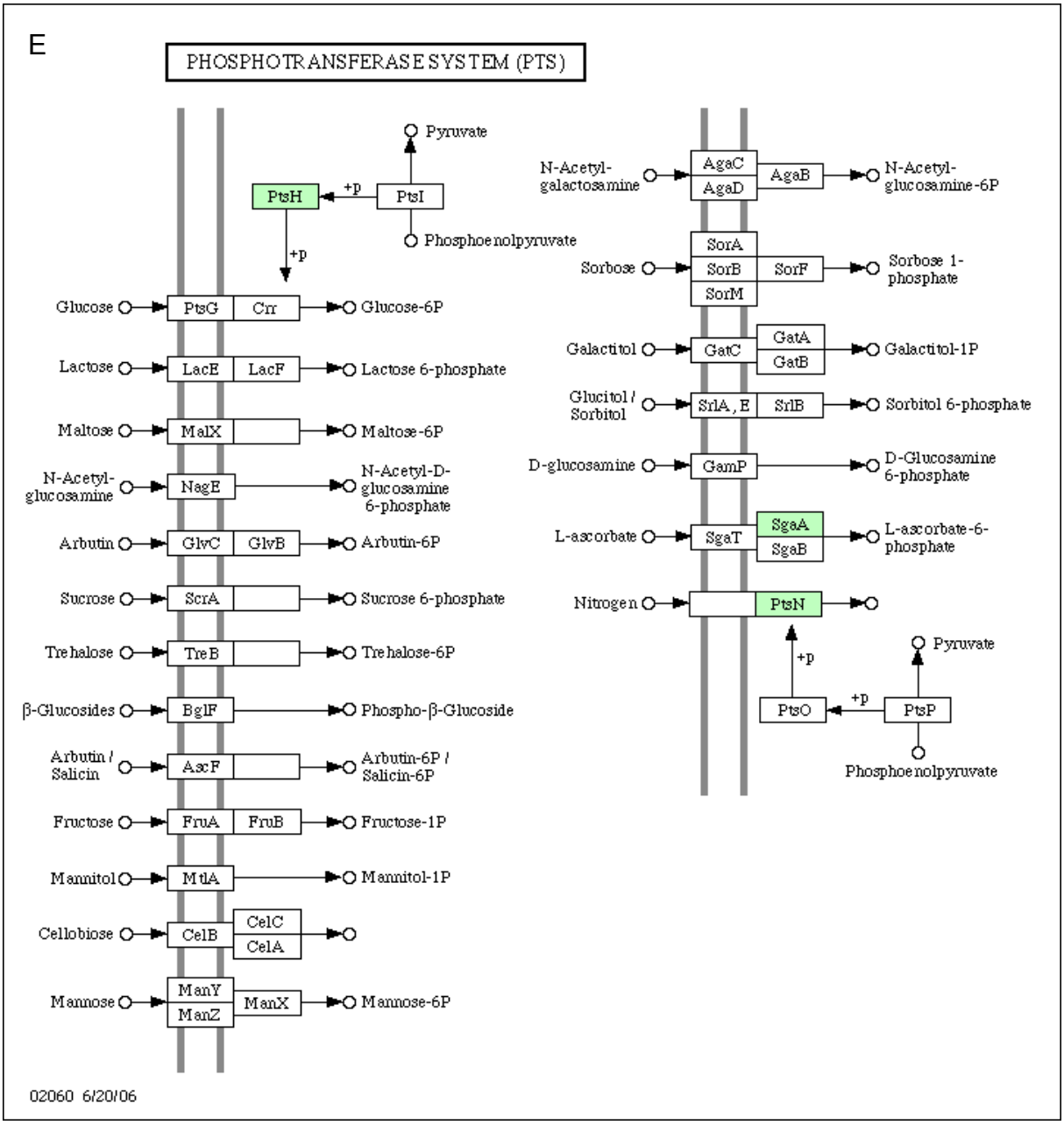
B – sistema de secreção do tipo II. As proteínas em verde são aquelas encontradas entre as ORF da anotação do genoma de H. seropedicae. Os sistemas estão bastante incompletos com exceção das proteínas de montagem do "pílus" (Type IV pílus assembly proteins).



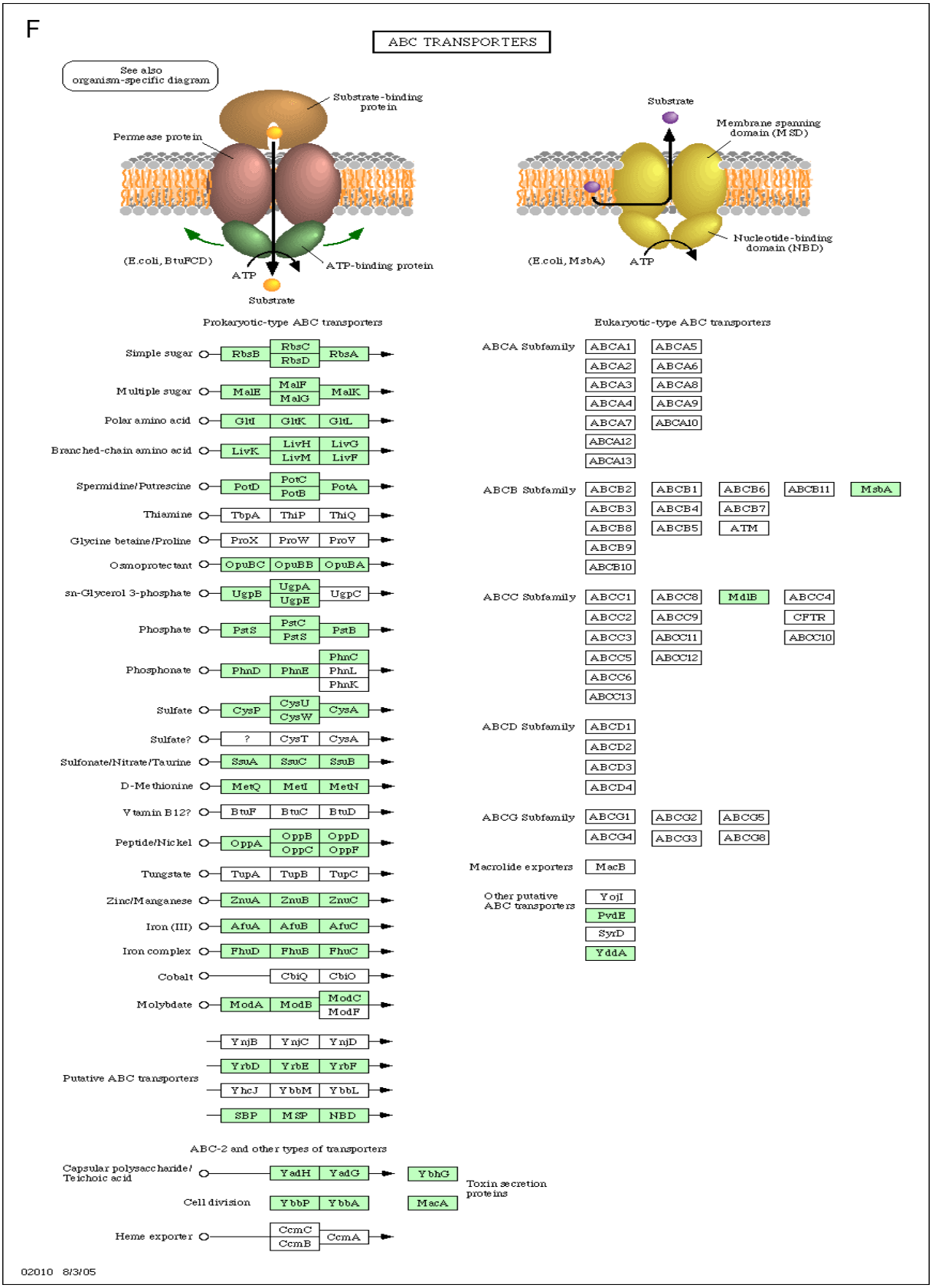
C – proteínas de excreção. As proteínas em verde são aquelas encontradas entre as ORF da anotação do genoma de H. seropedicae. Pode-se notar que o sistema de "preproteína translocase" (Sec dependent pathway) está praticamente completo. Também foram encontradas duas proteínas de "sinal de reconhecimento de partícula" (SRP); e as "proteínas translocases Sec-independentes", ou "sistema Tat".



D – montagem de flagelo. Praticamente todo o sistema foi encontrado, o que já era esperado, visto que H. seropedicae é uma bactéria flagelada.



E – PTS. Somente 3 ORF para sistema PTS foram encontradas em <u>H. seropedicae</u> pela análise do programa KAAS. A análise identificou um único componente proteína histidina (Hpr): a proteína PtsH; e ainda duas subunidades da enzima II, as proteínas SgaA, que participa do transporte de L-ascorbato, e PtsN, a qual participa de processo envolvendo nitrogênio.



F – família ABC. Pode-se notar vários transportadores completos (em verde) dessa família. Esses transportadores serão tratados melhor no tópico "Análise dos transportadores ABC" adiante.

Quanto ao mapa montado pelo programa KAAS para o Sistema de Secreção do Tipo III (TTSS) para *H. seropedicae,* esse foi comparado com o mapa montado para outras bactérias, visando verificar se nessas bactérias ocorre ausência das subunidades não encontradas em *H. seropedicae*.

Algumas dessas subunidades podem estar ausentes em vários organismos. Em *Pseudomonas syringae*, onde esse sistema é melhor estudado, estão ausentes as mesmas subunidades não encontradas em *H. seropedicae* pelo programa KAAS, e o sistema não deixa de ser funcional (GALAN & COLLMER, 1999). O resultado pode ser visto na tabela 7.

Tabela 7 – Subunidades constituintes do Sistema de Secreção do Tipo III ausentes em outras bactérias

ORGANISMO	SUBUNIDADES AUSENTES
Salmonella enterica	L
Shigella sonnei	D, L, Q, S
Shigella dysenteriae	F, L, Q, S
Erwinia carotovora	F, Q
Sodalis glossinidius	L
Xanthomonas campestris	F
Xanthomonas axonopodis	F
Xanthomonas oryzae	F
Pseudomonas syringae	F, L, Q
Burkholderia mallei	D, J
Chromobacterium violaceum	L

Parece comum ao Sistema de Secreção do Tipo III não estar completo em outras bactérias. Em Pseudomonas syringae, onde esse sistema é melhor estudado, estão ausentes as mesmas subunidades não encontradas em H. seropedicae pelo programa KAAS, e o sistema não deixa de ser funcional (GALAN & COLLMER, 1999).

5.5. Comparação entre possíveis proteínas transportadoras de H. seropedicae e H. rubrisubalbicans

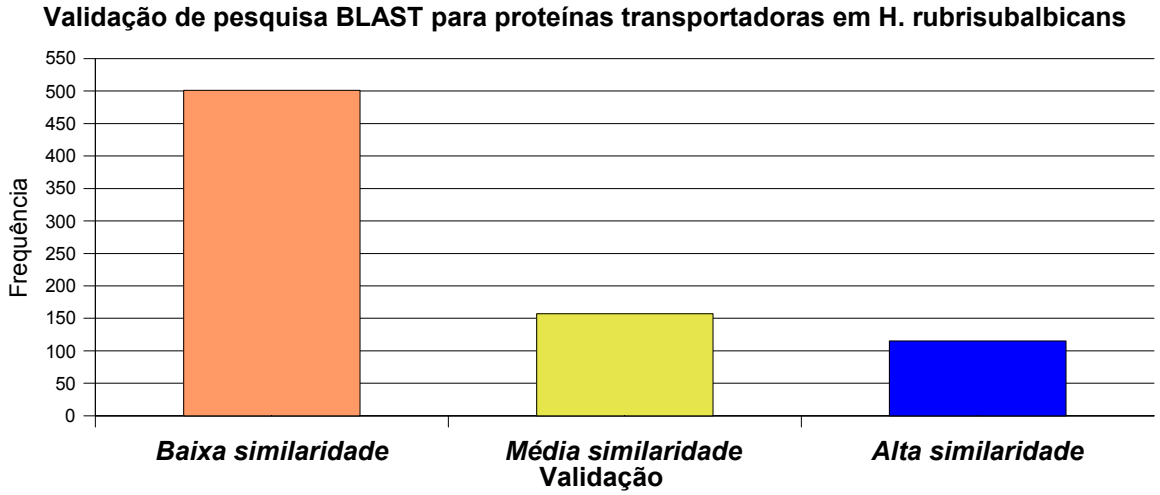
Após a montagem do genoma de *H. rubrisubalbicans* (com cerca de 20.000 seqüências) utilizando os programas PHRED para análise dos cromatogramas, CROSS_MATCH para filtro de vetor e seqüências do *operon rRNA,* e CAP3 para montagem dos *contigs.* 2.596 *contigs* foram obtidos. Em seguida, as 880 ORF de possíveis proteínas transportadoras de *H. seropedicae* foram submetidas a uma busca por similaridade com o programa TBLASTX (o qual converte seqüências de nucleotídeos das *queries* e *subjects* para seqüências de aa), contra um banco de dados formado pelas seqüências dos *contigs* de *H. rubrisubalbicans.* A análise foi feita com um limite de corte para o parâmetro EXPECT de 5×10^{-5}.

Um total de 774 ORF de *H. seropedicae* apresentaram alinhamentos. Esses alinhamentos foram também classificados por rede neuronal FAN, conforme mostram os resultados da figura 11.

O arquivo de treinamento da rede passou por algumas modificações em relação ao formato usado nas análises anteriores: valores de proporção em relação ao tamanho da seqüência *subject* (banco de dados) foram retirados, pois nesta análise foram usadas as seqüências dos *contigs* obtidos na montagem, sendo suas propriedades (proporção de tamanho, tamanho do alinhamento em relação a seqüência etc) diferentes daquelas esperadas para uma seqüência de proteína.

Essas modificações também podem explicar a alta proporção de ORF com baixa similaridade em relação ao banco, embora a rede tenha obtido uma taxa de acerto cuja porcentagem de média harmônica ficou em torno de 90,7%.

Figura 11 – Similaridade entre as ORF de possíveis transportadores em *H. seropedicae* e *H. rubrisubalbicans* através de pesquisa BLAST seguida de classificação através de rede neuronal FAN



A figura mostra que a maioria das ORF de proteínas transportadoras (501) de H. seropedicae possui baixa similaridade em relação aos contigs de H. rubrisubalbicans. 157 apresentaram média similaridade, e 115 alta similaridade. Isso pode ter ocorrido pelo fato do genoma de H. rubrisubalbicans estar bastante fragmentado (2.596 contigs); e também devido ao fato do treinamento da rede ter passado por modificações, nas quais foram retirados valores relativos ao subject do arquivo de treinamento.

Das 880 possíveis proteínas transportadoras de *H. seropedicae,* 773 possuem indícios de existência também em *H. rubrisubalbicans.* Esses indícios de proteínas foram também classificados segundo suas famílias, conforme o *site* TransportDB (tabela 8). A fragmentação do genoma de *H. rubrisubalbicans* pode ter reduzido o número de proteínas transportadoras encontradas (2.596 *contigs*).

Tabela 8 – Possíveis proteínas transportadoras de *H. seropedicae* com indícios em *H. rubrisubalbicans*

FAMÍLIA	*H. seropedicae*	*H. rubrisubalbicans*	FAMÍLIA	*H. seropedicae*	*H. rubrisubalbicans*	FAMÍLIA	*H. seropedicae*	*H. rubrisubalbicans*
ABC	528 (60%)	478	ArsB	1 (0,1%)	1	F-ATPase	11 (1,2%)	10
MFS	105 (12%)	89	H+-PPase	1 (0,1%)	1	Tat	5 (0,6%)	4
MscS	4 (0,4%)	1	GPTS	2 (0,2%)	2	APC	4 (0,4%)	2
CIC	8 (0,9%)	7	DASS	1 (0,1%)	1	GPH	1 (0,1%)	1
BenE	1 (0,1%)	1	PnuC	1 (0,1%)	1	MOP	6 (0,7%)	4
P-ATPase	13 (1,5%)	10	MIT	4 (0,4%)	3	Amt	19 (2,1%)	19
CHR	2 (0,2%)	2	BASS	3 (0,3%)	3	CPA	10 (1,1%)	9
NCS2	2 (0,2%)	0	GntP	3 (0,3%)	3	SSPTS	3 (0,3%)	2
SSS	2 (0,2%)	1	FeoB	5 (0,6%)	3	LIV-E	1 (0,1%)	1
AEC	2 (0,2%)	2	NhaA	2 (0,2%)	0	PiT	2 (0,2%)	1
TTT	3 (0,3%)	0	RhtB	5 (0,6%)	5	KUP	1 (0,1%)	1
DAACS	3 (0,3%)	2	NCS1	1 (0,1%)	1	AAA	4 (0,4%)	4
DMT	20 (2,3%)	13	SulP	2 (0,2%)	1	CitMHS	1 (0,1%)	1
RND	31 (3,5%)	31	Nramp	1 (0,1%)	1	OFeT	1 (0,1%)	1
TRAP-T	14 (1,6%)	13	MerTP	1 (0,1%)	0	CDF	3 (0,3%)	3
Oxa1	1 (0,1%)	1	OPT	1 (0,1%)	1	outras	35 (4%)	32
						TOTAL	880	773

Em preto estão os nomes das famílias, em verde o número de proteínas com sua porcentagem em relação ao número total de ORF do genoma de H. seropedicae; e em azul estão os números dos indícios dessas proteínas q foram encontrados em H. rubrisubalbicans.
*Das 880 proteínas de H. seropedicae, 773 possuem indícios encontrados em H. rubrisubalbicans. Esse número mais baixo pode ter sido causado pela grande fragmentação do genoma de H. rubrisubalbicans (*2.596 *contigs). A diminuição numérica refletiu nas famílias mais numerosas, como ABC, MFS, DMT. Alguns resultados foram mais tendenciosos, como na família RND, onde as 31 proteínas encontradas em H. seropedicae possuem indícios em H. rubrisubalbicans, em relação à familia TTT, por exemplo, onde as 3 proteínas encontradas em H. seropedicae não possuem indícios em H. rubrisubalbicans.*

5.6. Análise de preferência de códon

As 5.100 ORF anotadas pelo GENOPAR e as 880 ORF identificada como possíveis proteínas transportadoras na pesquisa BLASTX contra o banco de dados TransportDB, foram analisadas com os programas GCUA (MCINERNEY, 1998) e CODONW (PENDEN, 1999), visando obter o uso de códons no genoma e nas proteínas transportadoras de *H. seropedicae*.

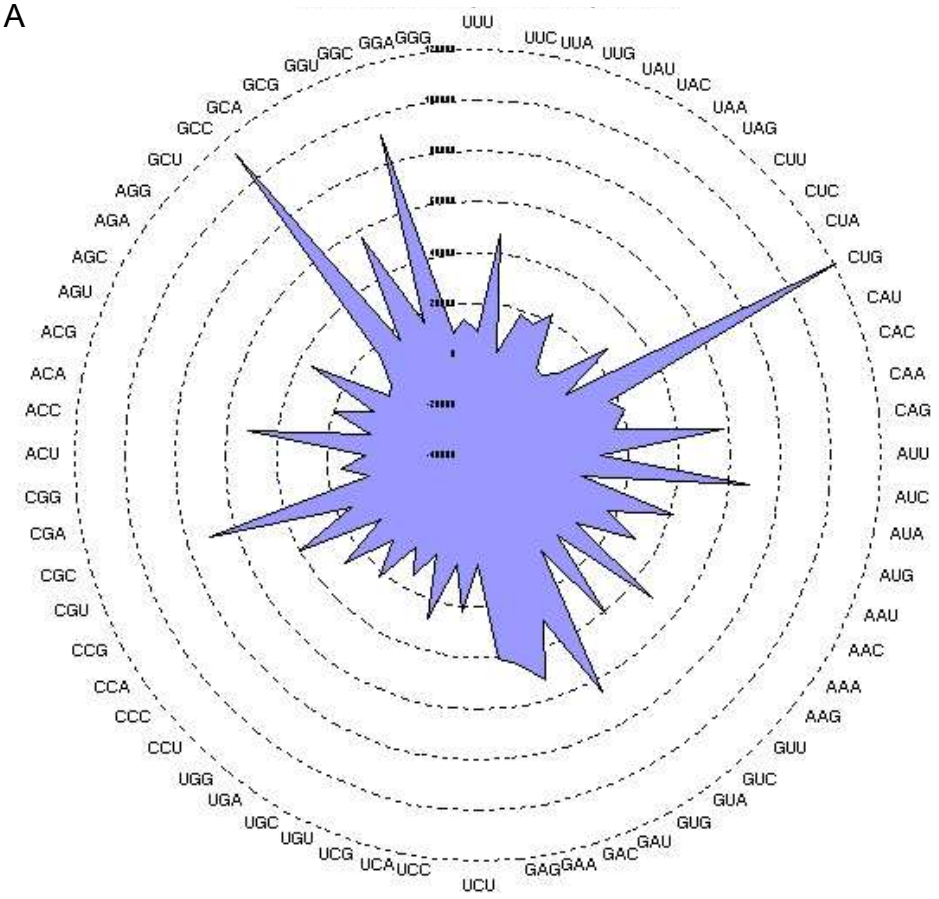
A figura 12 mostra os resultados de uso de códons realizada com o programa GCUA. Segundo essa análise, aparentemente não há grandes diferenças entre a distribuição do uso de códons entre todas as ORF presentes no genoma de *H. seropedicae,* e aquelas para possíveis proteínas transportadoras (figura 12).

A semelhança entre os dois gráficos sugere que essas proteínas transportadoras presentes no genoma de *H. seropedicae*, de forma geral, não foram adquiridas recentemente por transferência lateral. Entretanto, a ocorrência de poucas delas com uma freqüência de uso de códons diferenciada pode ter sido "mascarada" nesta análise.

Pode-se notar também a preferência de códons terminados em G ou C. Isso pode estar relacionado à característica do genoma de *H. seropedicae,* o qual apresenta um conteúdo de GC elevado (62,7%), indicando que a tendência no uso de códons sofre grande pressão do conteúdo GC e, provavelmente, menor pressão da eficiência traducional.

Figura 12 – Gráficos representando a distribuição do uso de códons pelo organismo *H. seropedicea*

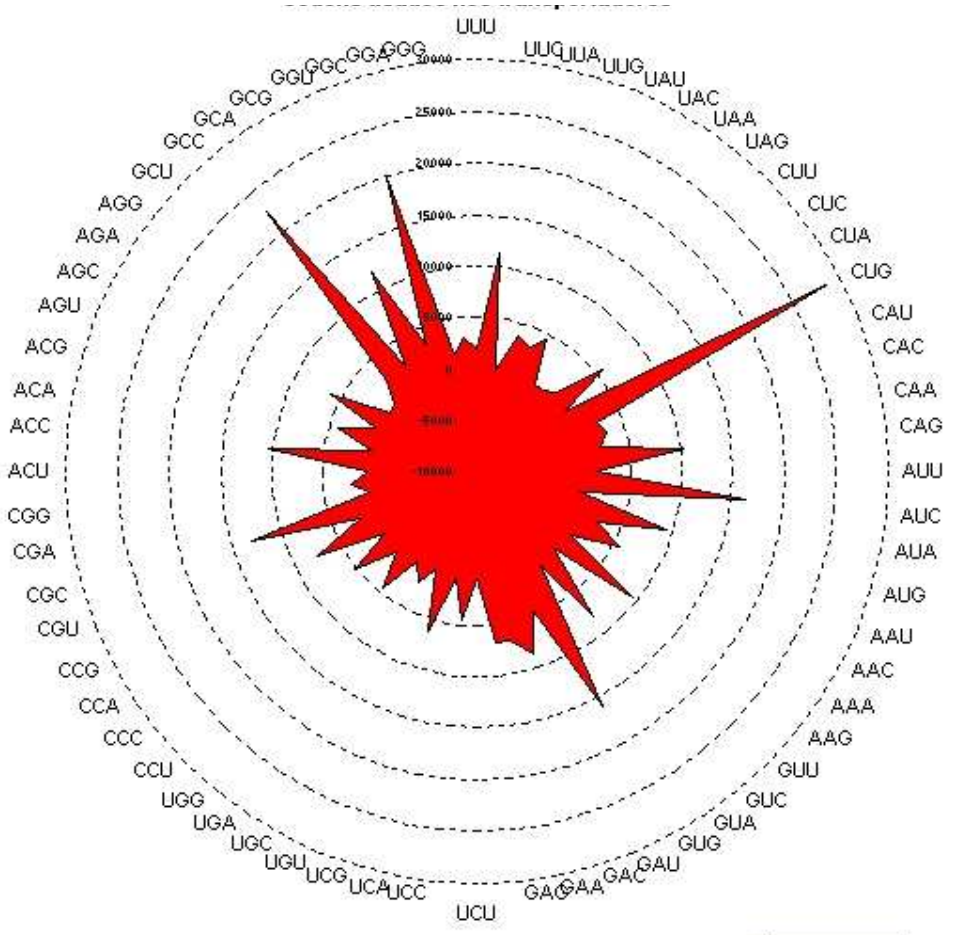
A



UUU	1382
UUC	8737
UUA	78
UUG	2711
UAU	2254
UAC	3216
UAA	134
UAG	43
CUU	418
CUC	4176
CUA	216
CUG	21475
CAU	1680
CAC	2167
CAA	1511
CAG	7029
AUU	1280
AUC	12118
AUA	127
AUG	7083
AAU	1830
AAC	4082
AAA	670
AAG	7112
GUU	339
GUC	5982
GUA	598
GUG	11965
GAU	3022
GAC	5684
GAA	4669
GAG	4437
UCU	311
UCC	3263
UCA	337
UCG	4619
UGU	224
UGC	1409
UGA	441
UGG	3434
CCU	706
CCC	3730
CCA	454
CCG	5514
CGU	1549
CGC	8753
CGA	238
CGG	1404
ACU	468
ACC	7666
ACA	311
ACG	2985
AGU	471
AGC	4194
AGA	83
AGG	262
GCU	1544
GCC	16483
GCA	1651
GCG	8618
GGU	2298
GGC	14760
GGA	996
GGG	2050

A: Uso de códons correspondente ao resultado obtido para todas as ORF de H. seropedicae. A tabela ao lado mostra o número de vezes que cada códon foi usado pelas proteínas transportadoras.

B



B: Uso de códons nas 880 proteínas transportadoras encontradas por pesquisa BLAST em relação ao banco de dados TransportDB.

Pode-se notar semelhança entre os dois gráficos, indicando que não há grandes diferenças no uso de códons por todas as ORF do genoma de H. seropedicae em relação às ORF de proteínas transportadoras; e também indica a preferência por códons com uso de GC na terceira base.

Estas ORF foram também usadas para o cálculo de índices de tendência de uso de códons com o programa CODONW (PENDEN, 1999). Partindo-se das 5.100 ORF totais, aquelas classificadas como "FRAMASHIFT:*yes*" e/ou "VALIDATION:*pending*" pela anotação do GENOPAR (957 ORF), foram retiradas da análise.

Sendo assim, primeiramente as ORF foram analisadas quanto à sua integridade, sendo que 85 apresentaram algum dos seguintes problemas: 65 ORF não iniciam com um códon de início reconhecido; 7 ORF possuem códon(s) não traduzíveis (devido à presença de bases indefinidas na seqüência); e 13 ORF possuem códon(s) de parada internos.

Somente as 20 ORF contendo códons não traduzíveis ou *stop codons* internos foram removidas das análises subseqüentes. Um total de 4.123 ORF foram analisadas quanto aos índices de tendência de uso de códons. Deste total, 627 ORF (das 880 ORF identificadas) correspondem às possíveis proteínas transportadoras.

Na figura 13 é mostrada a relação entre vários índices de tendência no uso de códons para as ORF de *H. seropedicae*. Em todos os casos foi observado valores correspondentes para as ORF de possíveis proteínas transportadoras com aqueles para as ORF totais. Entretanto, algumas ORF para essas proteínas apresentaram valores indicando alta tendência no uso de códons.

Na figura 13A, observa-se que ORF com alto CAI tendem a ter baixo ENc, ou seja, quanto mais espera-se que uma proteína seja expressa, maior é sua tendência na escolha por códons, cuja relação é confirmada pela posição das ORF marcadas como "genes altamente expressos" na figura. Estas ORF representam 30 genes que foram identificados a partir da análise proteômica de *H. seropedicae* gel 2D e

espectrometria de massa, e apresentaram alto nível de expressão em diferentes condições de cultivo da bactéria (SEIXAS, D., comunicação pessoal).

Nas figuras 13B e C, temos que quanto maior a tendência da proteína ser expressa (indicado por altos valores de CAI na figura 13B), maior é a tendência da escolha de códons com conteúdo GC na terceira base. Esta escolha é feita, principalmente pela terceira base, porque essa é mais variável entre os códons sinônimos. A análise mostra que as proteínas que possivelmente são altamente expressas, apresentam adaptação às características do genoma quanto ao conteúdo GC.

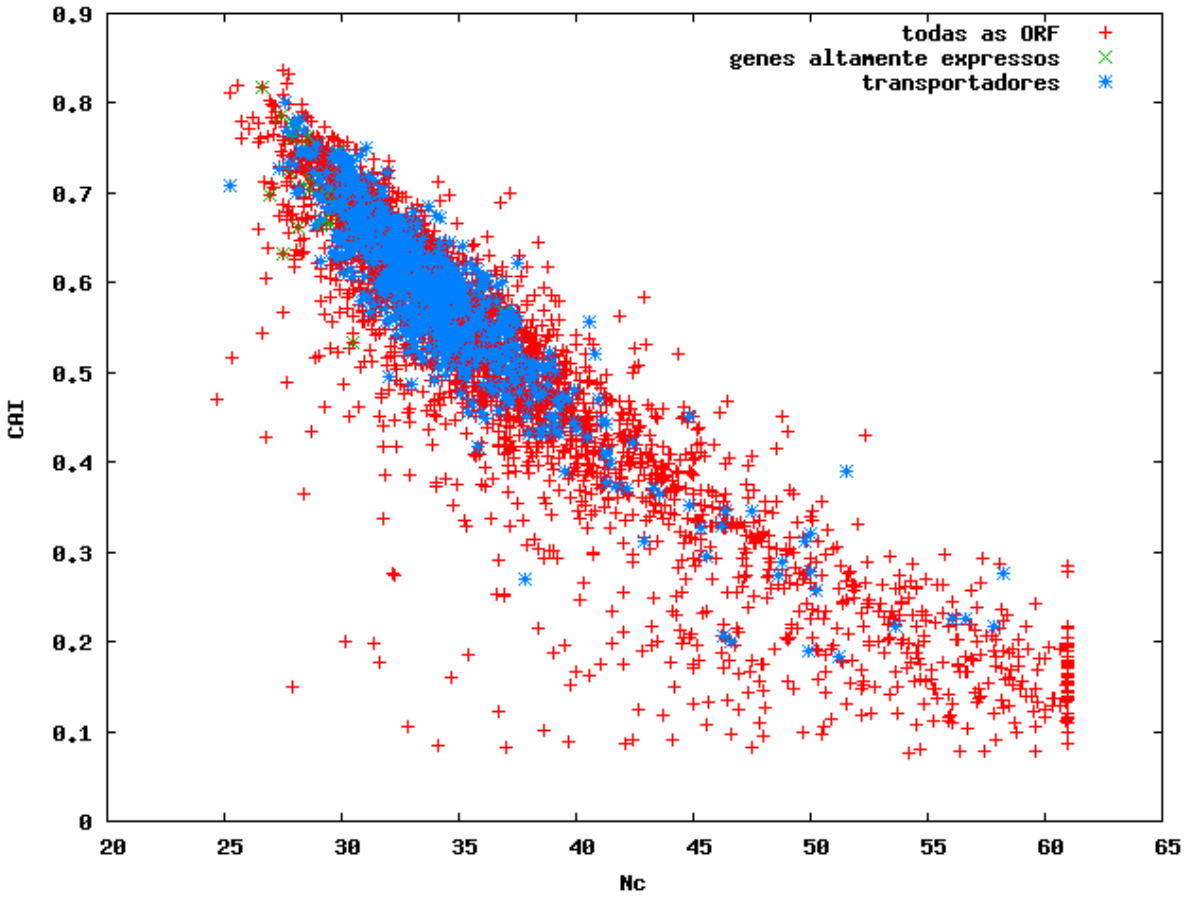
A mesma tendência adaptativa pode ser observada na figura 13C, onde as ORF com baixos valores de ENc apresentam, em geral, altos valores de GC na terceira base. Neste caso, os baixos valores de ENc indicam que as ORF tem alta tendência no uso de códons (não aleatório).

Aparentemente as ORF para possíveis proteínas transportadoras apresentam as mesmas tendências adaptativas da média apresentada por todas as ORF, sugerindo que estas proteínas estão adaptadas às características do genoma e devem ter sido adquiridas muito cedo em termos evolutivos, sendo transferidas verticalmente.

Algumas exceções são observadas entre as ORF para essas proteínas, por exemplo apresentando altos valores de ENc e baixos valores de CAI (figura 13A), o que indica uma escolha aleatória de códons e proteínas pouco expressas, podendo estar relacionados a proteínas adquiridas horizontalmente e que confiram vantagens adaptativas, como por exemplo, resistência a drogas e metais, sendo expressos somente em condições de estresse.

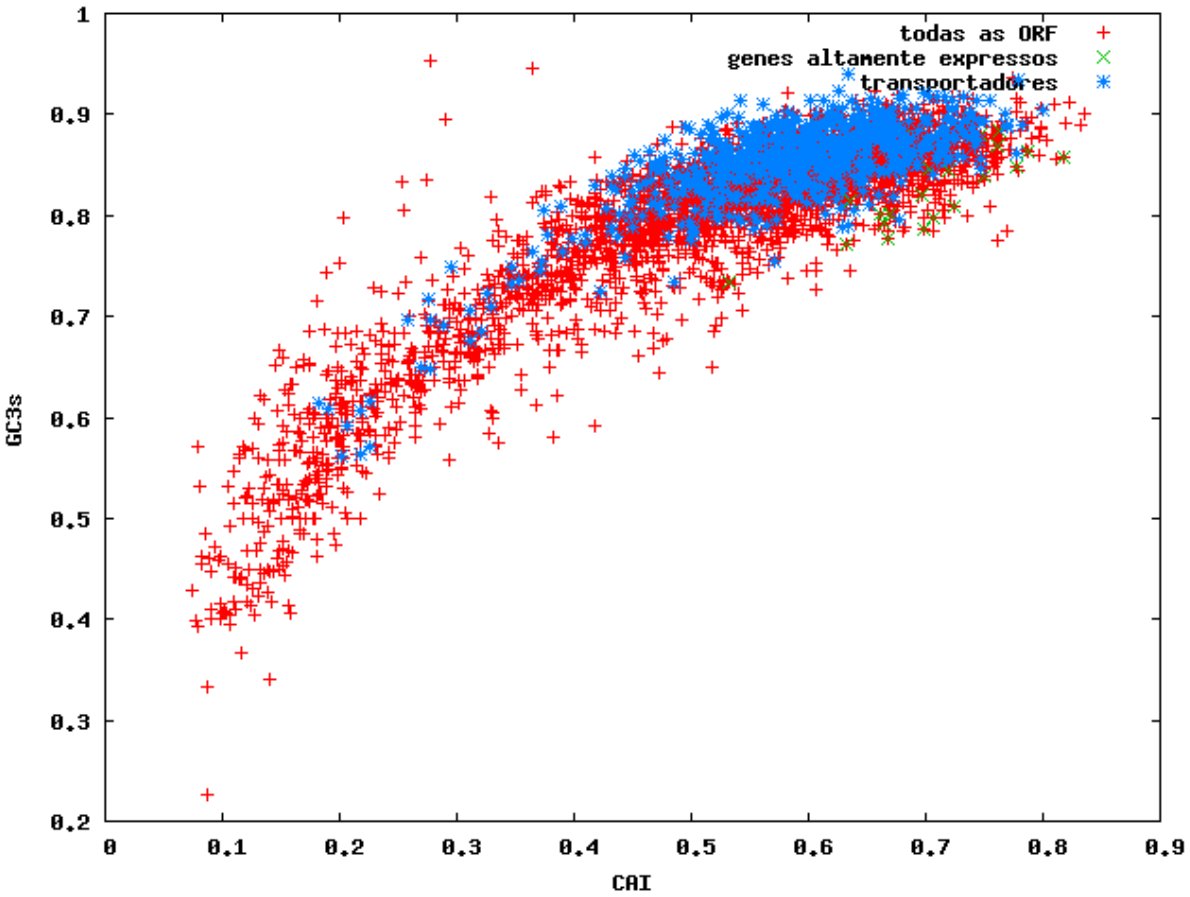
Figura 13 – Comparação entre índices de tendência no uso de códons para ORF anotadas de *H. seropedicae* e para ORF de possíveis proteínas transportadoras

A – Nc x CAI



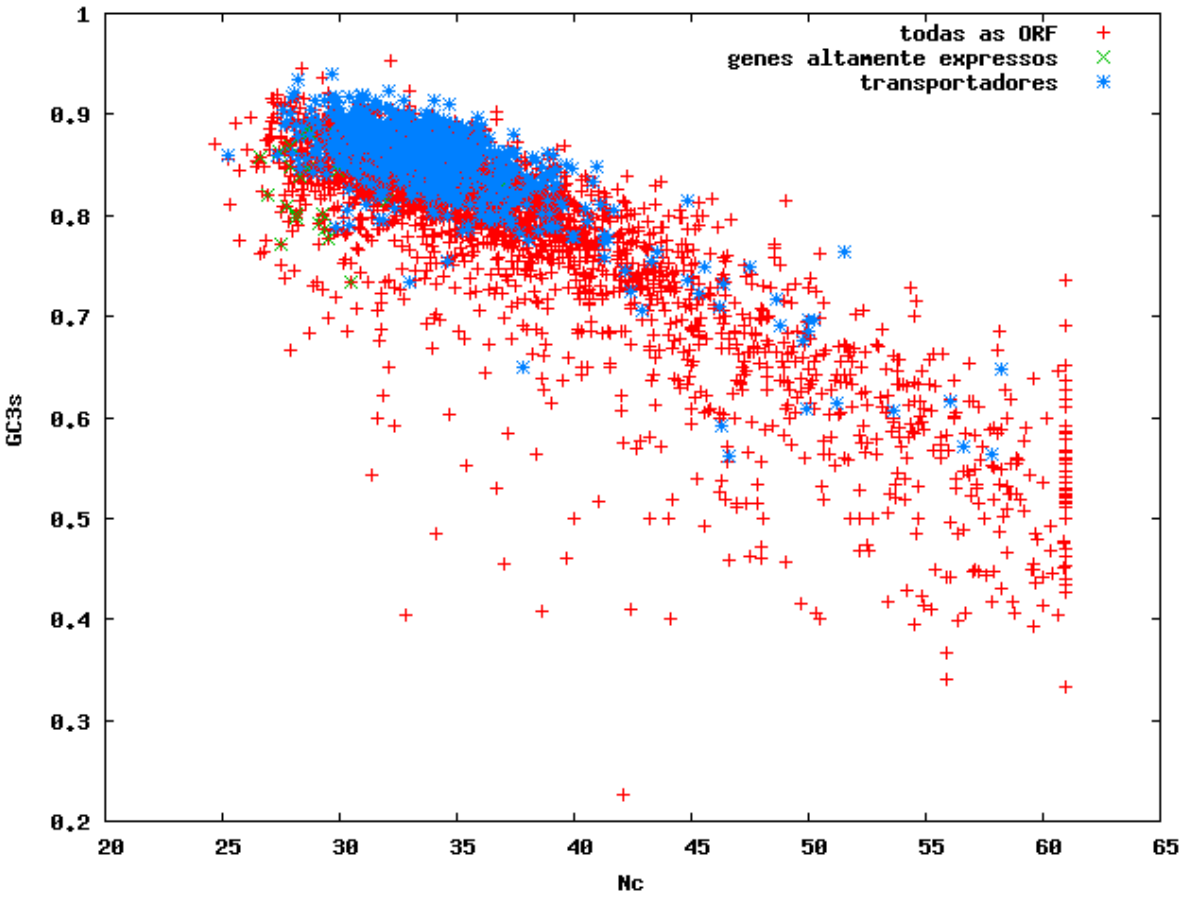
A: as ORF anotadas de H. seropedicae estão marcadas em vermelho; as ORF de possíveis proteínas transportadoras estão marcadas em azul; genes altamente expressos identificados através de análise de proteoma estão marcados em verde.
Observa-se que ORF com alto CAI tendem a ter baixo ENc, ou seja, quanto mais espera-se que uma proteína seja expressa, maior é sua tendência na escolha por códons. As proteínas transportadoras seguem a tendência do genoma de H. seropedicae.

B – CAI x GC3s



B: as ORF anotadas de H. seropedicae estão marcadas em vermelho; as ORF de possíveis proteínas transportadoras estão marcadas em azul; genes altamente expressos identificados através de análise de proteoma estão marcados em verde.
Esse gráfico mostra que quanto maior a tendência da proteína ser expressa (indicado por altos valores de CAI), maior é a tendência da escolha de códons com conteúdo GC na terceira base. A maioria das ORF está adaptada ao genoma, através do uso de GC na terceira base, e novamente as proteínas transportadoras seguem a tendência do genoma de H. seropedicae.

C – Nc x GC3s



C: Pode ser observada a mesma tendência adaptativa vista nos outros gráficos. As ORF com baixos valores de ENc apresentam, em geral, altos valores de GC na terceira base, ou seja, as ORF tem alta tendência no uso de códons.
As ORF de H. seropedicae estão, em geral, adaptadas ao genoma, sendo que as ORF de proteínas transportadoras seguem essa tendência adaptativa, e devem ter surgido muito cedo no processo evolutivo dessa espécie.

5.7. Análise dos transportadores ABC

O organismo *H. seropedicae* é capaz de transportar diversos substratos, tais como açúcares, aminoácidos, íons fosfato e sulfato, entre outros, conforme mostra o mapa da figura 10F.

A partir da reconstrução desse mapa para os sistemas de transporte realizados pela família ABC, os genes e as sub-unidades protéicas foram identificados e analisados em relação à vizinhança no genoma de *H. seropedicae* e à evidência de formação de possíveis *operons*. A tabela 9 mostra, detalhadamente, as ORF identificadas em relação aos sistemas de transporte da família ABC. As ORF visualizadas nessa tabela encontram-se agrupadas nos possíveis *operons*, que correspondem às unidades completas de transporte.

Algumas ORF mostraram similaridade com as proteínas *MsbA* (9 ORF) da subfamília ABCB, *MdlB* (1 ORF) da subfamília ABCC e *PvdE* (1 ORF) e *YddA* (2 ORF) de um transportador ABC putativo em eucariotos. Os demais componentes destes transportadores não foram identificados, porém foi observada uma alta similaridade dessas ORF com o banco de dados, com exceção de 6 nas quais foi verificada a presença de *frameshift* (5 ORF *MsbA* e uma *Mdlb*).

Pôde-se constatar que os transportadores ABC estão bem distribuídos pelo genoma de *H. seropedicae,* sendo a presença em maior número em alguns *contigs* devido ao tamanho destes *contigs.* Com isso, nenhuma das possíveis proteínas da família ABC foram identificadas nos *contigs* com menos de ~9,8 Kb.

Em alguns casos mais de uma ORF foi identificada como uma mesma proteína. Por exemplo, as ORF ORFID286.1415 e ORFID286.1467 foram

identificadas como a subunidade *GltI*, que é uma subunidade de ligação ao substrato no transporte de aminoácidos polares. Nesse caso, foi verificado a presença de *frameshift*, mas esse fator estava ausente, com exceção de um par (ORFID171.0002/ORFID171.0014), sugerindo a existência de mais de uma cópia de alguns genes no genoma.

Duas outras ORF (ORFID184.0136 e ORFID277.0338) não apresentaram similaridade na pesquisa BLASTX contra o banco de dados TransportDB, mas foram identificadas como transportadores ABC na análise realizada com o KAAS e assim constam na anotação de *H. seropedicae.*

O número de ORF para proteínas do transporte ABC corresponde a cerca de 4% das ORF de *H. seropedicae,* considerando somente as ORF presentes em possíveis *operons* completos, onde todas as subunidades protéicas foram encontradas próximas no genoma (cerca de 196 ORF), número acima do encontrado em outras bactérias, cuja média fica em torno de 2% (TOMII & KANEHISA, 1998).

Se levados em consideração as 528 ORF de transportadores ABC encontradas com a pesquisa BLASTX, boa parte não está em *operons* completos em relação ao encontrado em outras bactérias (TOMII & KANEHISA, 1998), pois os transportadores ABC podem apresentar subunidades ausentes, funcionando então com homodímeros de outras subunidades (GRANJEIRO et al., 2004).

Esse número elevado de transportadores ABC possivelmente reflete a grande quantidade de ORF de possíveis proteínas transportadoras encontradas no genoma de *H. seropedicae*.

5.7.1. Análise de domínios transmembrana

Os sistemas de transporte ABC apresentam uma ou mais sub-unidades protéicas transmembrana (permeases), as quais são classificadas, dessa maneira, pela análise do programa KAAS e na anotação do genoma de *H. seropedicae,* com exceção de cinco ORF: ORFID240.0070, ORFID265.0462, ORFID269.0235, ORFID275.0660 e ORFID287.0360.

As ORF de proteínas da família ABC foram submetidas a análise de hélices transmembrana com o uso do programa TMHMM (KROGH et al., 2001). Na sua grande maioria, observou-se que as subunidades identificadas como permeases realmente possuem hélices transmembrana, os quais variam entre 4 e 12 (tabela 10), dependendo do tipo de transporte. Outras, não necessariamente permeases, apresentaram uma estimativa de 1 hélice transmembrana, valor considerado pelo autor do programa TMHMM como um possível peptídio sinal (KROGH et al., 2001).

Tabela 9 – Sistemas de transporte da família ABC completos identificados no genoma de *H. seropedicae*. O número entre parênteses indica o número de hélices transmembrana previstas

COMPONENTE	ORF IDENTIFICADAS
1- AÇÚCAR SIMPLES	
RbsB(substrate-binding)	ORFID240.0019 (1)
RbsC(permease)	ORFID240.0045(4)
RbsD(permease)	ORFID240.0070
RbsA(ATP-binding)	ORFID240.0035
2- AÇÚCAR MÚLTIPLO	
MalE(substrate-binding)	ORFID171.0002/0014; ORFID183.0079; ORFID241.0076; ORFID246.0084/0095; ORFID287.0331
MalF(permease)	ORFID171.0023(6); ORFID183.0085(6) ; ORFID241.0086(6); ORFID246.0107(6); ORFID287.0352(6)
MalG(permease)	ORFID171.0029(6); ORFID183.0091(6); ORFID241.0093(6); ORFID246.0113(6); ORFID287.0346(6)
MalK(ATP-binding)	ORFID171.0044; ORFID183.0064; ORFID241.0065; ORFID246.0122; ORFID287.0360/0367
3- AMINOÁCIDOS POLARES	
GltI(substrate-binding)	ORFID249.0351; ORFID256.0283; ORFID264.0621; ORFID286.1415/1467; ORFID253.0465
GltK(permease)	ORFID249.0342(4); ORFID256.0268(3); ORFID264.0628(4);ORFID286.1436(5)/1442(5)/1478(5); ORFID253.0458(5)
GltL(ATP-binding)	ORFID249.0339; ORFID256.0265; ORFID264.0634; ORFID286.1455/1487; ORFID253.0446
4- CADEIA RAMIFICADA DE AMINOÁCIDOS	
LivK(substrate-binding)	ORFID204.0041(1); ORFID213.0164; ORFID218.0221; ORFID248.0482; ORFID250.043; ORFID281.0212; ORFID282.0421(1); ORFID287.0502
LivH(permease)	ORFID204.0059(8); ORFID213.0137(8); ORFID218.0211(7); ORFID248.0473(7); ORFID250.0416(8); ORFID281.0220(7); ORFID282.0452(7); ORFID287.0519(8)
LivM(permease)	ORFID204.0068(9); ORFID213.0149(9); ORFID218.0204(11); ORFID248.0466(8); ORFID250.0409(8); ORFID281.0230(8); ORFID282.0461(7); ORFID287.0512(10)
LivG(ATP-binding)	ORFID204.0074; ORFID213.0129; ORFID218.0199; ORFID248.0456; ORFID250.0444; ORFID281.0235; ORFID282.0466; ORFID287.0527
LivF(ATP-binding)	ORFID204.0081; ORFID213.0173; ORFID218.0193; ORFID248.0450; ORFID250.0438; ORFID281.0241; ORFID282.0473; ORFID287.0477
5-ESPERMIDINA/PUTRECINA	
PotD(substrate-binding)	ORFID257.0295(1)
PotC(permease)	ORFID257.0315(6)
PotB(permease)	ORFID257.0306(6)
PotA(ATP-binding)	ORFID257.0283
6-OSMOPROTETOR	
OpuBC(substrate-binding)	ORFID237.0293(1)
OpuBB(permease)	ORFID237.0284(6)/0295(6)

continua

	continuação
OpuBA(ATP-binding)	ORFID237.0304
7-SN-GLICEROL 3-FOSFATO	
UgpB(substrate-binding)	ORFID248.0284(1); ORFID268.0600(1)
UgpA(permease)	ORFID248.0291(6); ORFID268.0607(6)
UgpE(permease)	ORFID248.0300(6); ORFID268.0615(6)
UgpC(ATP-binding)	ORFID268.0629(?)
8-FOSFATO	
PstS(substrate-binding)	ORFID265.0462
PstC(permease)	ORFID265.0470(6)
PstS(permease)	ORFID265.0462
PstB(ATP-binding)	ORFID265.0480
9-SULFATO	
CysP(substrate-binding)	ORFID199.0109
CysU(permease)	ORFID199.0190(5)
CysW(permease)	ORFID199.0197(5)
CysA(ATP-binding)	ORFID199.0206
10-SULFONATO/NITRATO/TAURINO	
SsuA(substrate-binding)	ORFID181.0004; ORFID196.0083; ORFID199.0130(1); ORFID200.0147(1)/0156(1); ORFID226.0180; ORFID236.0354; ORFID249.0392; ORFID252.0404; ORFID272.0662; ORFID253.0329
SsuC(permease)	ORFID181.0023(6); ORFID196.0069(6); ORFID199.0151(5); ORFID200.0129(6); ORFID226.0140(6); ORFID236.0342(7); ORFID249.0412(6); ORFID252.0397(6); ORFID272.0656(6); ORFID253.0309
SsuB(ATP-binding)	ORFID181.0014; ORFID196.0059; ORFID199.0160; ORFID200.0137; ORFID226.0150; ORFID236.0335; ORFID249.0401; ORFID252.0413; ORFID272.0651; ORFID253.0322
11-D-METIONINA	
MetQ(substrate-binding)	ORFID215.0122; ORFID287.0064(1)
MetI(permease)	ORFID215.0124(6); ORFID287.0058(5)
MetN(ATP-binding)	ORFID215.0135; ORFID287.0053
12-PEPTÍDIO/NÍQUEL	
OppA(substrate-binding)	ORFID147.0015(1)/0043; ORFID256.0637(1); ORFID274.0504;ORFID271.0105(?)
OppB(permease)	ORFID147.0021(6); ORFID256.0623(6); ORFID274.0515(5); ORFID215.0183(4); ORFID271.0092(6)
OppC(permease)	ORFID147.0029(5); ORFID256.0615(5); ORFID274.0519(5); ORFID215.0189(6); ORFID271.0086(5)
OppD(ATP-binding)	ORFID215.0195; ORFID271.0076
OppF(ATP-binding)	ORFID256.0603; ORFID274.0538;ORFID271.0066
13-ZINCO/MANGANÊS	
ZnuA(substrate-binding)	ORFID205.0095
	continua

	continuação
ZnuB(permease)	ORFID205.0101(7)
ZnuC(ATP-binding)	ORFID205.0112
14-FERRO(III)	
AfuA(substrate-binding)	ORFID247.0046; ORFID280.0717
AfuB(permease)	ORFID247.0034(12); ORFID280.0701(12)
AfuC(ATP-binding)	ORFID247.0019; ORFID280.0689
15-COMPLEXO DE FERRO	
FhuD(substrate-binding)	ORFID283.0579
FhuB(permease)	ORFID283.0567(8)
FhuC(ATP-binding)	ORFID283.0556
16-MOLIBDATO	
ModA(substrate-binding)	ORFID269.0219(1); ORFID275.0676
ModB(permease)	ORFID269.0229(5); ORFID275.0667(5)
ModC(permease)	ORFID269.0235; ORFID275.0660
ModF(ATP-binding)	
17-TRANSPORTADORES ABC PUTATIVOS	
YrbD(substrate-binding)	ORFID284.0582(1)
YrbE(permease)	ORFID284.0576(6)
YrbF(ATP-binding)	ORFID284.0569
SBP(substrate-binding)	ORFID177.0080/0087
MSP(permease)	ORFID177.0066(8)
NBD(ATP-binding)	ORFID177.0060
18-POLISSACARÍDO CAPSULAR/ÁCIDO TEICÓICO	
Yadh	ORFID266.0137(6); ORFID284.0623(7)
YadG	ORFID266.0157(5); ORFID284.0617; ORFID184.0164(5)
YbhG	ORFID184.0136*
19-DIVISÃO CELULAR	
YbbP	ORFID205.0136(4); ORFID229.0111(4); ORFID277.0309(4)
YbbA	ORFID205.0127; ORFID229.0115; ORFID277.0347
MacA	ORFID277.0338*
20-FOSFONATO	
PhnD(substrate-binding)	ORFID192.0149
PhnE(permease)	ORFID192.0124/0131
PhnC(ATP-binding)	ORFID192.0139
PhnL(ATP-binding)	
PhnK(ATP-binding)	

**ORF não identificadas em pesquisa BLAST contra o banco de dados TransportDB.*

Entre parênteses estão marcados os números de hélices transmembrana encontrados nas subunidades. Pode-se perceber que esses números são encontrados principalmente em subunidades que aparecem classificadas como "permease", que são subunidades transmembrana.
Nota-se também que transportadores genéricos, como transportadores de açúcares múltiplos possuem várias cópias no genoma, enquanto transportadores mais específicos, como sulfato e fosfato, possuem apenas uma (ver também tabela 10).

Tabela 10 – Número de possíveis *operons* inteira ou parcialmente completos, e o número de hélices transmembrana encontrados na subunidade transmembrana (permease)

Substrato	Nº de transportadores	Domínios de ligação a substrato	Domínios transmembrana	Domínios de ligação a ATP	TMHMM
Açúcar simples	1	1	2	1	4 e 0
Açúcar múltiplo	5	1	2	1	6 e 6
Aminoácidos polares	5	1	1	1	3 a 5
Cadeia ramificada de aminoácidos	8	1	2	2	7 ou 8; 8 a 11
Espermidina/Putrecina	1	1	2	1	6 e 6
Osmoprotetor	1	1	1	1	6
SN-Glicerol 3-fosfato	1	1	2	1	6 e 6
Fosfato	1	1	2	1	6 e 0
Sulfato	1	1	2	1	5 e 5
Sulfonato/Nitrato/Taurino	10	1	1	1	5 a 7
D-Metionina	2	1	1	1	5 a 6
Peptídio/Níquel	5(?)	1	2	2	4 a 6
Zinco/Manganês	1	1	1	1	7
Ferro(III)	2	1	1	1	12
Complexo de Ferro	1	1	1	1	8
Molibdato	2	1	1	1	8
Putativo 1	1	1	1	1	6
Putativo 2	1	1	1	1	8
Polissacarídeo Capsular/ Ácido Teicóico	3(?)	1	1	1	5
Divisão Celular	1	1	1	1	4
Fosfonato	1	1	1	3(1)	?

O "número de transportadores" indica que todas as subunidades que compõe o transportador foram encontradas e estão próximas num mesmo contig de H. seropedicae, sendo considerados possíveis operons. Muitos transportadores ficaram "quebrados" (várias subunidades encontradas isoladas das demais), mas apresentaram no mínimo um transportador inteiramente completo como mostra o número na segunda coluna (número de transportadores); já as interrogações após os números de transportadores indicam que todas as subunidades do transportador foram encontradas, mas algumas não estão no mesmo contigs.
Os números em cada coluna correspondente aos domínios, indica a quantidade desses domínios que o transportador completo possui.
Na coluna TMHMM, o "e" ou ";" são usados quando o transportador possui mais de um domínio transmembrana (permease), por exemplo, no açúcar simples, 4 e 0 indicam que numa das permeases foram encontradas 4 hélices transmembrana com o programa TMHMM, e na outra nenhum. "Ou" ou "a", indica a variação no número de hélices, por exemplo, em "cadeia ramificada de aminoácidos", o qual possui duas permeases, numa foram encontradas 7 ou 8 hélices transmembrana em seus 8 transportadores completos, enquanto na outra permease foram encontradas de 8 a 11 hélices transmembrana.
A interrogação ("?") em relação ao fosfato representa que sua permease não foi analisada pelo programa TMHMM, isso porque não foi encontrado nenhum transportador completo: dos três domínios de ligação a ATP, somente um deles foi encontrado, conforme mostra o número entre parênteses nessa coluna.

5.7.2. Análise de similaridade com outros organismos

As subunidades de ligação a ATP, por serem as mais conservadas na família ABC (TOMMI & KANEHISA, 1998), foram escolhidas para serem analisadas quanto à similaridade com outros organismos. Nessa análise foi utilizado o banco de dados "não redundante" do NCBI (www.ncbi.nlm.nih.gov). Como esperado, essas proteínas apresentaram similares em sua maioria com Proteobacteria (tabela 11), distribuídas entre as classes Betaproteobacteria (grande similaridade com as Burkholderiales), Gammaproteobacteria (apresentando similaridades com as ordens Pseudomonadales e Enterobacteriales), Alfaproteobacteria (similaridade com membros das ordens Rhizobiales e Rhodobacterales).

Alguns transportadores apresentaram similaridade também com o filo Firmicutes (um dos transportadores de cadeias ramificadas de aminoácidos apresentou similaridade com membros da classe Bacilli e um transportador de sulfonato/nitrato/taurino com a classe Clostridia) e com o filo Cyanobacteria (similaridade com Chroococcales, e um transportador de nitrato/sulfonato/bicarbonato apresentou similaridade também com membros da ordem Oscillatoriales).

As ORF que apresentaram similaridade com esses organismos foram analisadas quanto a proporção de GC na ORF e GC na terceira base dos códons, segundo os resultados do programa CODONW.

A média da proporção de GC nos códons nas ORF de *H. seropedicae* é de 0,64, e a de GC na terceira base dos códons é de 0,83. A ORFID277.0338 esteve abaixo das médias (0,62 de GC e 0,74 de códons com GC na terceira base) e

apresentou valores de CAI 0,36 e Nc 44,97 o que mostra que esta ORF está fora da média dos valores de tendência (figura 13).

A ORFID253.0322 não apresentou resultados com o uso do programa devido a algum problema na seqüência e não pôde ser avaliada quanto ao uso de códons.

5.7.3. Análise de "motivo C"

O programa KAAS identificou 105 possíveis proteínas para o transporte ABC como sendo subunidades de ligação a ATP. Essas subunidades contém o motivo sinal, LSGGQ (*liker peptide*) ou motivo C (BIEMANS-OLDEHINKEL et al., 2006), usado como uma "assinatura" para identificar transportadores ABC (DAVIDSON & CHEN, 2004), passando então a ser procurado na seqüência de nucleotídeos dessas subunidades.

Das 105 ORF que correspondem às subunidades de ligação a ATP, o motivo C foi encontrado em somente 35 delas. O número de ORF que possuem o motivo C dentre as 548 possíveis proteínas ABC encontradas por pesquisa BLAST contra o banco de dados TransportDB, foi de 43.

Tabela 11 – Proximidade taxonômica entre as subunidades para ligação de ATP em transportadores ABC de *H. seropedicae* e proteínas do mesmo tipo em outros organismos

ORF	ORGANISMOS DE MAIOR SIMILARIDADE[1]
ORFID240.0035	*Pseudomonas entomophila,P. putida,P. syringae,P. fluorescens,P. aeruginosa*
ORFID171.0044	*Burkholderia xenovorans,B. phymatum,B. phytofirmans,B. pseudomallei, B. mallei,B. dolosa*
ORFID183.0064	*Bordetella parapertussis,B. bronchiseptica,B. pertussis,Desulfovibrio vulgaris,Acidovorax sp.,Lawsonia intracellularis,Comamonas testosteroni,Stappia aggregata,Delftia acidovorans*
ORFID241.0065	*Polaromonas sp.,Rhodopseudomonas palustris,Bradyrhizobium japonicum, Mesorhizobium loti,Xanthobacter autotrophicus,Rhizobium leguminosarum*
ORFID246.0122	*Hahella chejuensis,Marinomonas sp.,Pseudomonas mendocina,Rhizobium etli,Burkholderia ambifaria,B. cepacia,B. phymatum,Serratia proteamaculans,Burkholderia vietnamiensis*
ORFID249.0339	*Serratia proteamaculans,Yersinia frederiksenii,Y. intermedia,Erwinia carotovora,Polaromonas sp.,Polaromonas naphthalenivorans,Rhodoferax ferrireducens,Bacillus sp.,Halobacillus dabanensis,Bacillus clausii*
ORFID256.0265	*Ralstonia solanacearum,Burkholderia phymatum,Pseudomonas fluorescens,Burkholderia dolosa,B. cenocepacia,Bradyrhizobium japonicum,Burkholderia cenocepacia,Burkholderia xenovorans,Burkholderia cepacia*
ORFID264.0634	*Ralstonia pickettii,Rhodoferax ferrireducens,Polaromonas sp.,Ralstonia eutropha,Bordetella pertussis,B. parapertussis,B. bronchiseptica,B. phytofirmans,Verminephrobacter eiseniae,Polaromonas naphthalenivorans*
ORFID253.0446	*Xanthobacter autotrophicus,Agrobacterium tumefaciens,Rhizobium leguminosarum,Pseudomonas syringae,Rhizobium etli,Pseudomonas fluorescens,P. chlororaphis,P. putida*
ORFID250.043	*Ralstonia eutropha,R. metallidurans,Burkholderia sp.,B. cenocepacia,B. vietnamiensis,Comamonas testosteroni,B.pseudomallei*
ORFID204.0074	*Roseobacter sp.,Magnetospirillum gryphiswaldense,Aurantimonas sp.,Verminephrobacter eiseniae,Pseudomonas fluorescens,Roseovarius nubinhibens,Oceanicola batsensis,Stappia aggregata,Delftia acidovorans,Comamonas testosteroni*
ORFID213.0129	*Ralstonia eutropha,R. metallidurans,Verminephrobacter eiseniae,Azoarcus sp.,Dechloromonas aromatica,Rhodobacter sphaeroides,Dinoroseobacter shibae*
ORFID218.0199	*Burkholderia vietnamiensis,Ralstonia eutropha,Burkholderia multivorans,B. cenocepacia,B. sp.,R. metallidurans,B. thailandensis,B. ambifaria,B. cepacia*
ORFID248.0456	*Pseudomonas syringae,Delftia acidovorans,Polaromonas sp.,Methylobacillus flagellatus,Saccharophagus degradans,Methylibium petroleiphilum,Burkholderia xenovorans,Granulibacter bethesdensis,Ralstonia eutropha,Bradyrhizobium japonicum*
ORFID250.0444	*Delftia acidovorans,Ralstonia eutropha,Acidovorax avenae,Comamonas testosteroni,Ralstonia eutropha,Acidovorax sp.,R. metallidurans,R. pickettii*
ORFID281.0235	*Burkholderia phytofirmans,Burkholderia cenocepacia,Burkholderia xenovorans,Ralstonia metallidurans,Burkholderia vietnamiensis,Delftia acidovorans,Ralstonia pickettii,Acidovorax avenae,Paracoccus denitrificans*

continua

	continuação
ORFID282.0466	*Herminiimonas arsenicoxydans,Methylibium petroleiphilum,Polaromonas naphthalenivorans,Comamonas testosteroni,Polaromonas sp.,Acidovorax sp.,Acidovorax avenae,Rhodoferax ferrireducens,Ralstonia eutropha,Delftia acidovorans*
ORFID287.0527	*Agrobacterium tumefaciens,Mesorhizobium loti,Bradyrhizobium sp.,Paracoccus denitrificans,Xanthobacter autotrophicus,Dechloromonas aromatica,Verminephrobacter eiseniae*
ORFID204.0081	*Pseudomonas fluorescens,Polaromonas sp.,Acidovorax avenae,Azoarcus sp.,Dechloromonas aromatica,Cupriavidus necator,Delftia acidovorans,Pseudomonas chlororaphis,Bordetella pertussis,Verminephrobacter eiseniae,Delftia acidovorans*
ORFID213.0173	*Verminephrobacter eiseniae,Azoarcus sp.,Dechloromonas aromatica,Ralstonia eutropha,Ralstonia metallidurans,Bordetella avium,B. parapertussis,B. bronchiseptica,Polaromonas naphthalenivorans*
ORFID218.0193	*Burkholderia cenocepacia,B. sp.,B. vietnamiensis,B. cepacia,B. ambifaria,B. dolosa,B. multivorans,Ralstonia eutropha*
ORFID248.0450	*Pseudomonas syringae,Saccharophagus degradans,Methylobacillus flagellatus,Methylibium petroleiphilum,Polaromonas sp.,Burkholderia xenovorans,Bradyrhizobium japonicum,Bradyrhizobium sp.,Burkholderia cenocepacia*
ORFID250.0438	*Ralstonia eutropha,R. metallidurans,Burkholderia sp.,B. cenocepacia,B. vietnamiensis,Comamonas testosteroni,B.pseudomallei*
ORFID281.0241	*Burkholderia phytofirmans,Burkholderia vietnamiensis,Burkholderia xenovorans,Acidovorax avenae,Burkholderia cenocepacia,Ralstonia metallidurans,Ralstonia pickettii,Methylobacterium sp.,Delftia acidovorans*
ORFID282.0473	*Herminiimonas arsenicoxydans,Ralstonia eutropha,Ralstonia pickettii,Ralstonia eutropha,Ralstonia metallidurans,Polaromonas sp.,Ralstonia solanacearum,Rhodoferax ferrireducens,Acidovorax sp.*
ORFID287.0477	*Mesorhizobium loti,Agrobacterium tumefaciens,Bradyrhizobium sp.,Polaromonas naphthalenivorans,Stappia aggregata,Pseudomonas syringae,Roseobacter sp.,Silicibacter sp.*
ORFID257.0283	*Burkholderia multivorans,Burkholderia vietnamiensis,Ralstonia pickettii,Burkholderia sp.,Burkholderia phymatum,Burkholderia xenovorans,Burkholderia cenocepacia,Burkholderia multivorans,Burkholderia phytofirmans,Burkholderia cepacia,Burkholderia ambifaria*
ORFID237.0304	*Burkholderia xenovorans,Burkholderia sp.,Burkholderia cenocepacia,Ralstonia metallidurans,Ralstonia eutropha,Pseudomonas putida*
ORFID265.0480	*Herminiimonas arsenicoxydans,Phosphate import,Polynucleobacter sp.,Verminephrobacter eiseniae,Bordetella avium,Bordetella pertussis,Bordetella parapertussis,Bordetella bronchiseptica,Delftia acidovorans,Comamonas testosteroni,Thiobacillus denitrificans*
ORFID199.0206	*Herminiimonas arsenicoxydans,Ralstonia pickettii,Ralstonia solanacearum,Ralstonia metallidurans,Ralstonia eutropha,Polaromonas sp.,Polaromonas naphthalenivorans,Acidovorax avenae,Acidovorax sp.*
ORFID181.0014	*Rhodoferax ferrireducens,Ralstonia eutropha,Ralstonia metallidurans,Burkholderia phytofirmans,Burkholderia xenovorans,Burkholderia phymatum,Ralstonia pickettii,Ralstonia solanacearum,Bordetella parapertussis,B. bronchiseptica*
	continua

	continuação
ORFID196.0059	*Herminiimonas arsenicoxydans,Methylibium petroleiphilum,Polaromonas naphthalenivorans,Ralstonia pickettii,Polaromonas sp.,Ralstonia eutropha,Comamonas testosteroni,Rhodoferax ferrireducens,Ralstonia metallidurans,Delftia acidovorans*
ORFID199.0160	*Ralstonia metallidurans,Methylibium petroleiphilum,Comamonas testosteroni,Rhodoferax ferrireducens,Ralstonia eutropha,Ralstonia pickettii,Ralstonia solanacearum,Delftia acidovorans,Acidovorax avenae*
ORFID200.0137	*Burkholderia phymatum,Burkholderia phytofirmans,Burkholderia xenovorans,Burkholderia multivorans,Burkholderia vietnamiensis,Burkholderia thailandensis,Burkholderia sp.,Burkholderia ambifaria,Burkholderia cenocepacia*
ORFID226.0150	*Burkholderia xenovorans,Pseudomonas putida,Acinetobacter sp.,Comamonas testosteroni,Bradyrhizobium sp.,Synechococcus sp.,Trichodesmium erythraeum,Lyngbya sp.*
ORFID236.0335	*Methylobacterium sp.,Bradyrhizobium sp.,Ralstonia pickettii,Herminiimonas arsenicoxydans,Rhodobacter sphaeroides,Ralstonia solanacearum,Rhodobacter sphaeroides,Bradyrhizobium japonicum*
ORFID249.0401	*Pseudomonas syringae,Pseudomonas fluorescens,Pseudomonas entomophila,Pseudomonas fluorescens,Azotobacter vinelandii,Erwinia carotovora,Burkholderia ambifaria,Ralstonia metallidurans*
ORFID252.0413	*Pseudomonas stutzeri,Bordetella avium,Bordetella pertussis,Bordetella bronchiseptica,Bordetella parapertussis,Moorella thermoacetica,Desulfitobacterium hafniense,Methanosaeta thermophila,Desulfitobacterium hafniense*
ORFID272.0651	*Dechloromonas aromatica,Anaeromyxobacter dehalogenans,Magnetospirillum magnetotacticum,Yersinia pestis,Anaeromyxobacter sp.,Yersinia pseudotuberculosis,Xanthobacter autotrophicus,Bdellovibrio bacteriovorus,Methylobacterium sp.*
ORFID253.0322	*Ralstonia metallidurans,Ralstonia eutropha,Delftia acidovorans,Comamonas testosteroni,Verminephrobacter eiseniae,Azoarcus sp.,Acidovorax avenae,Agrobacterium tumefaciens,Xanthobacter autotrophicus,Synechococcus sp.*
ORFID215.0135	*Pseudomonas stutzeri,Pseudomonas,P. putida,Pseudomonas aeruginosa,Marinobacter sp.,Pseudomonas mendocina,Pseudomonas syringae*
ORFID287.0053	*Burkholderia multivorans,Ralstonia pickettii,Burkholderia vietnamiensis,Burkholderia xenovorans,Burkholderia phytofirmans,Burkholderia cenocepacia,Burkholderia phymatum,Burkholderia sp.,Bburkholderia dolosa,Ralstonia solanacearum*
ORFID215.0195	*Verminephrobacter eiseniae,Ralstonia pickettii,Ralstonia solanacearum,Bordetella avium, Pseudomonas syringae,Roseobacter sp.*
ORFID271.0076	*Verminephrobacter eiseniae,Marinomonas sp.,Pseudomonas syringae,Roseobacter sp.,Stappia aggregata,Roseovarius nubinhibens,Silicibacter sp.,Roseovarius sp.,Marinomonas sp.*
ORFID256.0603	*Burkholderia dolosa,Pseudomonas syringae,Burkholderia phytofirmans,Burkholderia xenovorans,Burkholderia ambifaria,Burkholderia cepacia,Burkholderia cenocepacia*
ORFID274.0538	*Herminiimonas arsenicoxydans,Ralstonia metallidurans,Ralstonia pickettii,Ralstonia solanacearum,Ralstonia eutropha,Ralstonia eutropha,Stappia aggregata,Burkholderia phymatum,Burkholderia sp.* ***continua***

	continuação
ORFID271.0066	*Verminephrobacter eiseniae,Marinomonas sp.,Pseudomonas syringae,Pseudomonas syringae,Roseovarius sp.,Roseovarius nubinhibens,Roseobacter sp.,Silicibacter sp.,Stappia aggregata*
ORFID205.0112	*Rhodoferax ferrireducens,Polaromonas naphthalenivorans,Rhizobium etli,Rhizobium leguminosarum,Pseudomonas syringae,Bradyrhizobium sp.,Nitrobacter hamburgensis,Xanthobacter autotrophicus*
ORFID280.0689	*Ralstonia solanacearum,Ralstonia pickettii,Ralstonia solanacearum,Bordetella avium,Serratia proteamaculans,Yersinia pestis,Yersinia pseudotuberculosis*
ORFID283.0556	*Bordetella bronchiseptica,Yersinia pseudotuberculosis,Bordetella pertussis,Bordetella parapertussis,Yersinia enterocolitica,Chromobacterium violaceum,Serratia proteamaculans,Photorhabdus luminescens,Erwinia carotovora,Enterobacter sp.*
ORFID284.0569	*Herminiimonas arsenicoxydans,Ralstonia eutropha,Azoarcus sp.,Ralstonia metallidurans,Ralstonia pickettii,Azoarcus sp.,Ralstonia solanacearum,Burkholderia dolosa,Burkholderia thailandensis*
ORFID177.0060	*Ralstonia eutropha,Ralstonia metallidurans,Ralstonia eutropha,Burkholderia sp.,Burkholderia cenocepacia,Burkholderia multivorans,Burkholderia pseudomallei,Burkholderia mallei,Burkholderia thailandensis,Ralstonia pickettii*
ORFID184.0136	*Burkholderia multivorans,Burkholderia thailandensis,Burkholderia vietnamiensis,Burkholderia phytofirmans,Ralstonia eutropha*
ORFID277.0338	*Bordetella bronchiseptica,Bordetella parapertussis,Vibrionales bacterium,Vibrio splendidus,Vibrio sp.,Marinomonas sp.,Vibrio shilonii,Photobacterium profundum, Marinomonas sp.,Desulfovibrio vulgaris,Synechococcus sp.*

*[1] Os organismos listados referem-se à fonte das seqüências que obtiveram melhores hits em uma pesquisa BLASTX contra o banco de dados **nr** (não redundante) do NCBI.*

Como esperado, essas proteínas apresentaram similares em sua maioria com Proteobacteria. Alguns transportadores apresentaram similaridade também com o filo Firmicutes (ORFID249.0339, ORFID252.0413) e com o filo Cyanobacteria (ORFID226.0150, ORFID253.0322).

5.7.4. Similaridade e conservação da "vizinhança" entre os genes de transportadores ABC em *H. seropedicae* e bactérias relacionadas

Através da análise do KAAS os transportadores constituintes das vias metabólicas encontradas em *H. seropedicae* foi possível encontrar as ORF que correspondem a cada proteína dessas vias, tais como sistemas de secreção tipo II e III, proteínas da família ABC e proteínas do tipo PTS.

Entre elas, o mapa dos transportadores da família ABC foi melhor investigado, encontrando-se cada uma das ORF correspondentes às subunidades de proteínas transportadoras completas. Sendo assim, essas ORF foram agrupadas em possíveis operons, conforme indícios de suas proximidades dentro do genoma de *H. seropedicae* (tabelas 9 e 10).

Esses agrupamentos foram submetidos à análise através da ferramenta STRING (http://string.embl.de/), visando observar as relações dessas ORF em outros organismos, além de outras informações, tais como a fusão dos genes e coocorrência dos genes/proteínas. A figura 14 mostra as vias de transporte realizadas por proteínas da família ABC.

Pode-se notar nos exemplos que os genes dessas proteínas possuem evidência de estar juntos coocorrendo nos outros genomas (*neighborhood* e *cooccurrence*) como já era esperado (KAAS), reforçando a idéia de que esses genes ocorram em *operons*.

Em alguns casos (ex., transportadores de aminoácidos polares) o gene que transcreve para a permease pode estar fusionado com o gene que transcreve para a

subunidade de ligação ao substrato, como ocorre nos genomas de *Burkholderia,* tal como no genoma de *B. mallei* , entre outros organismos (figura 15).

Figura 14 – Rede associativa para as proteínas da família ABC, realizada com o programa STRING

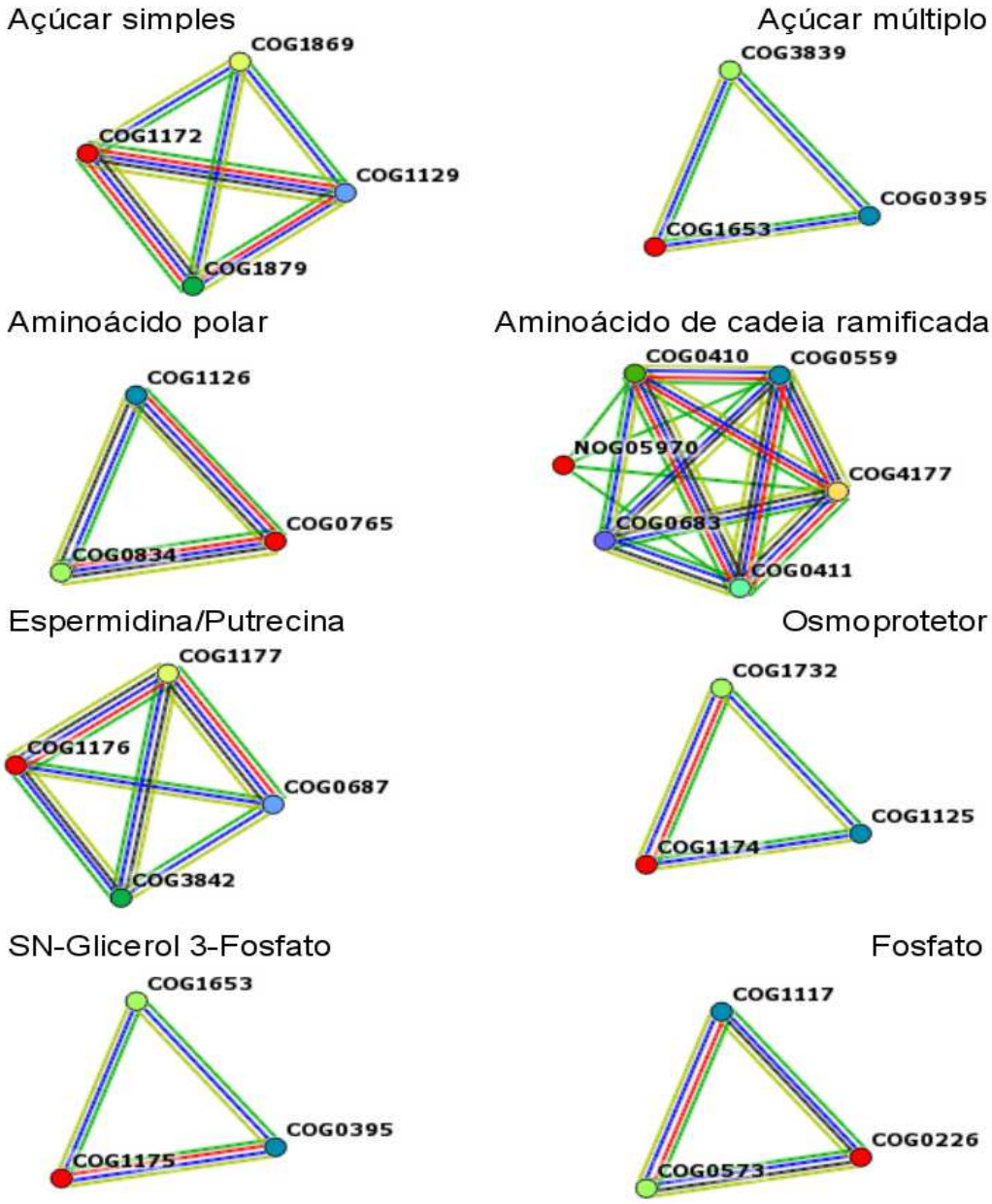
Cada linha indica um tipo de evidência usada na previsão das associações.

a) vermelha – indica a evidência de fusão entre os genes das proteínas;

b) verde – indica a evidência de vizinhança entre as proteínas;

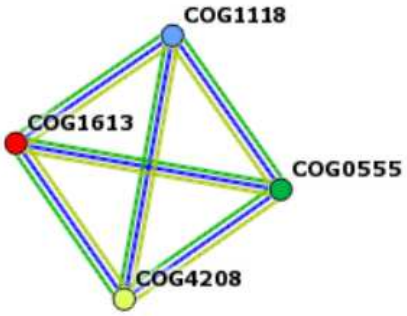
c) azul – indica a evidência de coocorrência das proteínas em outros organismos;

d) preto – indica a evidência de coexpressão das proteínas.

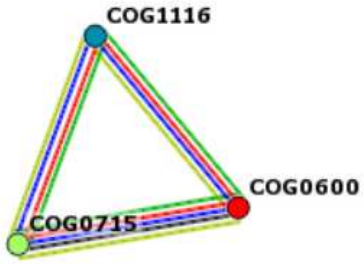


continua

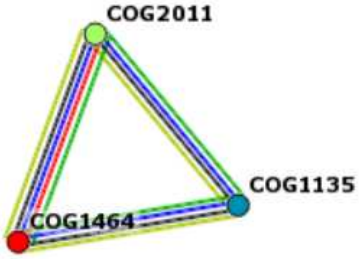
Sulfato



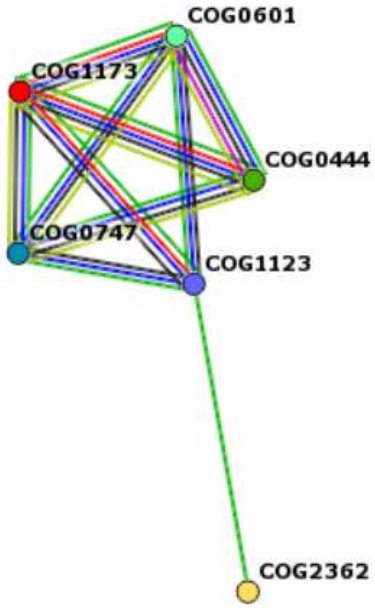
Sulfonato/Nitrato/Taurino



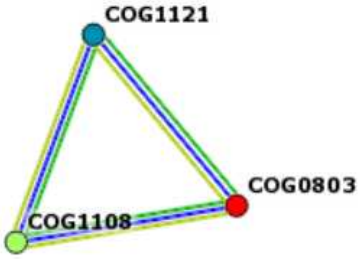
D-Metionina



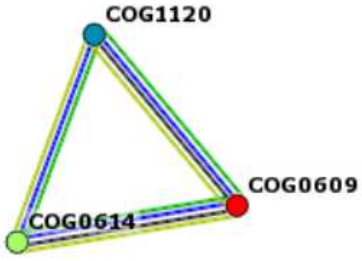
Peptídio/Níquel



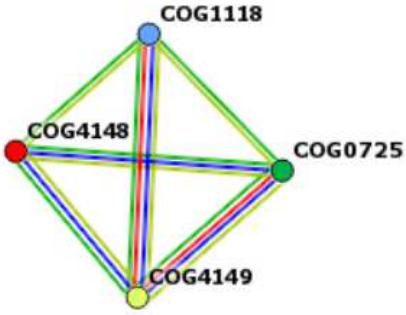
Zinco/Manganês



Complexo de Ferro

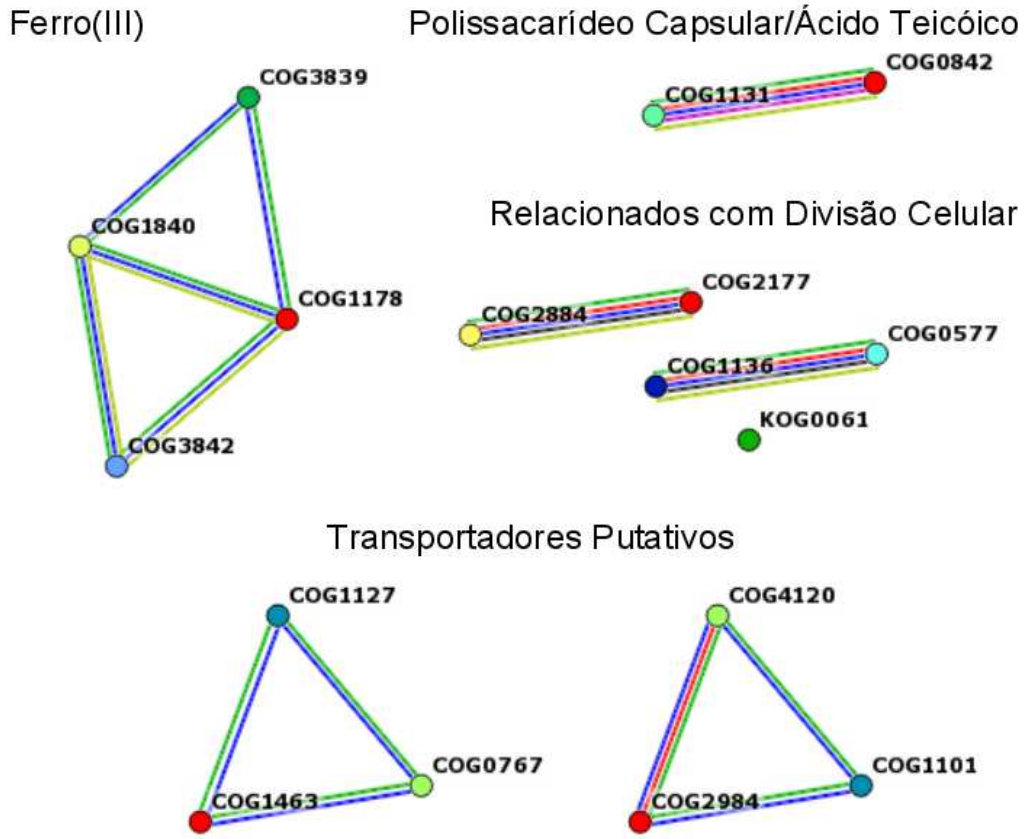


Molibdato



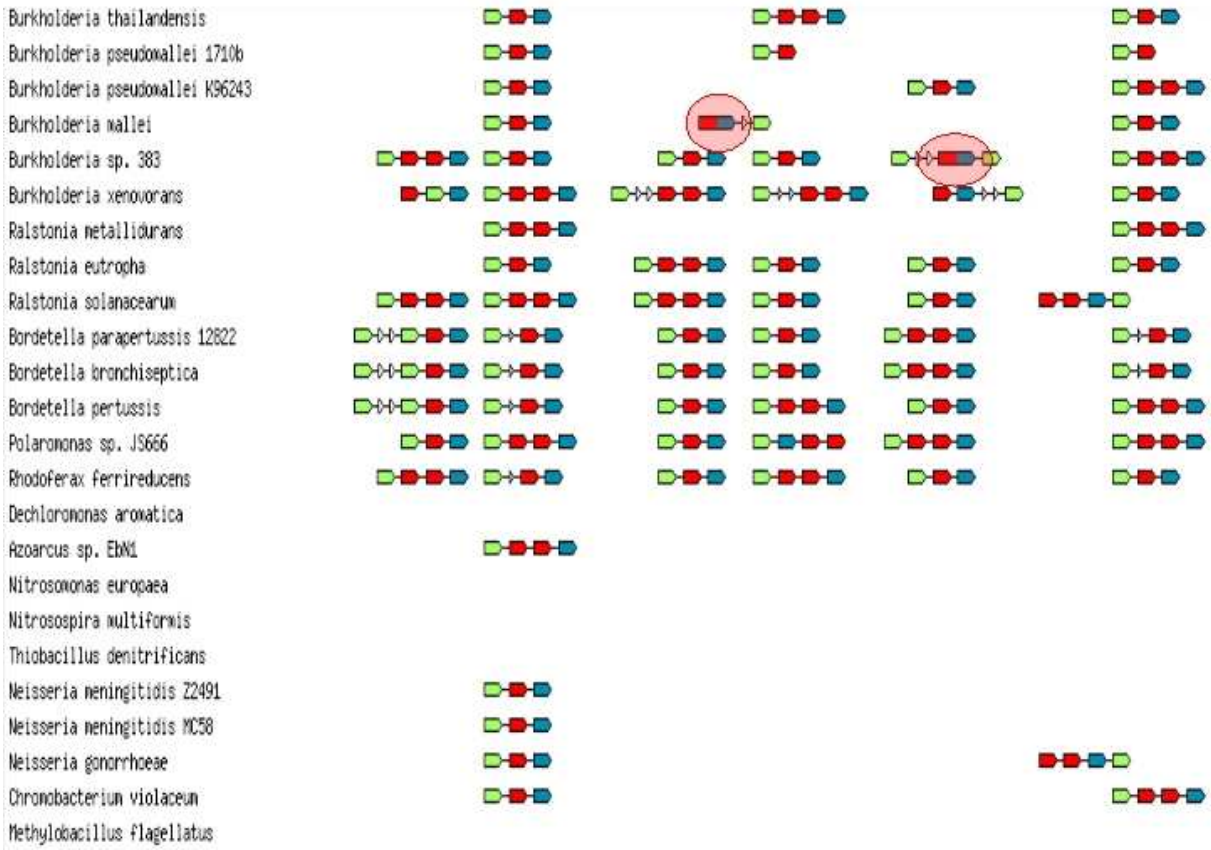
continua

continuação



A figura resume a rede de associações previstas para um grupo particular de proteínas. Os nós da rede são grupos de proteínas equivalentes (ex., ortólogos). As linhas de ligação entre os nós representam a associação funcional prevista na análise.

Figura 15 – Um exemplo de *neighborhood*



Essa é a disposição de alguns dos genes encontrados nos genomas de Betaproteobacteria e que transcrevem para as subunidades protéicas de transporte de aminoácidos polares, realizados por proteínas ABC. Exemplos de fusão gênicas estão marcados em círculos vermelhos.

5.7.5. Análise filogenética

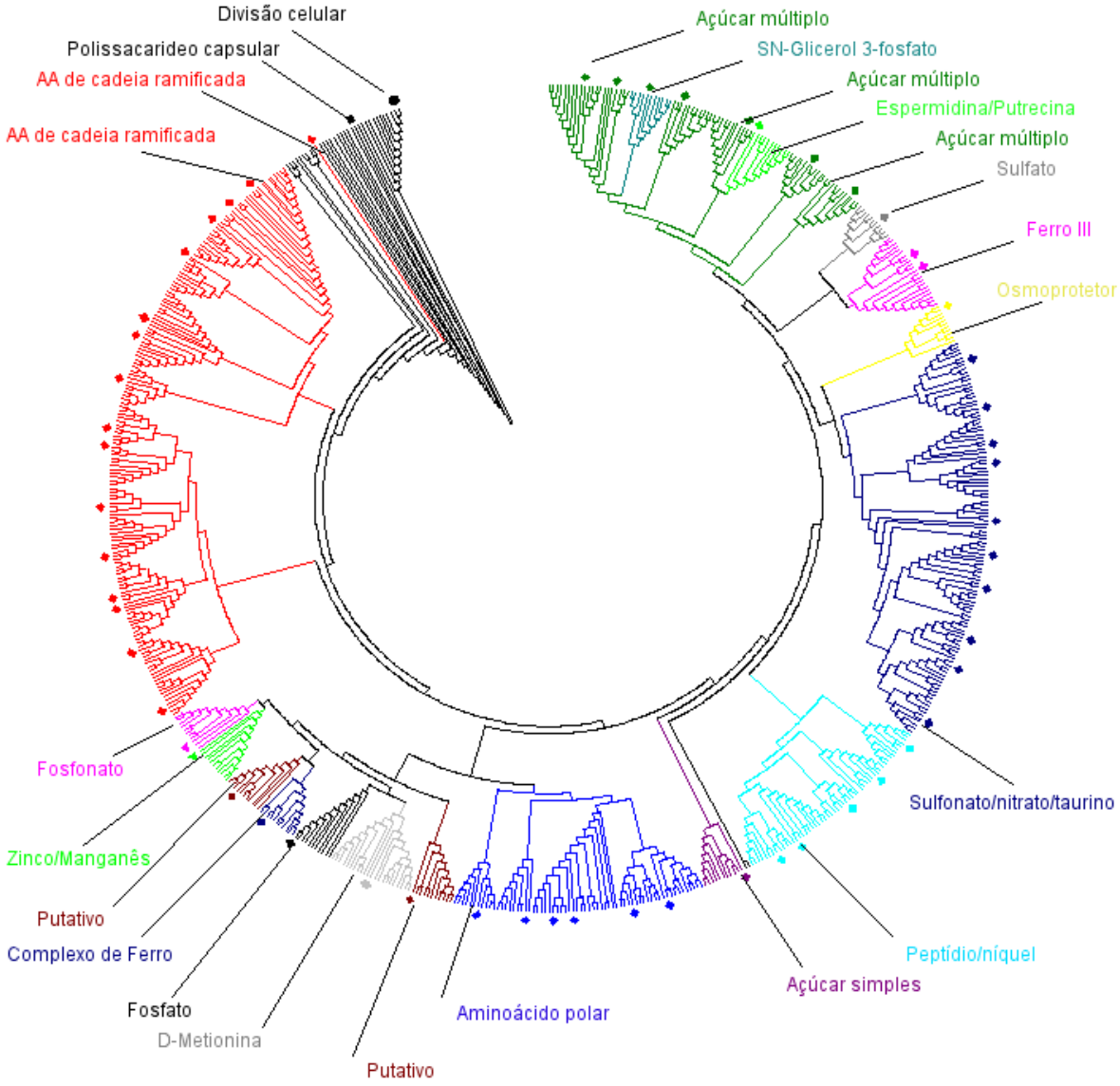
As 60 ORF identificadas como subunidades ligadoras de ATP dos *operons* completos ou parcialmente completos de transportadores ABC de *H. seropedicae,* foram submetidas a uma análise filogenética com seqüências relacionadas obtidas a partir de uma pesquisa BLASTX contra o banco de dados **nr** (não redundante) do NCBI (figura 16).

As ORF de *H. seropedicae* mostraram proximidade com seqüências para sub-unidades ligadoras de ATP de transportadores ABC de organismos da sub-classe das Proteobacteria (tabela 11). Entretanto, algumas ORF (ORFID249.0339; ORFID226.0150; ORFID252.0413; ORFID253.0322; ORFID277.0338) mostraram proximidade com organismos de grupos distantes das Proteobacteria, como os filos Firmicutes e Cyanobacteria, sugerindo uma origem evolutiva diferente para estas proteínas.

Em geral, os transportadores formaram grupos bem definidos, com as sub-unidades específicas participando de sub-grupos. Alguns grupos mostraram relativa proximidade evolutiva, como os transportadores para Ferro III e Sulfato, ou ficaram próximos a outros grupos, como os transportadores para Espermidina/Putrecina e sn-glicerol-3-fosfato (intercalados entre sub-grupos dos transportadores para Açúcar Múltiplo), sugerindo uma origem evolutiva comum para diferentes transportadores.

Os resultados obtidos para a análise filogenética das unidades ligadoras de ATP de transportadores ABC são preliminares e necessitam uma análise mais detalhada para a confirmação dos dados e reconstrução mais confiável da sua história evolutiva.

Figura 16 – Árvore filogenética das unidades ligadoras de ATP de transportadores ABC



*A análise foi realizada com um total de 655 seqüências de proteínas. As seqüências foram alinhadas com o programa ClustalW (THOMPSON et al., 1994); as distâncias genéticas calculadas com a matriz de substituição PAM (Point Accepted Mutation – Dayhoff et al., 1978) e a árvore obtida pelo método de Neighbour-Joining (*SAITOU & NEI, 1987*).*
Cada ponto no lado externo da figura corresponde a uma ORF de H. seropedicae; as linhas não marcadas com pontos são genes similares a elas. Genes para o transporte de um mesmo substrato são marcados com a mesma cor (com exceção da cor preta, que marca transporte de polissacarídeos capsulares e proteínas relacionados com divisão celular).
As proteínas que participam de divisão celular e as que fazem transporte de polissacarídeos, provavelmente não são unidades de ligação à ATP (o fato de serem proteínas de ligação a ATP não estava evidente no mapa montado pelo programa KAAS), e ocuparam a posição de "out group" ou "grupo externo". Transportadores de um mesmo substrato ficaram agrupados na árvore, embora alguns como SN-glicerol 3-fosfato e Espermidina/putrecina tenham ficado entre transportadores de açúcares múltiplos, sugerindo "parentesco" entre esses transportadores.

6. CONCLUSÕES

Embora 880 possíveis proteínas transportadoras tenham sido identificadas por pesquisa de similaridade BLAST contra o banco de dados TransportDB, em torno de 537 delas são comuns também a outros dois bancos de dados utilizados (TCDB e KEGG); 543 das 880 constam como válidas na anotação do projeto GENOPAR; 548 foram classificadas como “alta-similaridade” pela rede neuronal FAN; e 370 são comuns a todas as análises.

Isso sugere que o número de proteínas transportadoras de *H. seropedicae* esteja entre aproximadamente 550 e aproximadamente 370 proteínas.

Dentre as proteínas identificadas não foi encontrado sistema de transporte do tipo PTS, mas somente alguns de seus componentes (HPr, subunidades da EII). Uma proteína HPr encontrada com uso do programa KAAS pode indicar algum regulador e estar envolvida em algum outro processo; a subunidade PtsN deve fazer parte de alguma via metabólica envolvendo nitrogênio, e não necessariamente faz parte de uma via de transporte.

Bactérias relacionadas a *H. seropedicae* (*Xylella fastidiosa*, *Xanthomonas campestris*, *Neisseria meningitidis*, *Ralstonia solanacearum*, *Burkholderia pseudomallei*) também apresentam somente alguns componentes desse sistema.

Já o Sistema de Secreção do Tipo III foi encontrado em *H. seropedicae*, estando ausentes os componentes: YscF, YscL e YscQ. Em outras bactérias essas unidades também podem estar ausentes, como em *Pseudomonas syringae*, mas nem por isso o sistema deixa de ser funcional (GALAN & COLLMER, 1999).

A maioria das proteínas transportadoras identificadas seguem a tendência de

uso de códons do genoma, sofrendo forte pressão de seleção do conteúdo CG, o que mostra que estão adaptadas ao genoma de *H. seropedicae*, e surgiram cedo no processo evolutivo dessa bactéria .

As proteínas que não seguem essa adaptação talvez tenham sido adquiridas de outra bactérias por transferência lateral; no entanto essas proteínas teriam que ser melhor estudadas para reforçar essa hipótese.

Os organismos *H. seropedicae* e *H. rubrisubalbicans* possuem aparentemente capacidades de transporte semelhantes, no entanto essa análise poderia ser novamente realizada tendo-se em mãos uma melhor montagem do genoma de *H. rubrisubalbicans*, visto que a montagem utilizada possui mais de 2.000 contigs, o que pode ter dificultado a busca por similaridade.

Grande parte das proteínas transportadoras identificadas, em *H. seropedicae* (60%, considerando as 880), pertencem à família de transportadores ABC. 54 transportadores desse tipo (37,9%) estão completos (possuem todas as subunidades, conforme a montagem do mapa da via de transporte realizada pelo programa KAAS), formando grupos de genes que sugerem a existência de *operons*.

Filogeneticamente esses transportadores estão próximos dos encontrados em outras Proteobacteria, raro algumas exceções. Transportadores ABC para um mesmo substrato parecem estar mais próximos entre si na árvore filogenética montada para eles, e mantém relações de vizinhança em outros genomas, reforçando a idéia de que possam fazer parte de um mesmo *operon.*

7. REFERÊNCIAS BIBLIOGRÁFICAS

ALTSCHUL, S.F.; MADDEN, T.L.; SCHÄFFER, A.A.; ZHANG, J.; ZHANG, Z.; MILLER W.; LIPMAN, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Res.** 25:3389-3402.

ANNILO, T.; CHEN, Z-Q.; SHULENIN, S.; COSTANTINO, J.; THOMAS, L.; LOU, H.; STEFANOV, S.; DEAN, M. (2006). Evolution of the vertebrate ABC gene family: Analysis of gene birth and death. **Genomics** 88:1-11.

BALDANI, J.I.; BALDANI, V.L.D. (2004). History on the biological nitrogen fixation research in graminaceous plants: special emphasys on the Brazilian experience. **Anais da Academia Brasileira de Ciências** (2005) 77(3):549-579.

BALDANI, J.I.; BALDANI, V.L.D.; SELDIN, L.; DÖBEREINER, J. (1986). Characterization of *Herbaspirillum seropedicae* gen. nov., sp. nov., a root-associated nitrogen-fixing bacterium. **Int. J. Syst. Bacteriol.**, 36:86–93.

BALDANI, J.I.; CARUSO, L.; BALDANI, V.L.D.; GOI, S.R.; DÖBEREINER, J. (1997). Recent advences in BNF with non-legume plants. **Soil Biology & Biochemistry**, 29: 911-922.

BALDANI, J.I.; POT B.; KIRCHHOF G.; FALSEN E.; BALDANI, V.L.D.; OLIVARES F.L.; HOSTE B.; KERSTERS K.; HARTMANN A.; GILLIS M.; DÖBEREINER, J. (1996). Emended Description of Herbaspirillum; Inclusion of [*Pseudomonas*] *rubrisubalbicans*, a Mild Plant Pathogen, as *Herbaspirillum rubrisubalbicans* comb. nov.; and Classification of a Group of Clinical Isolates (EF Group 1) as Herbaspirillum Species 3. **International Journal of Systematic Bacteriology**, 46:802–810.

BARABOTE, R.V.; RENDULIC, S.; SCHUSTER, S.C.; SAIER, M.H.Jr. (2007). Comprehensive analysis of transport proteins encoded within the genome of *Bdellovibrio bacteriovorus*. **Genomics**, 90:424-446.

BARNEY, B.M.; LEE H-I.; DOS SANTOS, P.C.; HOFFMAN B.M.; DEAN D.R.; SEEFELDT L.C. (2006). Breaking the N2 triple bond: insights into the nitrogenase mechanism. **Dalton Trans.**; 21(19):2277-84.

BEDELL J.; KORF I.; YANDELL M. **BLAST**. O'Reilly & Associates, Inc.; 2003.

BIEMANS-OLDEHINKEL, E.; DOEVEN, M.K.; POOLMAN, B. (2006). ABC transporter architecture and regulatory roles of accessory domains. **FEBS letters**, v. 580, p. 1023-1035.

BIOPERL. Disponível em: <www.bioperl.org> Acesso em: março de 2007.

BUSCH W.; SAYER M.H.Jr. (2002). The transporter classification (TC) system, 2002. **Crit. Rev. Biochem. Mol. Biol.**; 37:287-337.

CAI Calculator. Freeland Lab, Biological Sciences Departament, UMBC. Disponível em: <http://www.evolvingcode.net/codon/cai/cai.php#> Atualizado em: 3 abr 2006; acesso em jul 2007.

DAVIDSON A.L.; CHEN J. (2004). ATP-Binding Cassete Transporters in Bacteria. **Annu. Rev. Biochem.**; 73:241-68.

DAYHOFF, M.O., SSHWARTZ, R.M., ORCUTT, B.C. (1978) A model of evolutionary change in proteins. **Atlas of Protein Sequence and Structure** 5(3) M.O. Dayhoff (ed.), 345-352.

DING L.; YOKOTA A. (2004). Proposals of *Curvibacter gracilis* gen. nov., sp. nov. and *Herbaspirillum putei* sp. nov. for bacterial strains isolated from well water and reclassification of [*Pseudomonas*] *huttiensis*, [*Pseudomonas*] *lanceolata*, [*Aquaspirillum*] *delicatum* and [*Aquaspirillum*] *autotrophicum* as *Herbaspirillum huttiense* comb. nov., *Curvibacter lanceolatus* comb. nov., *Curvibacter delicatus* comb. nov. and *Herbaspirillum autotrophicum* comb. nov. **International Journal of Systematic and Evolutionary Microbiology** 54:2223–2230.

DINH T.; PAULSEN I.T.; SAYER M.H.Jr. (1994). A Family of Extracytoplasmic Proteins That Allow Transport of Large Molecules across the Outer Membranes of Gram-Negative Bacteria. **Journal of Bacteriology**, 176:3825-3831.

EWING B.; GRENN P. (1998). Basecalling of automated sequencer traces using phred. I. Accuracy assessment. **Genome Research**, 8:175-85.

EWING B.; GRENN P. (1998). Basecalling of automated sequencer traces using phred. II. Error probabilities. **Genome Research**, 8:186-194.

FICHANT G.; BASSE M.J.; QUENTIN Y. (2006) ABCdb: an online resource for ABC transporter repertories from sequenced archaeal and bacterial genomes. **FEMS Microbiol Lett.**; 256(2):333-9.

FOSKETT J.K. (1998). CIC and CFTR Chloride Channel Gating. **Annu. Rev. Physiol.** 60:689-717.

GALAN, J.; COLLMER, A. (1999). Type III secretion machines; bacterial devices for protein delivery into host cells. **Science**, 284:1322-1328.

GARRETT, L.F.V; IGNÁCIO, F.A; KÜSTER, C.W.; LENFERS, F. P; ZOTTO, S. P. (2006). **EasyFan**. UFPR.

GENOPAR: Genoma estrutural da bactéria fixadora de nitrogênio endofítica *Herbaspirillum seropedicae*. Disponível em: <www.genopar.org/index.htm> Acesso

em fev 2007.

GIBAS, C.; JAMBECK P. **Developing Bioinformatics Computer Skills**. O'Reilly & Associates, Inc. 2001.

GRANGEIRO, T.B.; JORGE, M.M.; BEZERRA, W.M.; VASCONCELOS, T.R.; SIMPSON, A.J.G. (2004) Transport genes of *Chromobacterium violaceum*: an overview. **Genetics and Molecular Research**, 3:117-133.

HUANG, X.; MADAN, A. 1999. CAP3: A DNA Sequence Assembly Program. **Genome Res.**; 9: 868-877.

HURST A.C.; PETROV E.; KLODA A.; NGUYEN T.; HOOL L.; MARTINAC B. (2007). MscS, the bacterial mechanosensitive channel of small conductance. **The Internacional Journal of Biochemestry & Cell Biology.**

IM W-T.; BAE H-S.; YOKOTA A.; LEE S.T. (2004). *Herbaspirillum chlorophenolicum* sp. nov., a 4-chlorophenol-degrading bacterium. **International Journal of Systematic and Evolutionary Microbiology**, 54:851–855.

JACK D.L.; YANG N.M.; SAYER M.H.Jr. (2001). The drug/metabolite transporter superfamily. **Eur. J. Biochem.**, 268:3620-3639.

KAMMLER M.; SCHÖN C.; HANTKE K. (1993). Characterization of the ferrous iron uptake system of *Escherichia coli*. **J Bacteriol.** 175:6212-9.

KANEHISA M. (2002). The KEGG database. **Novartis Found Symp.**; 247:91-101.

KELLY D.J.; THOMAS G.H. (2001). The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea. **FEMS Microbiology Reviews** 25:405-424.

KIMBROUGH, T.G.; MILLER, S.I. (2002) Assembly of the type III secretion needle complex of *Salmonella typhimurium*. **Microbes Infect.**, 4: 75-82.

KIRCHHOFF G.; ECKERT B.; STOFFELS M.; BALDANI J.I.; REIS V.M.; HARTMANN A. (2001). *Herbaspirillum frisingense* sp. nov., a new nitrogen-fixing bacterial species that occurs in C4-fibre plants. **International Journal of Systematic and Evolutionary Microbiology**, 51:157–168.

KOMORIYA, K.; SHIBANO, N.; HIGANO, T.; AZUMA, N.; YAMAGUCHI, S.; AIZAWA, S. (1999). Flagellar proteins and type III-exported virulence factors are the predominant proteins secreted into the culture media of *Salmonela typhimurium*. **Mol. Microbiol.**, 34:767-779.

KONINGS W. N. (2006). Microbial transport: Adaptations to natural environments. **Antonie van Leeuwenhoek** 90:325–342.

KROGH A.; LARSSON B.; von HEIJNE G.; SONNHAMMER E. L. L. (2001). Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genome. **J. Mol. Biol.**, 305:567-580.

KUNDIG, W.; GOSH, S; ROSEMAN S. (1964). Phosphate Bound to Histidine in a Protein as an Intermediate in a Novel Phosphotransferase System. **Proc. Natl. Acad. Sci. USA** 52:1067-1074.

LENGELER, J.; DREWS, G.; SCHLEGEL, H. **Biology of the Prokariotes.** Blackwell Science Ltd. 1999.

LUNIN V.V.; DOBROVETSKY E.; KHUTORESKAYA G.; ZHANG R.; JOACHIMIAK A.; DOYLE D.A.; BOCHKAREV A.; MAGUIRE M.E.; EDWARDS A.M; KOTH C.M. (2006). Crystal structure of the CorA Mg2+ transporter. **Nature** 440:833-837.

MARKOVICH D.; MURER H. (2004). The SLC13 gene family of sodium sulphate/carboxylate cotransporters. **Pflugers Arch - Eur J Physiol** 447:594–602.

MCINERNEY, J. O. 1998. GCUA: General Codon Usage Analysis. **Bioinformatics**, 14:372-373.

MITCHELL W.J.; TEWATIA P.; MEADEN P.G. (2007). Genomic analysis of the phosphotransferase system in Clostridium botulinum. **J Mol Microbiol Biotechnol**; 12(1-2):33-42.

MORIYA, Y.; ITOH M.; OKUDA, S.; YOSHIZAWA A.C.; KANEHISA, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. **Nucleic Acids Res.**; 35:182-185.

NCBI - National Center for Biotechnology Information. Atualizada em 12 dez 2006. Disponível em: <http://www.ncbi.nlm.nih.gov> Acesso em fev 2007.

PAO S.S.; PAULSEN I.T.; SAIER M.H. Jr. (1998). Major Facilitator Superfamily. **Microbiology and Molecular Biology Reviews**, 62:1-34.

PAULSEN, I. T.; NGUYEN, L.; SLIWINSKI, M. K.; RABUS, R; SAIER, M. H. (2000). Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. **J. Mol. Biol.** 301:75-100.

PENDEN, J. F. (1999). **Correspondence Analysis of Codon Usage**. Nottingham, 1999. Thesis of Doctor of Philosophy. Dept of Genetics, University of Nottingham.

PHRAP/CROSS_MATCH/SWAT DOCUMENTATION. C1993-1996. Disponível em: <http://www.phrap.org/phredphrap/general.html> Acesso em: ago 2007.

POSTMA, P.W.; LENGELER J. W.; JACOBSON G. R. (1993). Phosphoenolpyruvate:

carbohydrate phosphotransferase systems in bacteria. **Microbiol. Rev.** 57:543-594.

POSTMA, P.W.; LENGELER J. W.; JACOBSON G. R. (1996) in: ***Escherichia coli* and *Salmonella*: cellular and molecular biology** (Neidhardt F.C., et al., Eds.), p 1149, ASM, Washington, DC.

POWELL R.J.; MORIYAMA E.N. (1997). Evolution of codon usage bias in Drosophila. **Proc Natl Acad Sci U S A**; 94:7784–7790.

QIAN H.; SAHLMAN L.; ERIKSSON P-O.; HAMBRAEUS C.; EDLUND U.; SETHSON I. (1998). NMR Solution Structure of the Oxidized Form of MerP, a Mercuric Ion Binding Protein Involved in Bacterial Mercuric Ion Resistance. **Biochemistry**, 37:9316-9322.

RAITTZ, R. **FAN 2002: um modelo neuro-fuzzy para reconhecimento de padrões**. Florianópolis, 2002. Tese de Doutorado em Engenharia de Produção. Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina.

REN, Q.; CHEN, K.; PAULSEN I.T. (2007). TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. **Nucleic Acids Res.**; 35:274-279.

REN, Q.; KANG K.H.; PAULSEN I.T. (2004). TransportDB: a relational database of cellular membrane transport systems. **Nucleic Acids Res.**; 32:284-288

RONCATO-MACCARI L.D.B.; RAMOS H.J.O.; PEDROSA F.O.; ALQUINI Y.; CHUBATSU L.S.; YATES M.G.; RIGO L.U.; STEFFENS M.B.R.; SOUZA E.M. (2003). Endophytic *Herbaspirillum seropedicae* expresses *nif* genes in gramineous plants. **FEMS Microbiology Ecology** 45:39-47.

ROTHBALLER M.; SCHMID M.; KLEIN I.; GATTINGER A.; GRUNDMANN S.; HARTMANN A. (2006). *Herbaspirillum hiltneri* sp. nov., isolated from surface-sterilized wheat roots. **Int. J. Syst. Evol. Microbiol.**, 56:1341-1348.

SAITOU, N.; NEI, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Mol. Evol. Biol.** 4:406-425.

SAURIN, W.; DASSA, E. (1994). Sequence relationships between integral inner membrane proteins of binding protein-dependent transport systems: Evolution by recurrent gene duplications. **Protein Sci.**, 3: 325-344.

SAYER, M.H.Jr. (1994). Computer-Aided Analyses of Transport Protein Sequences:Gleaning Evidence concerning Function,Structure, Biogenesis, and Evolution. **Microbiological Reviews** 58:71-93.

SAYER, M.H.Jr (2000). A Functional-Phylogenetic Classification System for

Transmembrane Solute Transporters. **Microbiology and Molecular Biology Reviews**, 40:354–411.

SAYER, M.H.Jr.; TRAN C. V.; BARABOTE R. D.; (2006). TCDB: the Transporter Classification Database for membrane transport protein analyses and information. **Nucleic Acids Research**, 34:181-186.

SHARP, P.M.; LI, W-H. (1987). The codon adaptation index: a measure of directional synonymous codon usage bias, and its potential applications. **Nucleic Acids Research**, 15:1281-1295.

SIEBOLD, C.; FLÜKIGER, K.; BEUTLER, R.; ERNI, B. (2001) Carbohydrate transporters of the bacterial phosphoenolpyruvate: sugar phosphotransferase system (PTS). **FEBS Letters**, 504:104-111.

TAKATA K.; MATSUZAKI T.; TAJITA Y. (2004). Aquaporins: water channel proteins of the cell membrane. **Progress in Histochemistry and Cytochemistry** 39:1-83.

TAMURA, K.; DUDLEY, J.; NEI, M.; KUMAR, S. (2007). MEGA4: Molecular Evolutionary Genetic Analysis (MEGA) Software Version 4.0. **Mol. Biol. Evol.** 24:1596-1599.

THOMPSON, J.D.; HIGGINS, D.G.; GIBSON, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Res**, 22:4673-80, 1994.

TOMII, K.; KANEHISA, M. (1998). A Comparative Analysis of ABC Transporters in Complete Microbial Genomes. **Genome Res.**, 8:1048-1059.

VALVERDE, A.; VELÁZQUEZ, E.; GUTIÉRREZ, C.; CERVANTES, E.; VENTOSA, A.; IGUAL, J-M. ***Her****baspirillum lusitanum* sp. nov., a novel nitrogen-fixing bacterium associated with root nodules of *Phaseolus vulgaris*. **International Journal of Systematic and Evolutionary Microbiology**, 53:1979–1983.

VON MERING, C.; HUYNEN, M.; JAEGGI D.; SCHMIDT S.; BORK P.; SNEL, B. (2003). STRING: a database of predicted functional associations between proteins. **Nucleic Acids Research**, 31:258-261.

VON MERING, C.; JENSEN, L.J.; KUHN, M.; CHAFFRON, S.; DOERKS, T.; KRUGER, B.; SNEL, B.; BORK, P. (2007). STRING 7—recent developments in the integration and prediction of protein interactions. **Nucleic Acids Research**, 35:358-362.

WRIGHT, F. (1990). The 'effective number of codons' used in a gene. **Gene**, 87:23-29.

WU, C.C.; GARDARIN A.; MARTEL A.; MINTZ E.; GUILLAIN F.; CATTY P. (2006). The cadmium transport sites of CadA, the Cd2+-ATPase from Listeria monocytogenes. **J Biol Chem**, 281:29533-29541.

8. APÊNDICES

8.1. Apêndice A – Programa BLAST (Basic Local Alignment Search Tool - ALTSCHUL et al.; 1997)

O programa BLAST realiza busca por similaridade entre seqüências que são submetidas e seqüências presentes num banco de dados. Para realizar essa busca, o programa BLAST funciona em três etapas:

1. “Semeadura” (seeding)

A primeira delas chama-se “semeadura/semeação” (*seeding*), na qual o programa fragmenta a seqüência submetida em todas as possíveis partes de 11 ou 3 caracteres (por padrão) se forem nucleotídeos (BLASTN) ou se forem aminoácidos (BLASTP), formando as chamadas “*word hits*”, ou simplesmente “palavras”. É então criada uma lista contendo todas as *word hits,* e em seguida, é verificada a ocorrência dessas *word hits* nas sequências presentes no banco de dados. Quando o BLAST encontra uma dessas “palavras” no banco de dados, diz-se que a seqüência foi “semeada”.

2. Extensão

Após “semear” as seqüências, inicia-se a etapa de extensão, onde o programa tenta ampliar o alinhamento das *word hits*. Nessa etapa, o programa BLAST usa um esquema de “*score*”. As seqüências semeadas então possuem um *score* inicial, e a cada novo caractere pareado é atribuído um valor de +1 (por exemplo) a esse *score*, da mesma forma que para cada diferença no pareamento atribui-se um valor de -1 ao *score*. Nesse caso, a extensão se encerraria quando o

score chegasse a zero. Essa estratégia serve para evitar o programa de encerrar a extensão logo nas primeiras discrepâncias de alinhamento.

3. Avaliação

A terceira etapa chama-se avaliação, e ocorre para verificar se os alinhamentos produzidos possuem significância estatística ou os caracteres alinharam ao acaso (por exemplo, existem quatro tipos de nucleotídeos numa seqüência de DNA, sendo então a chance de um deles ocorrer, numa determinada posição, 1 em 4) (GIBAS & JAMBECK, 2001; BEDELL et al., 2003).

8.2. Apêndice B - Scripts

8.2.1. Contigextract.sh

```
#!/bin/bash

# Extrai substrings (ex., ORF) de seqüências no formato FASTA (ex., contigs de uma montagem)
# Requer os scripts "seqextractor.sh" e "baseextract.sh"
# Dois arquivos devem ser fornecidos como argumento, contendo:
# arquivo 1) lista contendo: nome da ORF, nome do contig, início da ORF no contig e final da ORF no contig
#           deve conter uma ORF por linha e os valores tabulados;
# arquivo 2) seqüências, de onde as substrings serão retiradas, no formato FASTA
#
# Autores: Rodrigo Cardoso e Leonardo M. Cruz
# Data: 25 de setembro de 2006
#

ORFPOSITION="$1"     # Arquivo com os nomes das ORF e contigs e posições, inicial e final, das ORF
CONTIGSFASTA="$2"  # Arquivo de seqüências no formato FASTA
FASTAOUTFILE="seqfasta.out"            # Arquivo de saída

echo As seqüências serão escritas, no formato FASTA, no arquivo: $FASTAOUTFILE
touch $FASTAOUTFILE

IFS_old=$IFS    # Guarda o valor original da IFS
IFS=$'\n'
while read LINE# Lê o arquivo de nome e posições das ORF
        do

        IFS=" "
        COLUNAS=($(echo $LINE | tr "\t" " " | tr -s " "))  # Coloca as posições das ORF em um array
        ORFLENGTH=$((${COLUNAS[3]}-${COLUNAS[2]}+1))
        FLAG=

        echo Extraindo a seqüência do contig ${COLUNAS[1]}...
```

```
	CONTIGSEQ=$(bash seqextractor.sh ${COLUNAS[1]} $CONTIGSFASTA)	# Extrai a seqüência de um único contig


	echo Extraindo a seqüência da ORF ${COLUNAS[0]}...
	# Extrai uma substring correspondente à ORF e outras substrings com 200nt antes e depois da ORF, até
	# um máximo de 500nt
	# Altere estes valores da forma desejada
	for ((i=0; i<=500; i=$(($i+200))))
		do
		if [ ${COLUNAS[2]} -gt ${COLUNAS[3]} ]		# Se a ORF estiver invertida a extração será de forma diferente
			then
			bash baseextract.sh $CONTIGSEQ $((${COLUNAS[2]} + $i)) $(($ORFLENGTH + ($i * -2))) >> $FASTAOUTFILE
		else
			bash baseextract.sh $CONTIGSEQ $((${COLUNAS[2]} - $i)) $(($ORFLENGTH + ($i * 2))) >> $FASTAOUTFILE
		fi


		done


	done < $ORFPOSITION


IFS=$IFS_old


exit
```

8.2.2. Baseextract.sh

```
#!/bin/bash

# Extrai uma porção específica de uma seqüência no formato FASTA
#
# Autores: Rodrigo Cardoso e Leonardo M. Cruz
# Data: 07 de junho de 2006
# Atualizado em: 25 de setembro de 2006
#               Verifica se a ORF está invertida
#

SEQFILE="$1"  # Seqüência em FASTA
INIT="$2"         # Base inicial da substring
LENGTH="$3"  # Comprimento da substqring
END=$(( $INIT + $LENGTH - 1 ))          # Base final da substring

# Concatena a seqüência e coloca em $SEQBASE - a partir da variável
for LINE in `printf "$SEQFILE"`
        do

        if echo "$LINE" | grep ^\> > /dev/null
                then
                SEQTITLE=$LINE
        else
                SEQBASE=$SEQBASE$LINE
        fi

        done

# Cria substring da seqüência
if [ $LENGTH -lt 0 ]       # Se o comprimento da ORF for negativo (ORF invertida) o início e fim serão
invertidos
        then
        echo $SEQTITLE $INIT..$END
        LENGTH=$(($LENGTH * -1))
        printf "%s\n" "${SEQBASE:$END:$LENGTH}"

else
        echo $SEQTITLE $INIT..$END
        printf "%s\n" "${SEQBASE:$INIT:$LENGTH}"

fi

exit 0
```

8.2.3. Seqextractor.sh

```
#!/bin/bash

# Extrai uma seqüência especificada pelo usuário de um arquivo
# multi FASTA
#
# Autores: Rodrigo Cardoso e Leonardo M. Cruz
# Data: 25 de setembro de 2006
#

SEQ="$1"        # Nome da seqüência
FASTAFILE="$2"          # Arquivo que contém a seqüência

CONTIGNAME=$(grep $SEQ\$ $FASTAFILE)  # Encontra a seqüência no arquivo e
                                      # coloca na variável $CONTIG

sed -n "/$CONTIGNAME$/,/>/p" $FASTAFILE | sed -n '$!p'      # Imprime a seqüência (a última
                                                            # linha contém o título da próxima
                                                            # seqüência e é eliminada
                                                            # com o segundo "sed")

exit 0
```

8.2.4. Getsequences.sh

```
#!/bin/bash

# Extrai sequencias FASTA do banco de dados TransportDB
#
# Autores: Rodrigo Cardoso e Leonardo Cruz
#
# Data: 2 de março de 2007
# A variável múltipla "FAMILIAS" foi obtida através do código da página http://www.membranetransport.org/downloads/tree/faa

FAMILIAS=("APC" "P-ATPase" "MPT" "AAAP" "MIT" "Mid1" "ABC" "NCS2" "MFS" "NCS1" "MC" "GPH" "ZIP" "CytB" "Ctr2" "TDT" \
"PPI" "F-ATPase" "ClC" "MIP" "NSCC2" "RND" "OPT" "DMT" "Annexin" "Amt" "AE" "LCT" "SSS" "DASS" "CaCA" "MOP" \
"VIC" "Hsp70" "POT" "CDF" "CPA2" "CPA1" "ENT" "GUP" "SulP" "PiT" "ArsAB" "ThrE" "TRP-CC" "CHR" "MTC" "ACR3" \
"GPTS" " MPP" "CCC" "SSPTS" "CNT" "Nramp" "Trk" "NiCoT" "DAACS" "FP" "BASS" "IISP" "Oxa1" "FNT" "NSS" \
"Connexin" "PNaS" "GIC" "AEC" "RIR-CaC" "OAT" "PLM" "LIC" "Bcl-2" "RhtB" "MscL" "OFeT" "MgtE" "MscS" \
"H+-PPase" "ICC" "RFC" "CD20" "FBT" "CPA3" "O-ClC" "ENaC" "LysE" "OST" "KUP" "PCC" "FeT" "TRAP-T" "IRK-C" \
"UT" "Bestrophin" "ICln" "ACC" "Tat" "MET" "LPI" "BenE" "PUP" "NhaD" "NhaA" "MerTP" "E-ClC" "ArsB" "LctP" \
"FeoB" "GntP" "ESS" "AGCS" "BCCT" "PnuC" "CitMHS" "MSS" "LIVCS" "LIV-E" "SBT" "CadD" "ICT" "NhaC" "Tic110" \
"AAA" "HAAAP" "MEX" "Dcu" "CCS" "TTT" "UAC" "AAE" "KDGT" "AbgT" "DcuC" "PbrT" "Mtt" "Innexin" "NhaB" "Ctr3" "Ctr1")

for ((i=0; i<${#FAMILIAS[*]}; i++))
        do
        echo Copiando Familia ${FAMILIAS[$i]}
        wget http://www.membranetransport.org/downloads/tree/faa/${FAMILIAS[$i]}.faa
        done

cat *faa > transporterDB

exit 0
```

8.2.5. Separa_ORF

```
#!/bin/bash

# Extrai ORF de um arquivo fasta contendo diversas ORF e separa-as em arquivos
#
# Autores: Rodrigo Cardoso e Leonardo M. Cruz
# Data: 30 de março de 2007
#
IFS_old="$IFS"
IFS=$'\n'

while read LINE
        do

        if echo $LINE | grep ^\> > /dev/null
                then
                TITLE=$(echo $LINE | cut -d" " -f1 | cut -d"|" -f2)
        elif echo $LINE | grep ^[ABCDEFGHIJKLMNOPQRSTUVXZWY] > /dev/null
                then
                SEQ=$LINE
                echo ">$TITLE" >> /home/rodrigo/monografia/orfs_proteins/$TITLE.txt
                echo $SEQ >> /home/rodrigo/monografia/orfs_proteins/$TITLE.txt
                echo
        else
                continue
        fi
        done < /home/rodrigo/monografia/map_herbas/arquivos/orfsptn.fasta
exit 0
```

8.2.6. Blastparser.pl

```
#!/usr/bin/perl


# Extrai algumas características dos alinhamentos produzidos pelo programa BLAST, e calcula alguns parâmetros usando esses valores, colocando tudo num arquivo tabelado tipo CSV
#
# Autores: Vinicius Weiss e Rodrigo Cardoso
# Data: 21 de março de 2007
#
# Correr script: nome do script, seguido do nome do arquivo de saída do blast, seguido de um sinal ">" seguido de um arquivo de saída
use strict;
use Bio::SearchIO;
use Bio::SeqFeature::Generic;
use Bio::Search::Hit::BlastHit;
use Bio::Search::HSP::GenericHSP;



my $file = shift or die "Usage: render_blast4.pl <blast file>\n";



 my $searchio = new Bio::SearchIO(-format => 'blast',
                                  -file   => $file);

     my $query ;
     my $query_length;
     my $subject ;
     my $subject_length;
     my $score_bits ;
     my $score ;
     my $expect;
     my $identities ;
     my $positives ;
     my $query_start ;
           my $query_end;
           my $subject_start;
```

```
        my $subject_end;
    my $frame;
    my $gaps;
    my $length_aln;
    my $identities_query;
    my $identities_subject;
    my $positives_query;
    my $positives_subject;
    my $proportion;
    my $query_aa;
      my $align_lenght_query;
      my $align_lenght_subject;



                print "QUERY;SUBJECT;QUERY LENGTH;QUERY AA;SUBJECT LENGHT;ALIGN
LENGHT;IDENTITIES;POSITIVES;QUERY START;QUERY END;SUBJECT START;SUBJECT
END;SCORE;EXPECT;GAPS;ALIGN LENGHT QUERY;ALIGN LENGHT SUBJECT;IDENTITY
QUERY;IDENTITY SUBJECT;POSITIVES QUERY;POSITIVES SUBJECT;PROPORTION","\n";


        while( my $result = $searchio->next_result )
  {
    while( my $hit = $result->next_hit )
    {


      while( my $hsp = $hit->next_hsp )
       {
          $query = $hsp->seq_id;
          #$query = substr($query , 27);
          $subject = $hit->name;
          $score =  $hsp-> score; #score
          $expect =  $hsp->significance(); #  expect
          $identities =  $hsp->num_identical(); #  Identities
          $positives =  $hsp->num_conserved(); #  Positives
          $query_start = $hsp-> start('query'); #  inicio query
          $query_end = $hsp -> end;('query'); #  final query
                                $subject_start = $hsp -> start('hit'); # inicio subject
                                $subject_end = $hsp -> end('hit'); # final subject
          $query_length = $hit-> query_length(); #tamanho da query
```

```
            $subject_length = $hit-> length(); #tamanho subject
            $gaps = $hsp->gaps('query'); # gaps
            $length_aln = $hsp->length(); #alinhamento
                                        $frame = $hsp ->frame(); #frame
                    $align_lenght_subject = $length_aln / $subject_length;
                    $query_aa = ($query_length) /3;
                    $align_lenght_query = $length_aln / $query_aa;
            $identities_query = $identities / $query_aa;
            $identities_subject = $identities / $subject_length;
            $positives_query = $positives / $query_aa;
            $positives_subject = $positives / $subject_length;
            $proportion = $query_aa / $subject_length;


          print
"$query;$subject;$query_length;$query_aa;$subject_length;$length_aln;$identities;$positives;$query_
start;$query_end;$subject_start;$subject_end;$score;$expect;$gaps;$align_lenght_query;$align_leng
ht_subject;$identities_query;$identities_subject;$positives_query;$positives_subject;$proportion","\n";


                    }
    }
  }
```

8.2.7. Comparaorf.sh

```
#!/bin/bash

# Verifica se há repetições de ORF em arquivos CSV gerados pelo script Blastparser.pl
#
# Autores: Rodrigo Cardoso e Leonardo Cruz
# Data: 26 de julho de 2007

while read LINE
	do
	ORF2=$(echo $LINE | cut -d";" -f1)

	if [ "$ORF1" == "$ORF2" ]
		then
		continue
	else
		echo $LINE >> 1220_2.csv
		ORF1=$(echo $LINE | cut -d";" -f1)
	fi

	done < "$1"

exit 0
```