UNIVERSIDADE FEDERAL DO PARANÁ

VINICIUS ALMIR WEISS

ESTRATÉGIAS DE FINALIZAÇÃO DA MONTAGEM DO GENOMA DA BACTÉRIA DIAZOTRÓFICA ENDOFÍTICA *Herbaspirillum seropedicae* SmR1

CURITIBA
2010

VINICIUS ALMIR WEISS

ESTRATÉGIAS DE FINALIZAÇÃO DA MONTAGEM DO GENOMA DA BACTÉRIA DIAZOTRÓFICA ENDOFÍTICA *Herbaspirillum seropedicae* SmR1

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciências – Bioquímica, do Setor de Ciências Biológicas da Universidade Federal do Paraná, como requisito parcial para a obtenção do título de Mestre em Ciências – Bioquímica.

Orientador:
Profº Dr. Leonardo Magalhães Cruz
Co-orientador:
Profº Dr. Roberto Tadeu Raittz

CURITIBA
2010

TERMO DE APROVAÇÃO

VINÍCIUS ALMIR WEISS

Estratégias de finalização da montagem do genoma da bactéria diazotrófica endofítica *Herbaspirillum seropedicae* SmR1

Dissertação aprovada como requisito parcial a obtenção do título de Mestre em Ciências – Bioquímica, no Programa de Pós-Graduação em Ciências – Bioquímica, Setor de Ciências Biológicas da Universidade Federal do Paraná, pela banca examinadora formada pelos professores

Prof. Dr. Roberto Tadeu Raittz (Co-orientador)
Setor de Educação Profissional e Tecnológica
UFPR

Prof. Dr. Fábio de Oliveira Pedrosa
Departamento de Bioquímica e Biologia Molecular
UFPR

Prof. Dr. Emanuel Maltempi de Souza
Departamento de Bioquímica e Biologia Molecular
UFPR

Profª Drª Jeroniza Nunes Marchaukoski
Setor de Educação Profissional e Tecnológica
UFPR

Curitiba, 19 de Fevereiro de 2010

A Deus.

Aos meus pais, Almir e Simone e a minha irmã Pâmela.

AGRADECIMENTOS

Ao Professor Fábio de Oliveira Pedrosa, pela oportunidade de trabalhar no Núcleo de Fixação de Nitrogênio, pela ajuda e a atenção. Muito obrigado.

Aos meus orientadores, o Professor Leonardo Magalhães Cruz e Professor Roberto Raittz que foram responsáveis pelo meu ingresso no campo científico, sempre com apoio e paciência. Pelo auxílio, explicações e sugestões, na elaboração deste trabalho. Muito obrigado.

Ao Professor Emanuel Maltempi de Souza, pelas explicações, sugestões e atenção. Muito obrigado.

A Professora Maria Berenice Reynaud Steffens, pela ajuda desde minha preparação para o Mestrado. Muito obrigado.

Aos Professores do programa, pelo conhecimento adquirido e pela paciência. Muito obrigado.

Ao Grupo de Fixação Biológica de Nitrogênio da Universidade Federal do Paraná pelos seqüenciamentos.

Aos meus amigos Marco Antonio Seiki Kadowaki, Giovani Pisa e Felipe Renó Oliveira Pisa, pela convivência e ajuda. Muito Obrigado.

A todos os colegas de departamento. Muito obrigado.

Aos meus pais Simone e Almir Weiss, pela dedicação e amor, por nunca terem medido esforços para que eu seguisse sempre em frente. Por tudo que eles são e representam na minha vida. Muito obrigado.

A minha irmã Pâmela, por ser minha parceira para todas as horas. Muito obrigado.

A Deus por tudo.

RESUMO

Herbaspirillum seropedicae SmR1 e *Herbaspirillum rubrisubalbicans* M1 são bactérias diazotróficas endofíticas associadas a diversos grupos de plantas, principalmente gramíneas de importância agrícola. Neste trabalho foram utilizadas diferentes estratégias visando o fechamento do genoma do *H. seropedicae* estirpe SmR1. O genoma completo do *H*. *seropedicae* SmR1 finalizado contem 5.513.887 pb e 63,4%G+C . Foram identificados 4.737 genes compreendendo 88% do genoma, 55 tRNAS e três operons de RNA ribossomais. Os padrões de fragmentação in silico do DNA genômico do *H. seropedicae* em sítios de restrição específicos concordaram com análises por eletroforese em campo pulsado (PFGE), apoiando a ordenação da sequência genômica obtida. A sequência parcial do genoma do *H*. *rubrisubalbicans* analisado apresenta 3.291.242 pb, e um conteúdo GC de 61,3%, distribuídos em 2.703 sequências contíguas. A anotação deste genoma foi realizada com base em análises comparativas, utilizando o genoma de *H*. *seropedicae* como referência. Foram identificados 1.569 genes, 10 tRNAS e 1 operon de RNA ribossomal. Os produtos de tradução dos genes de ambos os genomas foram classificados em categorias funcionais de acordo com o banco de dados COG. A análise de códon mostrou uma forte tendência no uso de códons sinônimos contendo G e C na terceira base, refletindo o ao alto conteúdo de GC no genoma de *H*. *seropedicae*. Finalmente, os aminoácidos glicina, alanina e valina apresentam as maiores freqüências nas proteínas codificadas pelo genoma de H. seropedicae.

Palavras chave: sequenciamento de DNA, genoma, *Herbaspirillum seropedicae,* anotação, análise de códons,

ABSTRACT

Herbaspirillum seropedicae SmR1 and *Herbaspirillum rubrisubalbicans* M1 are endophitic diazotrophs associated with important agricultural crops. In the present work, different strategies to finish the genome of *H. seropedicae* were employed. The finished genome of *H. seropedicae* SmR1 contains 5,513,887 bp and 63.4% G+C. A total of 4,737 genes covering 88% of the genome, 55 tRNAs, and three ribosomal RNA operons were identified. The pattern of DNA *In silico* fragmentation at specific restriction sites of the *H. seropedicae* genome agreed well with Pulsed Field Gel Electrophoresis (PFGE) data, supporting the final ordering of the genomic sequence. The partial sequence of *H. rubrisubalbicans* genome analysed contains 3,291,242 bp and 61.3% of G+C distributed in 2,703 contigs. The annotation of the *H. rubrisubalbicans* genome was done by comparative analysis using the annotated genome of *H. seropedicae* as reference. In this genome 1,569 genes*,* 10 tRNAs *e* 1 ribossomal RNA operon were identified. The general functions of the translation products of both genomes were classified according to the COG database functional categories*.* The codon analysis showed a strong bias towards the use of G and C in the third base, reflecting the high GC genome content of *H. seropedicae.* Glycine, alanine, and valine showed the highest frequencies among the amino acids in the coded proteins.

Key-words: DNA sequencing, genome, *Herbaspirillum seropedicae,* annotation, codon analysis,

LISTA DE FIGURAS

FIGURA 1	Complexidade do processamento dos dados genômicos	13
FIGURA 2	Gráfico de crescimento do repositório público GenBank	15
FIGURA 3	Lista de países com maior quantidade de projetos genoma	16
FIGURA 4	Ciclo de sequenciamento pelo método de Sanger	19
FIGURA 5	Pirossequenciamento	21
FIGURA 6	Eletroforetograma	33
FIGURA 7	Regiões de alta e baixa qualidade	35
FIGURA 8	Distribuição das leituras de sequências pelo método Sanger	41
FIGURA 9	Mapa geral do genoma de *H. seropedicae*	53
FIGURA 10	Distribuição das ORF anotadas de *H. seropedicae* nas categorias funcionais COG	54
FIGURA 11	Distribuição do uso de códons no genoma de *H. seropedicae*	57
FIGURA 12	Frequência do uso de aminoácidos em proteínas codificadas pelo genoma de *H. seropedicae*	58
FIGURA 13	Distribuição das ORFs identificadas no genoma parcial de *H. rubrisubalbicans* nas categorias funcionais COG	61
FIGURA 14	Sobreposição do genoma parcial de *H. rubrisubalbicans* no genoma de *H. seropedicae*	62

LISTA DE TABELAS

TABELA 1	Resumo das características dos sequenciadores atuais	17
TABELA 2	Programas de Anotação	23
TABELA 3	Instituições participantes do Projeto GENOPAR	28
TABELA 4	Resultado da análise das leituras de sequências	42
TABELA 5	Estatísticas da MontagemV1	43
TABELA 6	Estatísticas da MontagemV2	44
TABELA 7	Estatísticas da MontagemV3	46
TABELA 8	Estatísticas da MontagemV4	47
TABELA 9	Estatísticas da MontagemV5	49
TABELA 10	Comparação do padrão de fragmentação *in silico* com enzima de restrição SwaI e por PFGE do genoma de *H. seropedicae*	50
TABELA 11	Características estruturais do genoma de *H. seropedicae*	52
TABELA 12	tRNAS presentes no genoma de H. *seropedicae* e seus anticódons	55
TABELA 13	Preferência no uso de Códon para o genoma de *H. seropedicae*	56
TABELA 14	Estatísticas para a montagem parcial da sequência genômica de *H. rubrisubalbicans* (MontagemHR)	59
TABELA 15	Características do genoma parcial de *H. rubrisubalbicans*	60

LISTA DE ABREVIATURAS

ATP	Adeninosina trifosfato
dNTP	Desoxirribonuleotídeo trifosfatado
DNA	Ácido desoxirribonucleico
EMBL	Laboratório de Biologia Molecular Europeu
FASTA	Formato utilizado para armazenar sequências de bases e de aminoácidos em arquivo texto
GENBANK	Banco de dados público do National Center for Biological Information, dos Institutos de Saúde dos Estados Unidos da America.
HGT	Elementos de transferência genética horizontal
Kb	Kilo bases ou mil pares de bases
lacZ	Gene repórter que codifica para a enzima beta galactosidase
MGE	Elementos genéticos móveis
Mpb	Mega pares de bases ou um milhão de pares de bases
ORF	Sequência de leitura aberta
pb	Pares de bases
PCR	Reação em cadeia da polimerase
PPi	Pirofosfato inorgânico
RNA	Ácido ribonucleico
tRNA	Ácido ribonucleico transportador

SUMÁRIO

1 INTRODUÇÃO 12
1.1 BIOINFORMÁTICA 12
1.2 SEQUENCIAMENTO DE DNA E SEQUENCIAMENTO GENÔMICO 14
1.2.1 Sequenciamento pelo Método Sanger 17
1.2.2 Sequenciamento pelo Método de Margulies (Pirossequenciamento) 20
1.3 PROGRAMAS PARA MONTAGEM DE CONTIGS 21
1.4 ANOTAÇÃO DO GENOMA 22
1.5 COMPARAÇÃO DE GENOMAS 23
1.6 *Herbaspirillum sp*. 24
2 OBJETIVOS 26
2.1 OBJETIVO GERAL 26
2.2 OBJETIVOS ESPECÍFICOS 26
3 MATERIAL E MÉTODOS 27
3.1 SEQUÊNCIAS DO PROGRAMA GENOPAR 27
3.2 SISTEMA OPERACIONAL 29
3.3 LINGUAGENS DE PROGRAMAÇÃO 29
3.3.1 Perl 29
3.3.2 Bash 30
3.3.3 Sql 30
3.3.4 Mysql 30
3.4 PROGRAMAS DE BIOINFORMÁTICA 31
3.4.1 Artemis 31
3.4.2 BLAST 31
3.4.3 ClustalW 31
3.4.4 Newbler 31
3.4.5 Phred/Phrap/Consed 32
3.4.6 SSAHA 34
3.5 MÉTODOS USADOS PARA A FINALIZAÇÃO DA MONTAGEM 35
3.5.1 Poda de bases com baixa qualidade 35
3.5.2 Montagem 35
3.5.3 Verificação de inconsistências 36
3.5.4 Análise das extremidades dos contigs 37
3.5.5 Inserção das leituras de sequências na montagem 37
3.5.6 Ressequenciamento dos clones e construção de *primers* 37
3.5.7 União entre contigs 38
3.6 ANÁLISE DE PREFERENCIA DE USO DE CÓDONS 38
4. HARDWARE 39
4.1 Desktop 39
4.2 Servidores 39
5. RESULTADOS E DISCUSSÃO 40
5.1 ANÁLISE DAS SEQUÊNCIAS USADAS NA MONTAGEM 40
5.2 GENOMA DE *H. seropedicae* 42
5.2.1 MontagemV1 42
5.2.2 MontagemV2 43
5.2.3 MontagemV3 45
5.2.4 MontagemV4 46
5.2.5 MontagemV5 47

5.2.6 Validação do Genoma de *H. seropedicae* 49
5.2.7 Anotação do genoma de *H.seropedicae* 51
5.2.8 Análise de uso de códon no genoma de *H. seropedicae* 55
5.3 GENOMA *DE H. rubrisubalbicans* 58
5.3.1 Anotação e análise de códons do genoma de *H. rublisubalbicans* 59
6 CONCLUSÕES 63
REFERÊNCIAS 64
ANEXOS 70

1 INTRODUÇÃO

1.1 BIOINFORMÁTICA

Dados biológicos, especialmente os de sequenciamento genômico, vêm sendo gerados em ritmo acelerado nas últimas décadas como resultado da automação dos equipamentos, o desenvolvimento de novas metodologias de alto rendimento e baixo custo. E Isto permitiu o desenvolvimento de centenas de projetos de sequenciamento de genomas, produzindo grande volume de informação. A fim de analisar este grande volume de dados os computadores tornaram-se instrumentos fundamentais para a pesquisa biológica, favorecendo o surgimento de um novo segmento no campo da ciência, a Bioinformática (LUSCOMBE *et al*., 2001).

A Bioinformática pode ser definida como a aplicação de técnicas computacionais para manipular dados biológicos (HUGHEY *et al*., 2001), ou também como Biologia computacional, aplicando técnicas quantitativas e analíticas à modelação de sistemas biológicos (GIBAS *et al.*, 2001).

A Bioinformática é interdisciplinar sendo que seus domínios permeiam vários campos do conhecimento como Biologia, Medicina, Matemática, Física, Ciência da Computação e Estatística. As Universidades já oferecem cursos com o objetivo de preparar profissionais qualificados para atuar como bioinformatas visando suprir a necessidade dos grandes projetos de pesquisa e empresas que trabalham com estas informações e sua alta complexidade (BAYAT, 2002).

Os profissionais de bioinformática podem atuar em várias áreas da Biologia Molecular, onde a análise e interpretação dos dados biológicos não se restringem apenas à Genômica, mas também à Proteômica e Transcriptomica, aumentando assim a complexidade da análise (Figura 1), destacando-se as seguintes áreas fundamentais:

- Análise de sequências de DNA – determinar genes que codificam proteínas específicas e áreas regulatórias.
- Análise de expressão genética – determinar o nível de expressão de um determinado gene (microarrays, SOLID, 454).
- Análise de expressão proteica – análise da expressão de proteínas em uma determinada condição através de identificação de pontos em géis de eletroforese 2D.

- Análise de regulação da expressão gênica – detectar, identificar e predizer regiões de regulação da expressão de grupos de genes.
- Predição de estrutura protêica – resolver a estrutura das proteínas através de cristalografia e modelagem computacional.
- Genômica comparativa – estabelecer relações entre os genomas de organismos próximos evolutivamente visando identificar suas particularidades, análise de polimorfismos e filogenia.

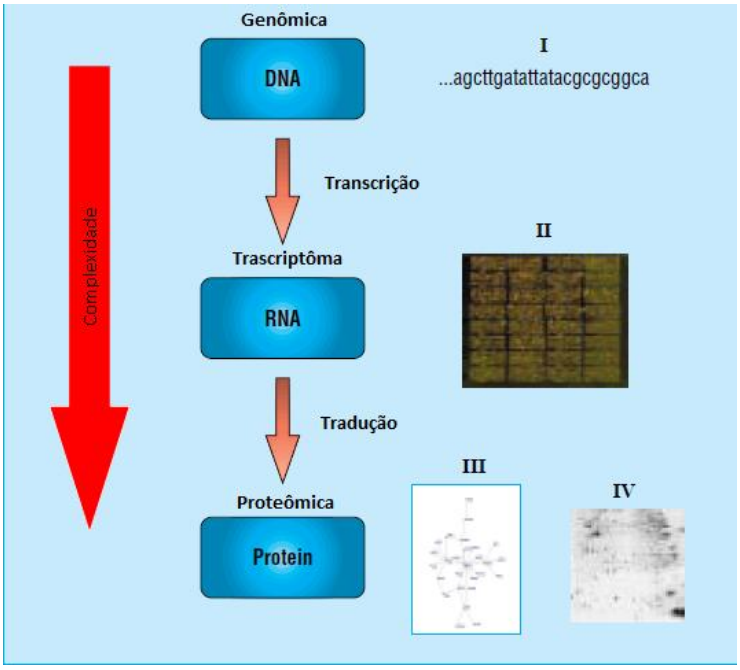


FIGURA 1. Complexidade do processamento dos dados genômicos.

Bayat 2002: Análises e interpretação dos dados biológicos consideram informação em todos os níveis do genoma (conteúdo total genético), do proteoma (conteúdo total de proteínas) e transcriptoma (conteúdo total de mensageiros RNA). As imagens numeradas apresentam exemplos de dados biológicos, onde I representa DNA (nucleotídeos), II RNA (microarray mostrando níveis de expressão gênica), III e IV Proteínas (onde III mostra a estrutura de uma única proteína e IV um gel de eletroforese 2D mostrando a separação de várias proteínas. Cada ponto representa cadeias de proteínas diferentes).

Assim segundo Luscombe *et al*., (2001), a Bioinformática tem como objetivos principais na manipulação da informação biológica preparar, organizar e disponibilizar a informação para estudo, facilitando a manipulação e edição de dados pelos pesquisadores através da criação de bancos de dados e redes de colaboração via Internet. Desenvolver ferramentas e recursos que resolvam os problemas e

facilitem a análise dos dados automatizando processos e aumentando a agilidade na obtenção dos resultados, proporcionando um tempo maior para os pesquisadores se concentrarem na análise. Utilizar a capacidade destas ferramentas para conduzir análises globais de todos os dados disponíveis, visando descobrir princípios comuns que se aplicam em outros sistemas.

1.2 SEQUENCIAMENTO DE DNA E SEQUENCIAMENTO GENÔMICO

O sequenciamento genômico permite a identificação e caracterização de genes e das proteínas por eles codificadas, permitindo um melhor entendimento dos processos biológicos. O estabelecimento de relações evolutivas entre estes organismos também é possível, bem como a identificação de elementos móveis, de genes adquiridos por transferência lateral e mapas metabólicos teóricos (NIERMAN *et al*., 2000).

O sequenciamento de genomas é relativamente recente na Biologia Molecular e Microbiologia, sendo que o primeiro genoma de um organismo não viral a ser sequenciado foi o da bactéria *Haemophilus influenzae,* em 1995 (FLEISCHMANN *et al*., 1995), com 1,8Mpb. Após a publicação deste trabalho, genomas de outras bactérias foram sequenciadas em um curto espaço de tempo como *Mycobacterium tuberculosis* (COLE *et al*., 1998), um dos patógenos humanos mais importantes, a *Escherichia coli* (BLATTNER *et al*., 1997) e a primeira *Archaea*, *Archaeoglobus fulgidus* (KLENK *et al*., 1997). Também foram sequenciados genomas de eucariotos como o do parasita responsável pela Malária, *Plasmodium falciparum* (GARNER *et al*., 2002a; GARNER *et al*., 2002b; HALL *et al*., 2002; HYMAN *et al.*, 2002). Estes projetos junto com o sequenciamento do genoma de mamíferos como o genoma humano (LANDER *et al*., 2001), do rato (WATERSON et al., 2002) e do chimpanzé (MIKKELSEN *et al* 2005) foram determinantes para a obtenção massiva dos dados disponíveis até hoje (HALL, 2007). Atualmente mais de 1000 genomas bacterianos completos estão depositados no GenBank, o banco de dados público do National Center for Biological Information, dos Institutos de Saúde dos Estados Unidos da America. Bancos de dados similares e espelhos encontram-se na Europa e no Japão.

A determinação da sequência de genomas só foi possível com a introdução

de métodos de sequenciamento automáticos. O primeiro sequenciador automático utilizava a metodologia de Sanger (SANGER *et al*., 1977) modificada por (EDWARDS *et al*., 1990) e esta metodologia liderou a pesquisa genômica por 30 anos. Esta realidade mudou com o surgimento da nova geração de sequenciadores a partir de 2005, com a publicação do método de pirossequenciamento (MARGULIES *et al*., 2005) utilizado o sequenciador 454 (Roche). Outros métodos foram desenvolvidos a partir deste período, os mais expressivos são o método Polony (SHENDURE *et al*., 2005) utilizado no sequenciador SOLID (AppliedBiosystems) e o método de amplificação em ponte (BENNETT *et al*., 2005) utilizado no sequenciador Genome Analyser (Ilumina). Estes métodos promoveram um aumento exponencial na quantidade de dados gerados, dobrando a quantidade de bases submetidas em bancos públicos de sequencias como NCBI (http://www.ncbi.nlm.nih.gov/) a cada 18 meses. O NCBI conta atualmente com mais de 350.000 submissões resultando em um total de mais de 100 trilhões de bases depositadas (Figura 2).

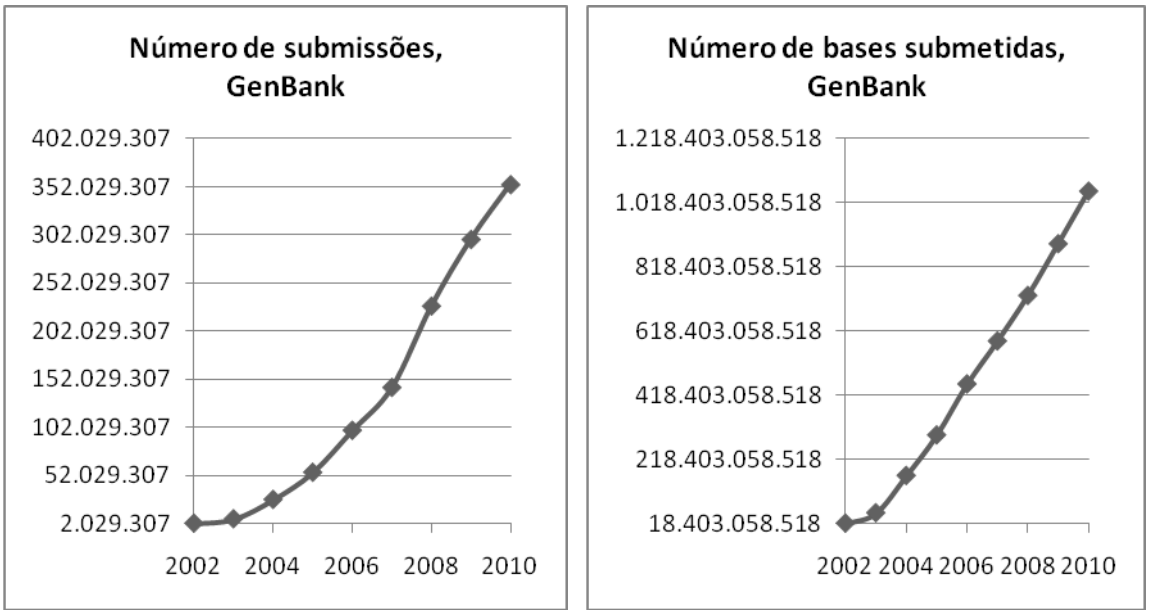


FIGURA 2. Gráfico de crescimento do repositório público GenBank.
Fonte: http://www.ncbi.nlm.nih.gov

Outro fator determinante para o aquecimento da área foi a redução no custo do sequenciamento (MARDIS *et al*., 2007) através da introdução destas novas tecnologias, possibilitando assim que laboratórios menores liderassem seus projetos

genomas. Atualmente o Brasil encontra-se na nona posição com 37 projetos, em relação aos países com os maiores números de projetos na área (Figura 3).

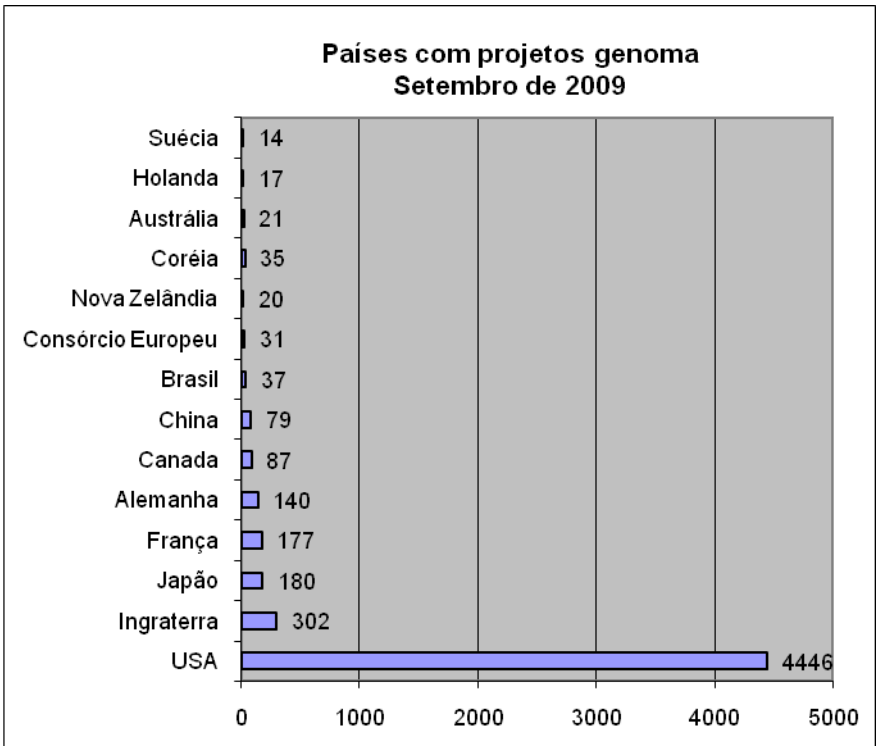


FIGURA 3. Lista de países com maior quantidade de projetos genoma.
Fonte: http://genomesonline.org/gold_statistics.htm

Na Tabela 1 encontra-se um resumo dos sequenciadores mais utilizados atualmente e a redução do custo do sequenciamento com o uso destas novas tecnologias.

TABELA 1. RESUMO DAS CARACTERÍSTICAS DOS SEQUENCIADORES ATUAIS

	Sanger	454	Solid	Illumina
Química do Sequenciamento	Terminadores de cadeia marcados com fluoróforos	Pirosequenciamento	Sequenciamento por ligação	Sequenciamento por síntese com base em polimerase
Abordagem de amplificação	Clonagem in vivo	PCR em emulsão	PCR em emulsão	Amplificação em ponte
Tamanho das leituras	~650 pb	~200-500 pb	~25-35 pb	35-100 pb
Número de leituras por corrida	384	200.000-400.000	200 milhões sequencias de pares	8X 10.000.000
Números de dados gerados por corrida	290 kb	~300 Mb	> 15Gb	~7 Gb
Tempo por corrida	1 h	5 hrs	10 dias	3-7 dias
Sequências de Pares	sim	sim	sim	sim
Custo por corrida	£ 192	£ 3500	£ 7000	£ 6000
Custo por bases	£ 1500	£ 87,50	£ 0.90	£ 1.70
Acurácia do sequenciamento	>99,9	99,5	99,94	98,5

Fonte: Thomas D. Otto, Patogen Genomics.

1.2.1 Sequenciamento pelo Método Sanger

Método de sequenciamento enzimático onde a marcação foi inicialmente realizada com os isótopos radioativos ^{32}P ou ^{35}S e a cadeia interrompida com dideoxinucleotídeos terminadores. Os fragmentos sintetizados eram separados por eletroforese em gel desnaturante de poliacrilamida-uréia e a seguir autoradiografados. Devido a este tipo de gel ter alto poder de resolução, foi possível a separação de fragmentos de DNA com a diferença de uma única base. Desta forma, a técnica permitia analisar originalmente de 200 a 300 nucleotídeos por gel (SANGER *et al*., 1977). Posteriormente a marcação com isótopos radioativos foi substituída por fluoróforos (SMITH *et al.*, 1986). Quatro diferentes fluoróforos foram empregados e uma vez excitados por um feixe de laser emitem luz em diferentes comprimentos de onda. Foi possível marcar com estes fluoróforos o oligonucleotídeo iniciador usado na reação de sequenciamento ou então cada um dos dideoxinucleotídeos terminadores. Assim, uma vez que em cada uma das reações de incorporação de base (A,C,T,G) for empregado um fluoróforo diferente é possível

juntar estes produtos e realizar a corrida em uma única raia do gel de sequenciamento. Os produtos da reação de sequenciamento marcados, ao serem submetidos à eletroforese passam pelo feixe de laser, que promove a excitação dos fluoróforos (PROBER *et al.*, 1987). A luz emitida pelos fluoróforos é detectada por um fotomultiplicador e a informação é processada através do sequenciador como, por exemplo, o ABI da empresa Aplied Biosystems e MegaBACE da empresa GE Healthcare (Figura 4). O tamanho médio das sequências aumentou de 450 pb para 850 pb.

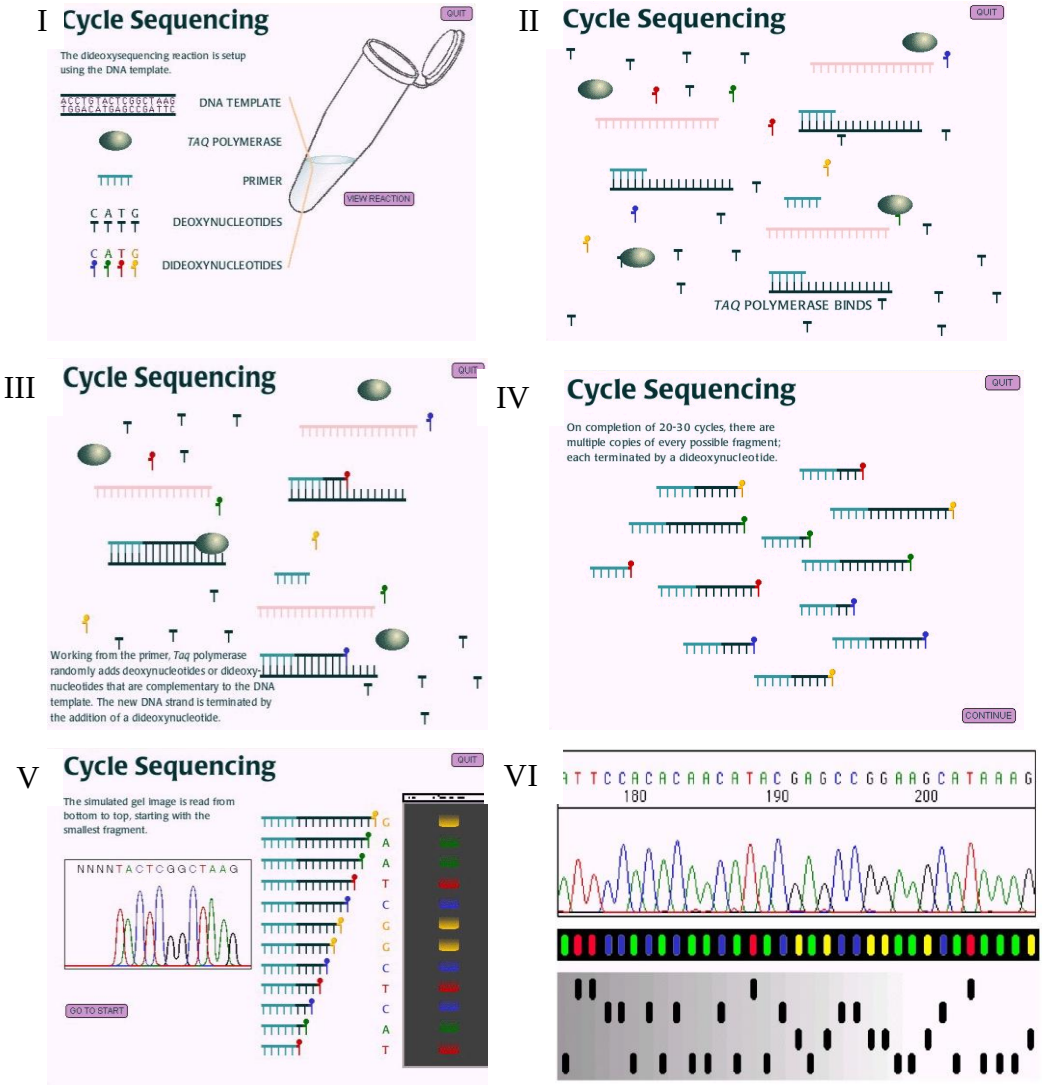


FIGURA 4. Ciclo de sequenciamento pelo método de Sanger.

I – Reação de sequenciamento. É necessária a presença de deoxinucleotídeos, dideoxinucleotídeos terminadores marcados com fluoróforos, primers, TAQ polimerase e o DNA genômico.
II – Anelamento dos primers no DNA genônico e ligação da TAQ polimerase polimerizando aleatoriamente um deoxinucleotídeo ou dideoxinucelotídeo terminador de cadeia.
III – TAQ insere um dideoxinucleotideo terminador de cadeia marcado com um fluoróforo que não contem 3´OH interrompendo a polimerização.
IV – Criação de fragmentos de DNA clonados de tamanho aleatório.
V – Separação dos fragmentos em gel de poliacrilamida. A leitura das bases é realizada através do dideoxinucleotídeo terminador marcado com fluoróforo.
VI – Os diferentes fluoróforos são excitados emitindo comprimentos de onda em frequências diferentes captadas por um fotomultiplicador.

Uma das estratégias de sequenciamento se baseia na clonagem massal de fragmentos de DNA seguido de sequenciamento. Como a clonagem é totalmente ao acaso, este método é denominado em inglês de “shotgun” por analogia aos disparos

quase aleatórios de um rifle. O método consiste na fragmentação do DNA genômico por quebra mecânica (sonificação ou nebulização) e a inserção de fragmentos em um vetor de clonagem. Os vetores contendo genes de resistência, genes reporters e o inserto são transformados em *E.coli* para a clonagem *in vivo* construindo as bibliotecas de clones. Após a clonagem é feita a seleção dos clones transformados através da resistência a ampicilina e da expressão do gene repórter lacZ que permite a seleção das colônias transformadas que contém o incerto através da diferença de coloração.

1.2.2 Sequenciamento pelo Método de Margulies (Pirossequenciamento)

A tecnologia do sequenciador 454 (Roche) dispensa a necessidade de clonagem, tirando vantagem de um eficiente método de amplificação *in vitro* de DNA, conhecido como PCR em emulsão (MARGULIES *et al*., 2005). Segundo Morozova *et al.*, 2008, na PCR em emulsão fragmentos individuais de DNA obtidos por nebulização são ligados a adaptadores marcados com biotina, e capturados individualmente em nanoesferas cobertas superficialmente por estreptavidina em emulsão. Estas nanoesferas agem como reatores de amplificação individual produzindo cópias de um único molde. Cada molde é subsequentemente transferido para um poço individual presente na lâmina de picoreação, e um clone molde relacionado é analisado usando reação de pirossequenciamento. O uso desta lâmina permite milhões de reações de pirossequenciamento realizadas em paralelo, aumentando massivamente o sequenciamento. A abordagem de pirossequneciamento é uma técnica de “sequenciamento por síntese” que mede a liberação de pirofosfato inorgânico (PPi) por quimioluminescência. O DNA molde é imobilizado e soluções de dNTPs são adicionadas uma de cada vez. A liberação de PPi com a incorporação do nucleotídeo complementar é convertido a ATP pela enzima ATP sulfurilase que fornece energia para a oxidação da luciferina e consequentemente geração de luz. As sequências do DNA molde são determinadas por um “pirograma”, que corresponde à ordem correta de nucleotídeos que é incorporado e a intensidade do sinal de quimioluminescência é proporcional ao total de moléculas de pirofosfato liberadas (Figura 5).

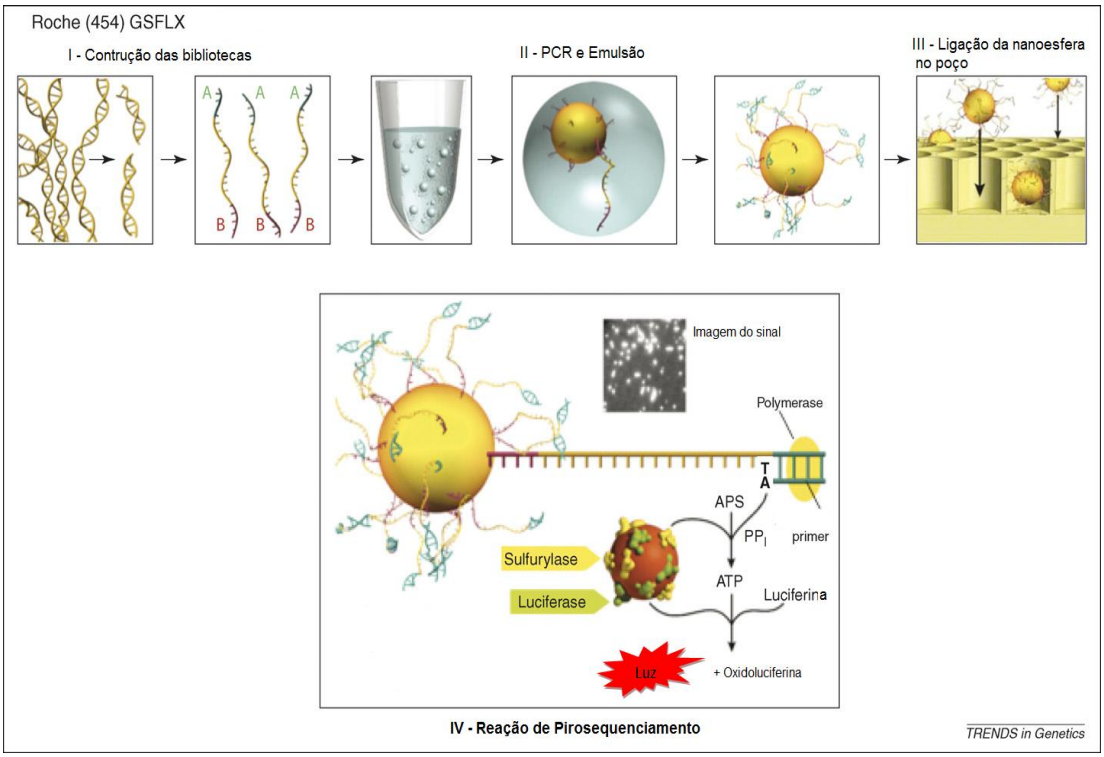


FIGURA 5. Pirossequenciamento.

Fonte: Mardis, Trends in Genetics, v. 24, n.3, 2007.

I Construção das bibliotecas – Genoma é fragmentado por nebulização, não existe a necessidade de selecionar colônias. O DNA simples fita é criado com adaptadores A e B que são usados como primers para ligação nas nanoesferas. A seleção dos fragmentos A/B é feita através de purificação por streptavidina-biotina

II PCR em Emulsão – O DNA simples fita é capturado pelas nanoesferas. A emulsão contendo os reagentes de PCR formam microreatores para a clonagem do fragmento.

III Ligação da nanoesfera no poço – Cada poço recebe uma nanoesfera com os fragmentos de DNA simples fita clonados. O diâmetro de cada poço possue 44 µm contendo 200.000 leituras em paralelo.

IV Reação de Pirosequenciamento – Anelamento da polimeraze no primer, incorporação dos dNTPs e liberação do PPi que é convertido a ATP pela sulfurilase. O ATP fornece energia a luciferase oxidar a luciferina em oxiluciferina gerando luz. Por fim a apirase hidrolisa os dNTPs excedentes limpando o poço para um nova incorporação.

1.3 PROGRAMAS PARA MONTAGEM DE CONTIGS

Devido ao grande volume de informação gerado e a limitação da técnica de sequenciamento, na qual é possível sequenciar apenas fragmentos do genoma original, o agrupamento destas sequências visando à reprodução do genoma do organismo não é uma tarefa trivial.

A “montagem” do genoma depende da sobreposição das sequências para

reproduzir uma sequência contínua única denominada contig. Para garantir que este resultado seja confiável torna-se necessário que cada posição de base do cromossomo seja representada várias vezes, fazendo com que o número de bases sequenciadas atinja até dez vezes ou mais o tamanho do genoma original (BOUCK *et al*., 1998). Dentro deste cenário vários problemas podem ocorrer como agrupamento de regiões repetidas, regiões onde o sequenciamento obteve baixa qualidade, compressão de base no sequenciamento, e até regiões com baixa cobertura devido ao caráter aleatório do sequenciamento (EWING *et al*., 1998).

Para tratar todas estas variáveis, junto com os sequenciadores foram desenvolvidos programas de montagem chamados em inglês de *Assemblers* que permitem a “montagem” deste “quebra-cabeça”. Dentre eles destacam-se o pacote Phred/Phrap/Consed (GORDON *et al*., 1998), CAP/PCAP (HUANG *et al*., 2003), ARACHNE (BATZOGLOU *et al*., 2007), Celera Assembler (MYERS *et al*., 2000), “Genome Analyzer” (Illumina) e o “GS De Novo Assembler” (Roche).

1.4 ANOTAÇÃO

Uma vez obtida à sequência genômica é necessário agregar informação biológica a ela. Esta caracterização é chamada de anotação onde são identificadas regiões codificadoras com base na parte estrutural (códon de início, códon de término, quantidade de GC) e regulatória (região promotora, sítio de ligação de ribossomo). Vários programas podem ser utilizados nesta etapa para anotar uma sequência genômica como os apresentados na Tabela 2.

TABELA 2. PROGRAMAS DE ANOTAÇÃO

Programa	**Característica principal**
Orfinder http://www.ncbi.nlm.nih.gov/projects/gorf/ Jorfinder (MENDES *et al*., 2007) Glimmer (SALZBERG *et al*., 1998)	Busca por ORF/genes
BLAST (ALTSCHUL *et al*., 1990**)**	Comparações de candidatos a genes com sequências depositadas em bancos de dados públicos
Mummer (KURTZ *et al*., 2004) ClustalW (HIGGINS *et al*., 1988)	Análises comparativas entre genomas
ClustalW (HIGGINS *et al*., 1988) Mega (KUMAR *et al*., 2008) Phylip (BAUM., 1989)	Inferências filogenéticas e estruturais dos genes/genoma
KAAS (MORYA *et al*., 2007)	Determinação de vias metabólicas e sistemas de transporte

1.5 COMPARAÇÃO DE GENOMAS

Genoma é todo material genético contido no organismo incluindo toda a informação necessária para manutenção e transferência da vida. A informação biológica contida em um genoma é codificada em ácidos ribonucléicos (DNA) e é dividida em unidades chamadas genes, que por sua vez codificam proteínas e RNAs (LENINGHER, 2002). A grande quantidade de sequências genômicas nos diversos

bancos de dados públicos tem contribuído para inúmeros avanços em diversos temas como a diversidade bacteriana, genéticas das populações, a estrutura de *operons*, de elementos genéticos móveis (do inglês, *mobile genetic elements* – MGE) e a transferência genética horizontal (do inglês, *horizontal gene transfer* – HGT) (BINNEWIES *et al*., 2006).Por exemplo a sequência de bactérias patogênicas e comensais possibilita uma análise detalhada de aspectos como, por exemplo, as interações entre elas e seus hospedeiros.

Na anotação de um genoma, é importante compará-lo com outros genomas de organismos filogeneticamente relacionados devido à presença de genes pertencentes a famílias ancestrais conservadas. Nos genomas dos Domínios *Bacteria* e *Archeae*, o número de genes conservados é de 70% (TATUSOV *et al*., 2000). Além da identificação de genes a comparação entre genomas possibilita a identificação de exons, introns, sequências regulatórias, RNAs, entre outros. Dependendo do objetivo, estas comparações podem ser realizadas de maneiras diferentes utilizando programas específicos. Para o alinhamento entre os genomas pode ser utilizado o Mummer, programa que apresenta bom desempenho para alinhamentos locais; o programa SSEARCH (PEARSON *et al*., 1991), que traz estatísticas sobre o alinhamento ou BLAST (ALTSCHUL *et al*., 1990) que apresenta a sequência alinhada. Para identificação das diferenças entre genes similares, domínios funcionais, diferenciação de proteínas, inserções e deleções nos genomas pode ser utilizado o programa ClustalW (THOMPSON *et al*., 1994).

1.6 *Herbaspirillum spp*.

As Beta-proteobactérias *Herbaspirillum seropedicae* e *Herbaspirillum rubrisubalbicans* são organismos diazotróficos, associativos e endofíticos (BALDANI *et al*., 1986, 1996), encontrados principalmente no interior de plantas de interesse agrícola, como milho, trigo, sorgo, cana-de-açúcar e bananeira (BALDANI *et al*., 1986, BALDANI *et al*., 1997., CRUZ et al., 2001). Trabalhos mostraram que *Herbaspirillum rubrisubalbicans e Herbaspirillum seropedicae,* tem capacidade de colonizar raízes, caules e folhas de diferentes grupos de plantas, principalmente gramíneas (BALDANI *et al.*, 1992). A expressão dos genes *nif*, responsáveis pela fixação biológica de nitrogênio, foram identificadas em *Herbaspirillum seropedicae*

por técnicas de biologia molecular (JAMES *et al*., 2002; RONCATO-MACCARI *et al*., 2003; YOU *et al*., 2005). A contribuição da fixação biológica de nitrogênio por *Herbaspirillum sp.* foi demonstrada em arroz e cana-de-açúcar através da incorporação de N_2 (BALDANI *et al*., 2000; ELBELTAGY *et al*., 2001; JAMES *et al*., 2002; OLIVEIRA *et al*., 2002). Com base nestes resultados estas bactérias podem contribuir para uma diminuição de custo de produção destas culturas sendo uma alternativa aos fertilizantes industriais (COCKING *et al*., 2003). Estudos também mostram que esses organismos são capazes de produzir análogos a fitormônios como auxinas e giberelinas em associação com plantas (BASTIÁN *et al*., 1998). O genoma estrutural da bactéria *Herbaspirillum seropedicae* foi obtido pelo Programa Genoma do Paraná (GENOPAR) (http://www.genopar.org), sendo o primeiro genoma a ser sequenciado no estado do Paraná. O GENOPAR também sequenciou parcialmente o genoma da bactéria *Herbaspirillum rubrisubalbicans.* Entretanto, a sequência genômica dos dois organismos ainda não foram finalizadas.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Finalizar a montagem do genoma de *Herbaspirillum seropedicae*, utilizando sequenciamento de tecnologia Sanger e 454 e utilizar a sequência como referência para a identificação de genes de *Herbaspirillum rubrisubalbicans* a partir de sequenciamento genômico parcial utilizando tecnologia Sanger.

2.2 OBJETIVOS ESPECÍFICOS

- Montagem de sequências contíguas do genoma da bactéria *H. seropedicae* utilizando sequenciamento de tecnologia Sanger e 454;
- Determinação da ordem das sequências contíguas obtidas
- Estimativa do tamanho das lacunas entre os contigs remanescentes e desenho de oligonucleotídeos iniciadores para fechamento a partir de amplificação por PCR e sequenciamento;
- Identificação das regiões repetidas do genoma contendo operons de RNA ribossomais (rRNA) e sua ligação nas sequências contíguas obtidas na montagem;
- Transpor a anotação das ORFs do genoma de *H. seropedicae* para a montagem final;
- Montar parcialmente o genoma da bactéria *H. rubrisubalbicans*;
- Identificar genes no genoma de *H. rubrisubalbicans*, a partir da montagem parcial, utilizando o genoma completo de H. seropedicae como referência.

3 MATERIAL E MÉTODOS

3.1 SEQUÊNCIAS DO PROGRAMA GENOPAR

As sequências obtidas pelo método de terminadores dideoxinucleotídeos marcados utilizadas neste trabalho foram produzidas pelo Programa GENOPAR (http://www.genopar.org) e foram obtidas a partir de diversas bibliotecas construídas por clonagem em plasmídeos, com insertos de ~1000 a ~3.000 pb, e cosmídeos com insertos de ~40.000 pb. Os fragmentos de DNA clonados foram sequenciados pelo método de terminadores fluorescentes e analisados nos sequenciadores automáticos MegaBACE 1000 (Amersham Biosciences) e ABI 377 (Applied Biosystems).

O sequenciamento contou com a coordenação geral do Prof. Dr. Fabio de Oliveira Pedrosa do Departamento de Bioquímica e Biologia Molecular da Universidade Federal do Paraná e a participação de vários laboratórios de instituições dos estados do Paraná e Santa Catarina, que alimentaram o banco de dados do projeto, centralizado no laboratório de Bioinformática do Núcleo de Fixação de Nitrogênio da Universidade Federal do Paraná – UFPR (Tabela 3).

TABELA 3. INSTITUIÇÕES PARTICIPANTES DO PROJETO GENOPAR

Instituição	Coordenador	Estado
Universidade Federal do Paraná (UFPR)	Prof. Dr. Fábio de Oliveira Pedrosa Prof[a]. Dra. Maria Luiza Petzl-Erler	Paraná
Pontifícia Universidade Católica do Paraná (PUC-PR)	Prof. Dr. Humberto Maciel França Madeira	Paraná
Instituto Agronômico do Paraná (IAPAR)	Dr. Luiz Gonzaga Esteves Vieira	Paraná
Universidade Estadual de Londrina (UEL)	Prof[a]. Dra. Maria Helena Pelegrinelli Fungaro	Paraná
Centro Nacional de Pesquisa de Soja da Embrapa (Embrapa – CNPSo)	Dra. Mariângela Hungria	Paraná
Universidade Estadual de Maringá (UEM)	Prof. Dra. Maria Aparecida Fernandez	Paraná
Universidade Estadual de Ponta Grossa (UEPG)	Prof. Dr. Ricardo Antônio Ayub	Paraná
Universidade Paranaense (UNIPAR)	Prof. Dr. Nelson Barros Colauto	Paraná
Universidade Estadual do Oeste do Paraná (UNIOESTE)	Prof[a]. Clarice Aoki Osaku	Paraná
Universidade Federal de Santa Catarina (UFSC)	Rubens Onofre Nodari Hernan Terenzi Edmundo Carlos Grisard	Santa Catarina

Fonte: GENOPAR.

No total foram obtidas 132.287 leituras de sequência pelo método Sanger para o genoma de *H. seropedicae,* totalizando 121 megabases (Mb).

Além destas foram obtidas duas corridas no sequenciador 454 GS FLX Titanium realizado nos Estados Unidos da América pela empresa Creative genomics (http://www.creative-genomics.com/) resultando em 1.220.352 leituras de

sequências, totalizando 456 Mb.

Já para o genoma de *H. rubrisubalbicans* foram obtidas 24.757 leituras de sequências pelo método Sanger, totalizando aproximadamente 22 Mb sequenciados pelo consórcio GENOPAR.

3.2 SISTEMA OPERACIONAL

O sistema operacional utilizado foi baseado na plataforma GNU/Linux sendo utilizadas as distribuições Kurumin-NG (http://www.gdhpress.com.br/kurumin-ng/) baseada no Ubuntu 8.0 (http://www.ubuntu.com/) para desktop e Debian (http://www.debian.org/) para os servidores.

3.3 LINGUAGENS DE PROGRAMAÇÃO

Para processar os dados de sequenciamento e análises do genoma foram desenvolvidos vários *scripts* para integração dos programas de Bioinformática utilizados. As linguagens utilizadas foram Perl, Bash, e Sql.

3.3.1 Perl

Perl (*"Practical Extraction And Report Language"*) é uma linguagem de programação estável e multiplataforma usada em todos os setores, como no desenvolvimento de aplicações web e *scripts* em ambiente UNIX sendo também portável a outros sistemas como Windows, MSDOS, *Macintosh*, entre outros (http://perldoc.perl.org/perlintro.html). Além da fácil manipulação de dados em texto, que a linguagem proporciona, também contem um pacote de ferramentas de código livre, voltadas para o tratamento de dados biológicos, o BIOPERL (http://www.bioperl.org/wiki/Main_Page).

3.3.2 Bash

Desenvolvido para o projeto GNU (*GNU is Not Unix*) da *Free Software Foundation* (http://www.gnu.org), Bash é um interpretador de comandos largamente utilizado em plataforma Unix e GNU/Linux onde o usuário pode realizar sequências de comandos ou automatizar tarefas na forma de *scripts*. Esta característica é útil para manipulação de dados armazenados em arquivos em modo texto, para formatação de dados e para a automação de fluxo de execução de programas, onde os resultados gerados e gravados em arquivos de saída de um programa são usados como dados de entrada para outro programa (*Pipelines*).

3.3.3 Sql

Visando a segurança e integridade dos dados gerados, o Programa Genopar tem parte de suas informações armazenadas em bancos de dados. Para acessar o seu conteúdo foi utilizada a linguagem Sql *(Structured Query Language)* que é padrão para bases relacionais, e apresenta uma sintaxe de pesquisa declarativa que também permite interação com outras linguagens (CHAMBERLIN *et al*., 1981).

3.3.4 Mysql

Mysql é um programa de gerenciamento de banco de dados que utiliza linguagem Sql. É um banco de alta portabilidade, compatível com as plataformas e sistemas operacionais mais utilizados como Linux, Windows, Solaris, MacOS (http://www.mysql.com/).

3.4 PROGRAMAS DE BIOINFORMÁTICA

3.4.1 Artemis

Ferramenta desenvolvida pelo Instituto Sanger (http://www.sanger.ac.uk/Software/Artemis/) de anotação de genoma que permite a visualização das sequências e as anotações feitas nos diversos formatos utilizados. O programa Artemis (RUTHERFORD *et al*., 2000*)* é escrito em Java, e está disponível para diversas plataformas, como UNIX*,* Macintosh e Windows. O programa utiliza sequências armazenadas em texto nos formatos EMBL e GENBANK ou FASTA.

3.4.2 BLAST

BLAST (*Basic Local Alignment Search Tool*) considera regiões de similaridade entre as sequências locais. O programa compara sequências de nucleotídeos ou proteínas com sequências de uma base de dados e calcula a significância estatística dos alinhamentos. O programa BLAST pode ser usado para inferir relações evolutivas e funcionais entre as sequências, assim como para ajudar a identificar os membros de famílias protéicas (ALTSCHUL *et al*., 1990**).**

3.4.3 ClustalW

Programa usado para o alinhamento global múltiplo de sequências de DNA ou proteínas (HIGGINS *et al*., 1988). Pode ser utilizado na identificação de regiões conservadas das sequências e para inferências filogenéticas.

3.4.4 Newbler

Programa para montagem de genomas, desenvolvido especificamente para processar os dados gerados pelo pirossequenciador 454-GS-series (ROCHE). O

programa pode ser utilizado tanto através da linha de comando quanto via interface, utilizando Java GUI. Tem como dado de entrada o arquivo binário gerado pelo sequenciador 454, onde se encontram todas as informações referentes ao pirossequenciamento. O programa também permite o uso de sequências de outras químicas de sequenciamento desde que estejam no formato FASTA. O montador se utiliza de sobreposições para a formação de contigs utilizando a comparação direta entre os *flowgrams,* equivalentes aos eletroforetogramas da química Sanger, através do módulo *Overlapper.* O módulo *Unitigger* agrupa as leituras de sequências que possuem sobreposições consistentes em *unitigs.* Cada *unitig p*ode ser obtido devido a presença de regiões de repetição no genoma ou por não haver sequência para continuar a extensão. Por fim o módulo *Multialigner* faz o processo de otimização dos *unitigs* em contigs onde qualquer deficiência identificada nas outras etapas é corrigida e os contigs são estendidos.

3.4.5 Phred/Phrap/Consed

Phred/Phrap/Consed é um pacote de programas que utiliza arquivos de sequenciamento de DNA (eletroforetograma). O programa Phred é responsável por fazer a chamada de bases e estimar a probabilidade de erro de cada base da sequência (GORDON *et al*., 1998). O programa utiliza parâmetros como amplitude e espaçamento entre os picos em relação ao sinal de fundo (background) do eletroforetograma para discriminar erros na chamada de bases (Figura 6). Os valores de Phred variam de 0 a 97 sendo calculados pela fórmula:

$$q = -10 \times \log_{10}(p)$$

onde *q* é a qualidade e *p* é a probabilidade de erro estimado para uma base (Ewing *et al*., 1998)., Um valor *p* igual a 30 resulta em uma base errada em cada mil bases sequenciadas e um valor *p* igual a 20 resulta em uma base errada em cada cem bases sequenciadas. Quanto maior a o valor Phred melhor a qualidade da base.

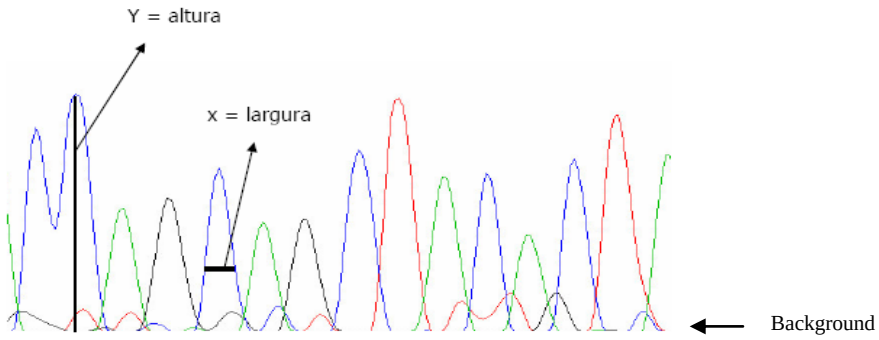


FIGURA 6. Eletroforetograma.

O programa Phrap é um montador para sequências de DNA e utiliza a parâmetros de alinhamento e as qualidades atribuídas pelo programa Phred para realizar a montagem do genoma em contigs, a partir das sobreposições entre as leituras do sequenciamento de DNA. Junto com as sequências consenso, formada por contigs, o programa fornece informações sobre a montagem ajudando no tratamento de eventuais problemas encontrados. A ferramenta *Consed/Autfinish* (GORDON *et al*., 1998) é uma ferramenta para visualização, edição e finalização das montagens criadas com os programas Phred/Phrap. Entre os recursos de finalização destacam-se: a capacidade de escolha e sugestão de oligonucleotídeos iniciadores, a identificações de regiões com problemas de montagem como baixa qualidade nas sequências, fornece sugestões de sequenciamentos adicionais e agrupa os contigs em scaffolds (contigs ordenadas por evidências de ligação).

3.4.6 SSAHA

SSAHA é um programa para alinhamento de sequências que utiliza vetores e programação dinâmica, sendo capaz de mapear rapidamente leituras de sequências em uma sequência de referência. Suporta leituras de sequências das novas tecnologias de sequenciamento e também sequências de pares. Pode ser utilizado para o mapeamento de leituras do método Sanger sem perder a referência dos pares, pois permite ao usuário inserir valores de distância a fim de verificar a consistência do mapeamento de cada sequência (NING *et al*., 2001).

3.5 MÉTODOS USADOS PARA A FINALIZAÇÃO DA MONTAGEM

3.5.1 Poda de bases com baixa qualidade

O processo de poda é realizado com auxílio do programa Phred responsável por fazer a chamada de bases do arquivo resultante do sequenciamento (eletroforetograma) (Figura 7). Este programa permite fazer uma varredura das sequências a fim de preservar o maior segmento de bases de alta qualidade. O descarte das regiões de baixa qualidade normalmente ocorre nas extremidades. Para realizar a poda o programa Phred utiliza os valores de probabilidade de erro calculados a partir do valor de qualidade, e subtrai de cada base um valor de corte (0.05 por padrão), obtendo um valor de base final. Então o algoritmo percorre a sequência em janelas de 20 pb apartir de cada base e procura o segmento de maior valor da sequência onde este valor é a soma dos valores das bases,. A base é então descartada se apresentar um valor negativo (EWING *et al*., 1998).

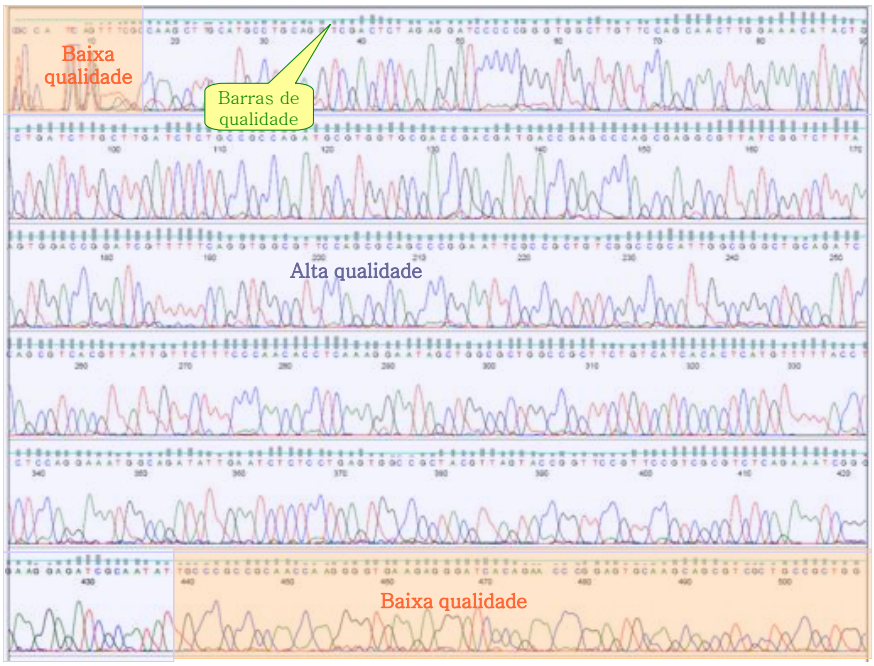


FIGURA 7. Regiões de alta e baixa qualidade.
Eletroforetograma mostrando regiões de alta qualidade na região central (azul) e baixa qualidade nas extremidades. Pode-se observar a que os picos não são bem definidos nestas regiões. Com a realização da poda as bases em vermelho são descartadas.
FONTE: o autor.

3.5.2 Montagem

As montagens foram realizadas utilizando o montador Phrap que foi desenvolvido inicialmente para trabalhar com dados da química Sanger. Este montador consegue aproveitar ao máximo as informações destas leituras de sequências. Além de gerar os contigs através da sobreposição, ele aproveita também, com o objetivo de aumentar a certeza das sobreposições, a referência entre a sequência amplificada pelos primers universais e reversos de cada clone. O algoritmo trabalha atribuindo pontuações aos alinhamentos e foi usado com os parâmetros padrões de *minmatch* 20, *penalty* -2 e *minscore* 20. O parâmetro *Minmatch* é responsável pelo tamanho mínimo da sequência de bases que o algoritmo irá tentar alinhar, visando a sobreposição e extensão da sequência contígua. Quanto maior este número, maior a rigorosidade do alinhamento e menor o tempo de execução. No parâmetro *Penalty* o algoritmo irá utilizar o valor para penalizar uma troca de base no alinhamento. Por padrão um alinhamento de base perfeito recebe uma pontuação de +1, uma troca de base recebe a pontuação de -2,

a falta de uma base a pontuação de -4, e a falta de uma base extendida por um resíduo recebe o valor de -3 (GORDON *et al*., 1998). Assim, quanto maior o valor do número negativo, maior a rigorosidade da montagem. Por fim, o parâmetro *minscore* avalia o valor do alinhamento obtido, pelo fato do alinhamento ser pontuado como comentado anteriormente, quanto maior o valor utilizado neste parâmetro maior a rigorosidade.

Outro montador utilizado, o *Newbler*, que faz parte do pacote “GS *De Novo* Assembler” desenvolvido pela Roche foi utilizado para alinhar as leituras do sequenciador 454. Os parâmetros utilizados para a montagem foram *Seed step* 12, *Seed length* 16, *Minimum overlap length* 25, *Minimum overlap identity* 90%, *Alignment identity score* 2, *Alignment difference score* -3, *Large contig Threshold* 100, onde *Seed step* é a distância entre os locais escolhidos para seleção de regiões para comparação; *Sedd length* é o número de bases usado para cada intervalo de comparação; *Minimum overlap length* é o número mínimo de sobreposição utilizado pelo montador; *Minimum overlap identity* é o percentual mínimo de identidade entre os alinhamentos utilizado pelo montador; *Alignment identity score* é utilizado quando sobreposições múltiplas são identificadas a fim de ordenar os valores das sobreposições para o uso na progressão do alinhamento; *Alignment difference score* é utilizado quando sobreposições múltiplas são identificadas a fim de ordenar os valores de diferença entre as sobreposições para o uso na progressão do alinhamento; *Large contig Threshold* é o tamanho mínimo que o contig deve conter para ser separado para análise, embora o montador crie contigs menores.

3.5.3 Verificação de inconsistências

Após criação da montagem foi realizada uma varredura nos contigs criados pelo montador visando validar o posicionamento das leituras de sequências. Como os fragmentos genômicos clonados eram aproximadamente de 1-3 kb para plasmídeo e 40 kb para cosmídeos e o sequenciamento do fragmento é parcial, compreendendo apenas cerca de 400 pb de suas extremidades, para cada fragmento existe um par de sequências obtido com o *primer* universal, denominado “b” e outro obtido com o *primer* reverso, denominado “g” resultando uma região

central não sequenciada. O programa se utiliza desta informação dos pares e identifica as leituras de sequência que estão posicionadas em distâncias não condizentes com a calculada, sendo chamados de inconsistentes. Estas regiões então foram re-analisadas.

3.5.4 Análise das extremidades dos contigs

Foi criado um arquivo contendo 500pb das extremidades dos contigs. Um banco BLAST foi criado utilizando as leituras de sequências que até então não tinham sido utilizadas na montagem. Como foi realizada mais de uma montagem, foram usadas estratégias diferentes e em algumas tentativas não partimos do conjunto total de leituras de sequência para criar montagem e sim de um grupo selecionado realizando adições sucessivas do restante das leituras de sequências nos contigs. Para isso o arquivo com as extremidades dos contigs foi submetido à busca por similaridade utilizando-se o programa BLAST, com valor E igual a 10^{-10} visando obter apenas os melhores alinhamentos. Quanto menor é o valor de E, menor é a chance da ocorrência ao acaso de outro alinhamento com a mesma similaridade, este parâmetro está relacionado à rigorosidade do alinhamento. Esta estratégia permitiu a extensão dos contigs e fechamentos das falhas na sequência.

3.5.5 Inserção das leituras de sequências na montagem

As leituras de sequências selecionadas foram inseridas na montagem através da ferramenta *addNewReads* do pacote *PhredPharpConsed*. Esta ferramenta consiste na adição de um bloco de sequências fornecido pelo usuário, sendo que o montador o ancora na montagem obedecendo aos critérios de identidade sem a necessidade de iniciar uma nova montagem.

3.5.6 Ressequenciamento dos clones e construção de *primers*

Leituras de sequências com baixa qualidade não fornecem respostas

concretas nos alinhamentos porque apresentam maior chance de erro de sequenciamento. Com a realização da poda algumas leituras de sequência não foram utilizadas na montagem. Visando a recuperação das sequências de DNA destas regiões, foram realizados ressequenciamentos de clones selecionados.

O ressequenciamento foi realizado pelo grupo de Fixação de Nitrogênio do departamento de Bioquímica de Biologia Molecular da UFPR e foi utilizado nos casos onde as leituras de sequências garantiam a ligação física entre dois contigs, através das leituras de sequências “b” e “g”, respeitando sempre a distância estipulada para cada biblioteca.

Para as regiões onde o ressequenciamento dos clones não obteve resultado positivo foram desenhados *primers* a fim de eliminar a falha na sequência.

3.5.7 União entre Contigs

Primeiramente foi realizada uma comparação entre as regiões utilizando a ferramenta *cross_match* que faz parte do pacote Phred/Pharp/Consed. Esta ferramenta é semelhante à ferramenta BLAST, que identifica sequências idênticas ou similares. A vantagem do programa *cross_match* é considerar os valores de qualidades das bases das sequências atribuídos pelo programa Phred além da indenidade das sequências para compor o alinhamento.

Os resultados passaram por uma verificação manual, e a região a ser unida foi avaliada com relação aos parâmetros de qualidade da sequência e distância dos pares “b” e “g”.

A união foi realizada utilizando a ferramenta *Join Contig* presente no programa Consed que une as regiões através da sobreposição apontada pelo *cross_match.*

3.6 ANÁLISE DE PREFERÊNCIA DE USO DE CÓDONS

As ORFs anotadas no genoma de *H. seropedicae* foram submetidas a análise de códons, com auxílio dos programas GCUA (Graphical Codon Usage Analyser –MCLNERNEY, 1998) e CodonW (PENDEN, 1999). A análise permitiu

estabelecer a preferência de uso de códon das ORFs a partir do cálculo de diferentes índices que verificam distorções na utilização de códons sinônimos.

Foram analisados os índices de conteúdo de GC na terceira posição do códon (*CG3s*), o número efetivo de códons (*Enc*), índices de hidrofobicidade de cada aminoácido (*Gravy*), índice de adaptação de códons para genes altamente expressos (CAI) e análise de correspondência (*COA*) a fim de analisar a diferença códons.

4. HARDWARE

4.1 Desktop

Sistema Operacional – GNU/Linux – Kurumin NG
Kernel 2.6.24-18-generic
Arquitetura – i686
Processadores – Intel Pentium Dual Core 2.00GHz
Memória – 2 GB

4.2 Servidores

- Athos

Sistema Operacional – GNU/Linux Debian 4.0 x64
Kernel – 2.6.18-5-amd64
Arquitetura – x86_64
Processadores – Intel Core2 Quad 2.4 GHz Q6600
Memória – 8GB

5 RESULTADOS E DISCUSSÃO

5.1 ANÁLISE DAS SEQUÊNCIAS USADAS NA MONTAGEM

As leituras de sequências obtidas pelo consórcio GENOPAR para o sequenciamento genômico das bactérias *H. seropedicae* e *H. rubrisubalbicans* pelo método Sanger e aquelas obtidas por pirossequenciamento, foram analisadas em relação ao seu tamanho e qualidade. Foram obtidas 137.287 sequências com um tamanho médio de 966 pb (Figura 8.A) utilizando o método de Sanger, totalizando 121 Mb para o genoma do *H. seropedicae*. Visando a remoção das extremidades de baixa qualidade o conjunto de leituras foi submetido à poda resultando em um comprimento médio de aproximadamente 600 pb (Figura 8.B) com qualidade média de Phred 33.

Outro conjunto de dados analisado para o genoma de *H. seropedicae* foi as leituras de sequências obtidas através do sequenciador “454 GS FLX Titanium”, totalizando 1.220.352 leituras de sequências com um comprimento médio de 430 pb, resultando em aproximadamente 456 Mb sequenciados com qualidade média Phred 35 para as bases e não precisaram da poda pois a pré-análise já tinha sido feita pela empresa que o sequenciou.

O genoma de *H. rubrisubalbicans* contém atualmente 24.757 leituras de sequência pelo método Sanger, totalizando aproximadamente 22 Mb com um comprimento médio por leitura de 891 pb (Figura 8.C). As leituras de sequências também foram analisadas a fim de eliminar bases com baixa qualidade fazendo com que o tamanho médio fosse reduzido para 600 pb (Figura 8.D).

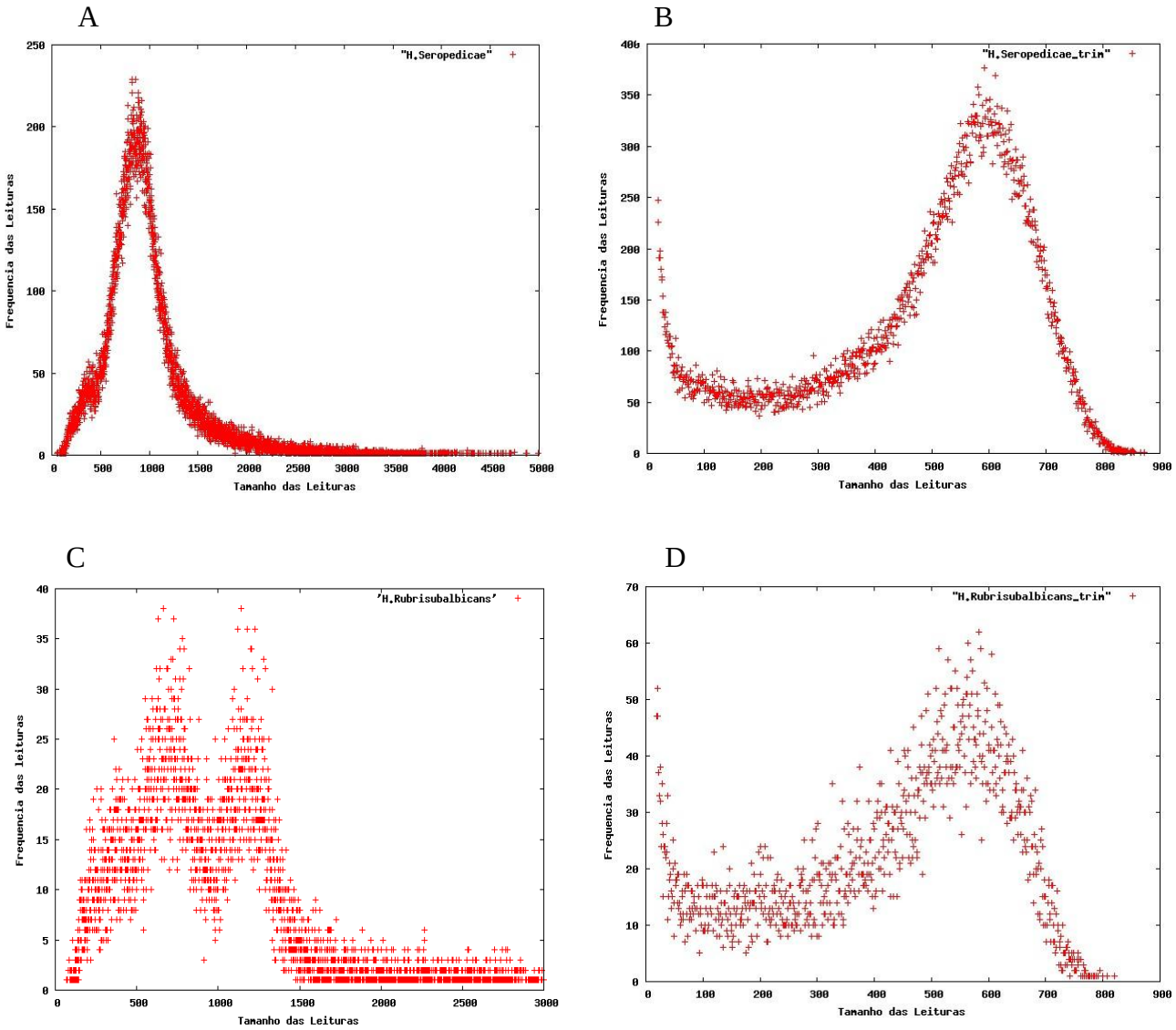


FIGURA 8. Distribuição das leituras de sequências pelo método Sanger.
Diferença em relação ao tamanho e frequência das leituras de sequências de *H. serpopedicae* A e B e *H. Rubrisubalbicans* C e D. A e C representam o conjunto de dados antes da poda e B e D depois da poda.

Como podemos ver na diferença entre a Figura 8.A e 8.B e 8.C e 8.D a poda das sequências causa um grande impacto nas características dos dados. O sequenciamento utilizando o método de Sanger produz sequências de no máximo 900 pb com qualidade Phred 20 (MOROZOVA *et al*., 2008). As sequências com tamanho maior certamente contem bases com baixa qualidade que não serão aproveitadas na montagem justificando sua retirada. O resumo das análises das sequências pode ser observado na Tabela 4.

TABELA 4. RESULTADO DA ANÁLISE DAS LEITURAS DE SEQUÊNCIAS.

	Sanger HS	Sanger HR	454 HS
Número de leituras de sequências	137287	24747	1220352
Total de leituras de sequências em Mb	121 Mb	22 Mb	430 Mb
Tamanho médio das leituras de sequências	966 pb	891 pb	430 pb
Tamanho médio das leituras de sequência após a poda	600 pb	600 pb	-

5.2 O GENOMA DE *H. seropedicae*

Durante o processo de análise foram criadas quatro montagens, pois a MontagemV1 já estava pronta. As duas primeiras denominadas MontagemV2 e MontagemV3 utilizaram apenas sequências de leituras pelo método Sanger enquanto a quarta (MontagemV4) e a quinta (MontagemV5) montagem corresponderam a montagens híbridas utilizando as sequências Sanger e 454. Cada uma das montagens citadas acima foi processada de forma diferente, conforme como descrito a seguir.

5.2.1 MontagemV1

Esta é a primeira montagem do genoma de *H. seropedicae* e utilizou 91.546 leituras de sequências de um total de 132.287 obtidas pelo método Sanger, gerando 287 contigs (Tabela 5). A montagem desenvolvida por Augusto Favetti foi utilizada para a anotação do genoma de *H. seropedicae.* Dentre suas características destacam-se a um número alto de contigs e leituras de sequência inconsistentes, podendo indicar erros de montagem. A consistência da montagem foi avaliada através do cálculo do tamanho dos insertos em cada biblioteca genômica a partir da distância média entre os pares de leituras de sequências na montagem. Os contigs obtidos nesta montagem foram anotados utilizando o programa Glimmer onde foram identificados 5.100 genes.

A montagem foi editada com a retirada das leituras de sequências inconsistentes e utilizada como base para o fechamento do genoma de *H.*

*seropedicae. S*uas características estão resumidas na Tabela 5. As análises, melhorias e criação de novas montagens foram realizadas a fim de corroborar com as informações contidas nesta primeira montagem e finalizá-la.

TABELA 5. ESTATÍSTICAS DA MONTAGEMV1

	MontagemV1
Número de bases	5593394 pb
Número de contigs	287
Maior contig	169353 pb
Menor contig	363 pb
Tamanho médio dos contigs	19489 pb
Número de Scaffolds	52
Leituras de sequências totais	132287
Leituras de sequências na montagem	91546
Singlets	22147
Leituras de sequências com vetor	25703
Total de clones	54209
Clones inconsistentes	6069

Fonte: GENOPAR.

5.2.2 MontagemV2

A MontagemV2 visou obter um número de bases não redundantes similar a MontagemV1 porém aumentando sua qualidade geral, principalmente eliminando as inconsistências dos pares “b” e “g” das leituras de sequência. Para este fim as leituras de sequências do método Sanger foram podadas, o que não ocorreu na MontagemV1. Os contigs foram obtidos utilizando o programa Phrap, através do pacote Phred/Phrap/Consed.

Inicialmente a montagem obteve um número excessivo de contigs (639), o

que já era esperado devido à diminuição do comprimento das leituras de sequências após a poda. Como o mesmo conjunto de dados foi utilizado, o principal objetivo foi a identificação do maior número de contigs ligados consistentemente pelos pares “b” e “g”. Os contigs ligados pela referência dos pares “b” e “g” de uma mesma leitura são denominados *scaffolds.* A ferramenta *Reorder Contigs* do programa Consed possibilitou o agrupamento destas sequências manualmente fornecendo assim uma visão mais clara das possíveis ligações. Por fim, visando eliminar as lacunas presentes, foram desenhados *primers* nas extremidades dos contigs para a amplificação e o sequenciamento do fragmento faltante, e inseri-lo na montagem.

A montagem evoluiu, atingindo-se 467 contigs e eliminando 90% das leituras de sequências inconsistente, porém a montagem foi interrompida pelo excesso de contigs em comparação a MontagemV1 (Tabela 6).

TABELA 6. ESTATÍSTICAS DA MONTAGEMV2.

	MontagemV1	MontagemV2
Número de bases	5593394	5473761 pb
Número de contigs	287	467
Maior contig	169353 pb	111317
Menor contig	363 pb	224 pb
Tamanho Médio dos contigs	19489 pb	11722 pb
Número de Scaffolds	52	66
Leituras de sequências totais	132.287	132.287
Leituras de sequências na montagem	91546	83516
Singlets	22147	2869
Leituras de sequências com vetor	25703	6300
Total de Clones	54209	48773
Clones Inconsistentes	6069	508

5.2.3 MontagemV3

Diferentemente da MontagemV2, na MontagemV3 as leituras de sequências não foram podadas. As leituras de sequência que apresentavam os pares (“b” e ”g”) e valores de qualidade maior ou igual à Phred 20 foram aproveitadas primeiramente e submetidas ao programa Phrap. Esta abordagem possibilitou a criação de contigs mais longos que a MontagemV2, onde foi realizada poda, e também com maior qualidade que a MontagemV1 pela seleção da qualidade. Outra diferença foi a cobertura inferior em relação às outras tentativas devido ao número de sequências de leituras utilizados ser menor. A MontagemV3 foi iniciada com 108 contigs, porém utilizando inicialmente 50.741 leituras de sequências de um total de 91.546 utilizadas na MontagemV1. Em seguida foi realizada a extensão destes contigs através da análise das extremidades dos contigs. Assim as sequências que não foram adicionadas na etapa inicial foram inseridas na montagem através da ferramenta *add new reads* do programa Consed visando a extensão dos contigs. Esse processo foi executado para todas as extremidades e após a seleção pelo BLAST e inserção foi verificada a existência de sobreposição entre as extremidades estendidas. Quando houve sobreposição entre estas extremidades estendidas pelas leituras de sequência adicionadas, os contigs foram unidos através da ferramenta *Join contigs* do programa Consed.

As inserções foram bem sucedidas, porém surgiu a necessidade de incorporar as leituras de sequências que não entraram na etapa de criação da montagem e também na análise das extremidades dos contigs. Apesar de algumas destas leituras de sequência terem conseguido incorporar os contigs existentes outras formaram contigs próprios aumentando o número de contigs totais da montagem chegando a 414 (Tabela 7). Esta montagem foi interrompida para a utilização das leituras de sequências do sequenciador 454.

TABELA 7. ESTATÍSTICAS DA MONTAGEMV3

	MontagemV1	MontagemV3
Número de bases	5593394 pb	5686616 pb
Número de contigs	287	414
Maior contig	169353 pb	168302 pb
Menor contig	363 pb	550 pb
Tamanho médio dos contigs	19489 pb	13735 pb
Número de *Scaffolds*	52	146
Leituras de sequências totais	132.287	132.287
Leituras de sequências na montagem	91.546	80119
Singlets	22147	262
Leituras de sequências com vetor	25703	9275
Total de clones	54209	48265
Total de clones inconsistentes	6069	2800

5.2.4 MontagemV4

As sequências de DNA obtidas por pirossequenciamento foram submetidas ao montador Newbler utilizando o programa *GS de novo Assembler*. A montagem criada foi importada no programa Consed para análise e edição.

Com a utilização das duas corridas 454 que perfazem um total de 1.220.352 sequências de leituras foram gerados 233 contigs. Esta montagem foi criada com a expectativa de finalizar a sequência do genoma de *H. seropediace* com pela quantidade de leituras de sequências utilizada, mas a montagem obteve um número alto de contigs (Tabela 8).

TABELA 8. ESTATÍSTICAS DA MONTAGEMV4

	MontagemV1	MontagemV4
Número de bases	5593394 pb	5477389
Número de contigs	287	233
Maior contig	169353 pb	170028 pb
Menor contig	363 pb	571 pb
Tamanho médio dos contigs	19489 pb	23508 pb
Número de *Scaffolds*	52	233
Leituras de sequências totais	132287	1220352
Leituras de sequências na montagem	91546	1196200
Singlets	22.147	14357
Leituras de sequências com vetor	25703	-
Total de clones	54209	-
Total de clones inconsistentes	6069	-

5.2.5 MontagemV5

As quatro montagens do genoma de *H. seropedicae* utilizando abordagens diferentes foram comparadas entre si com o objetivo de identificar a melhor. Através das características como maior cobertura e menor quantidade de contigs pudemos observar que a MontagemV1 manteve um melhor resultado. Como a Montagemv2 e Montagemv3 foram desenvolvidas como alternativas visando identificar problemas estruturais nos contigs da MontagemV1, e resolver a presença de clones inconsistentes, foi realizada a comparação entre as três montagens utilizando o programa BLAST. Os contigs da MontagemV1 foram utilizadas como objeto de comparação contra os contigs da MontagemV2 e MontagemV3. Para garantir alinhamentos com alta identidade foi utilizado um valor de E igual a 10^{-10}. O resultado obtido foi um valor de 99% de identidade para ambas as montagens e 88% de cobertura para a MontagemV2 e 62% para a MontagemV3 em relação os contigs

da MontagemV1 (Anexo 1).

Como as três montagens utilizaram as leituras de sequência do método Sanger obtiveram alto grau de identidade entre si, foi decidido repetir a comparação, utilizando a MontagemV1 que continha o menor número de contigs e maior cobertura, contra a MontagemV4, que possuía apenas as leituras de sequências do 454. Os resultados mostraram cobertura média entre os contigs das montagens de 90% e identidade de 99%, sendo que apenas 17 contigs de aproximadamente 1 Kb não obtiveram alinhamento (Anexo 1).

Deste modo foi possível confirmar a qualidade dos contigs da MontagemV1, já que as sequências de DNA gerados pelo sequenciador 454 apresentam alta qualidade e os contigs cobertura elevada devido o grande número de sequências usadas. Outro dado relevante foi o fato de duas montagens obtidas a partir de conjunto de dados e montadores distintos obterem resultados similares.

Assim os contigs consenso da MontagemV1 e MontagemV4 foram utilizados e submetidos de novo ao montador Phrap*.* Para isso foram criados 520 *fake traces* destes consensos utilizando o programa Phred para que o montador os utilizasse. Após a execução do montador a montagem obteve 31 contigs, mostrando uma melhora significativa comparada com as outras separadamente e também confirmando a identidade entre os contigs observados no BLAST.

Utilizando esta montagem como base, a conclusão da montagem do genoma consistiu na análise das extremidades dos contigs. Como foram utilizados os *fake traces* dos contigs, perdeu-se a referência fornecida pelos pares “b” e “g”, assim as leituras de sequências pelo método Sanger presentes nestas regiões de extremidades foram adicionadas utilizando a ferramenta *AddNewReads*. O programa *Cross_match* foi utilizado para identificar sobreposições nas extremidades dos contigs e a união destas foi realizada utilizando a ferramenta *Join Contigs.* Por fim as regiões de falha de sequências foram resolvidas através da referência entre os pares “b” e “g”, onde os clones foram ressequenciados e os contigs unidos pela ferramente *Join Contigs.*

Outro fator determinante para a união dos contigs foi a anotação do genoma realizada na MontagemV1. Com base nesta anotação foi possível identificar qual gene estava contido na região a ser unida o que nos permitiu comparar os genes presentes na vizinhança com organismos próximos ao *H. seropedicae*. Esta etapa

do trabalho foi realizada pelos Professores Drs Fabio de Oliveira Pedrosa e Emanuel Maltempi de Souza.

A MontagemV5 do genoma de *H. seropedicae* foi finalizada com uma cobertura de 5,5 Mb e suas características podem ser vistas na Tabela 9.

TABELA 9. ESTATÍSTICAS DA MONTAGEMV5

	MontagemV5
Número de bases	5513887 pb
Numero de contigs	1
Leituras de sequências totais	1352639
Leituras de sequências na montagem	1287746

5.2.6 Validação do Genoma de *H. seropedicae*

A MontagemV5 do genoma de *H seropedicae* foi submetida a restrição *in silico* visando reproduzir os resultados obtidos em laboratório por Ramos (2003) que submeteu o DNA genômico de *H. seropedicae* a digestão com a endonuclease de restrição de corte raro (SwaI - New England Biosciences) e separação dos fragmentos obtidos por eletroforese em gel agarose em campo pulsado (PFGE). A restrição *in silico* foi realizada utilizando um *script* em PERL que identificou os sítios de corte da enzima, produziu um conjunto de 12 fragmentos teóricos (Tabela 10).

O eletroforetograma obtido por Ramos (2003) foi reanalisado revelando 10 macrofragmentos (Tabela 10), aos invés de 14 originalmente propostos por Ramos. Esta diferença se deveu ao fato de Ramos ter ignorado uma banda >1800 Kb presente no gel e também pelo fato de Ramos ter considerado as bandas de 682, 550, 367 e 166 Kb como triplas ou duplas, devido ao seu formato alargado. Após reanálise do eletroforetograma estas bandas foram consideradas bandas únicas. A Tabela 10 mostra a correspondência entre os fragmentos observados e os fragmentos teóricos. Com exceção das bandas teóricas de 76 e 31 Kb, todas as outras bandas obtidas *in silico* possuem uma banda experimentalmente comprovada correspondente dentro da faixa de 1 desvio padrão. As duas bandas sem

correspondência no eletroforetograma são muito pequenas para serem detectadas pelo método utilizado. Estes resultados indicam fortemente que a Montagem V5 do genoma de *H. seropedicae* está correta.

TABELA 10. COMPARAÇÃO DOS FRAGMENTOS OBTIDOS APÓS RESTRIÇÃO COM A ENZIMA SwaI *IN SILICO* E POR PFGE DO GENOMA DE *H. seropedicae*.

Análise *in silico*	*PFGE
Fragmentos (Kb)	Fragmentos (Kb)
1995	>1800
728	682 ± 32,35
557	550 ± 20,97
486	504 ± 36,02
406	367 ± 13,63
317	295 ± 13,93
271	267 ± 13,93
261	253 ± 9,88
229	232 ± 10,37
139	166 ± 18,27
76	Não detectada
31	Não detectada
Total 5513	5511

* Fragmentos de macrorrestrição com enzima Swa I obtidos por RAMOS, 2003. Tamanho dos fragmentos em Kb ± desvio padrão.

A anotação realizada na MontagemV1 também foi fundamental para a validação da montagem do genoma. Com ela foi possível identificar a disposição dos grupos de genes conservados e compará-los com organismos próximos ao *H.seropedicae*.

5.2.7 Anotação do genoma de *H.seropedicae*

Anotar é o processo de agrupar todas as informações disponíveis e relacionar com as sequências de DNA, como por exemplo, inferir a função de uma determinada ORF por comparação com genes conhecidos (Gregory *et al*., 2006). A anotação do genoma de *H. seropedicae* foi realizada no início do Projeto GENOPAR utilizando MontagemV1 através da plataforma de anotação automática GAnM (Genome Annotation Module) desenvolvida por Augusto Favetti. A plataforma utilizou os programas Glimmer (SALZBERG *et al* ., 1998) para identificar genes candidatos, RBS Finder (SUZEK *et al*., 2001) para identificar os sítios de ligação de ribossomo e BLAST (ALTSCHUL *et al*., 1990**)** para comparação com sequências depositadas em bancos de dados públicos como o NCBI (http://www.ncbi.nlm.nih.gov/). Após a finalização, o genoma de *H. seropedicae* foi reanotado onde foram identificados novos genes. A anotação também foi revisada manualmente pelos Professores Fabio de Oliveira Pedrosa e Emanuel Maltempi de Souza que também efetuaram a correção de erros de fase de leitura (*frameshifts*). O programa Artemis (RUTHERFORD *et al*., 2000) foi usado para a edição e reanotação, resultando em 4.737 genes, três operons ribossomais contendo os genes 16SrRNA-23SrRNA-5SrRNA (nesta ordem) e 55 tRNA. Cerca de 88% do genoma de *H. seropedicae* consiste de regiões codificadoras (Tabela 11; Figura 9). Os genes identificados no genoma também foram classificados em suas categorias funcionais utilizando a base de dados Cluster of Ortholog Groups (COG) (Figura 10).

TABELA 11. CARACTERÍSTICAS ESTRUTURAIS DO GENOMA DE *H. seropedicae*

Número de bases	5513887 pb
Conteúdo de G+C	63,4%
Total de regiões codificadoras	88%
tRNAS	55
Operons rRNA 16S-23S-5S	3
ORFs codificadores para proteínas	4737
ORFs com função conhecida	3617
Tamanho médio das ORFs	1029 pb
Maior ORF	27483 pb

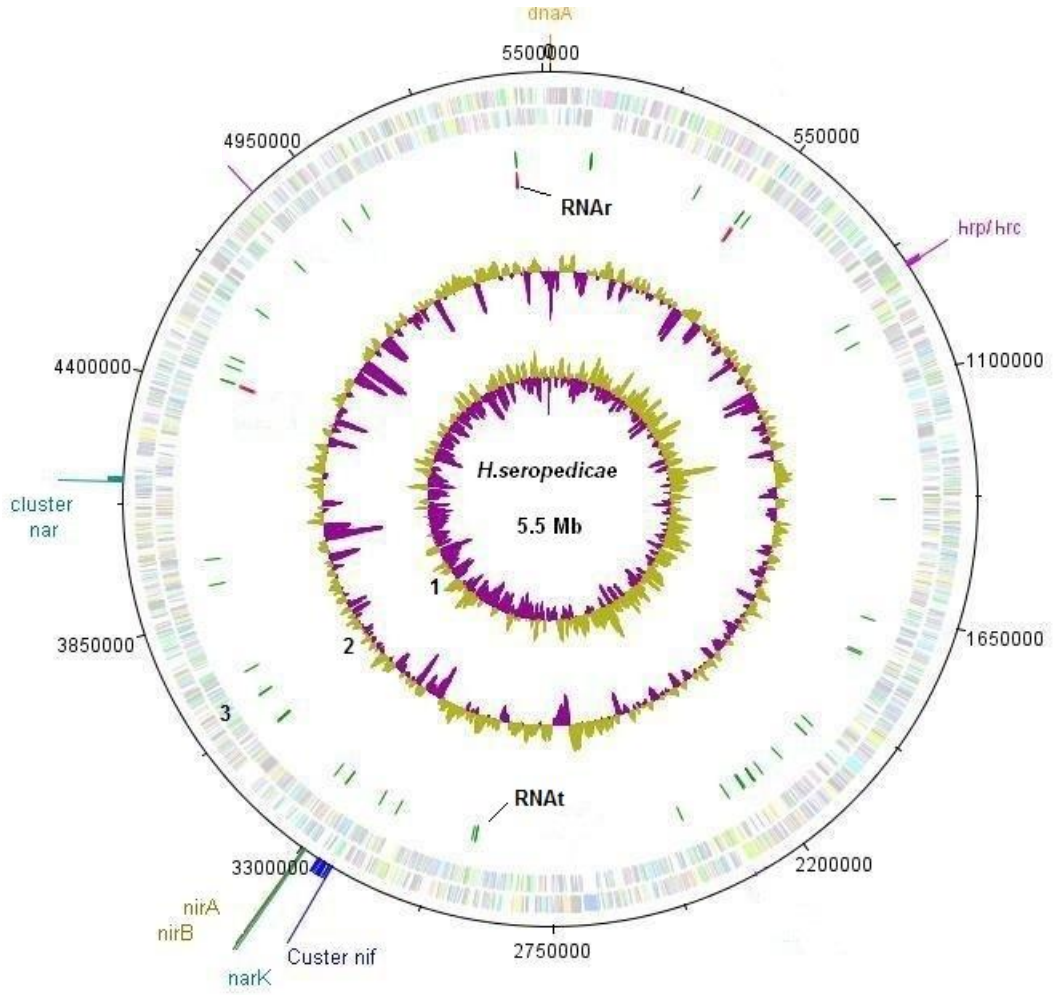


FIGURA 9. Mapa geral do genoma de *H. seropedicae.*

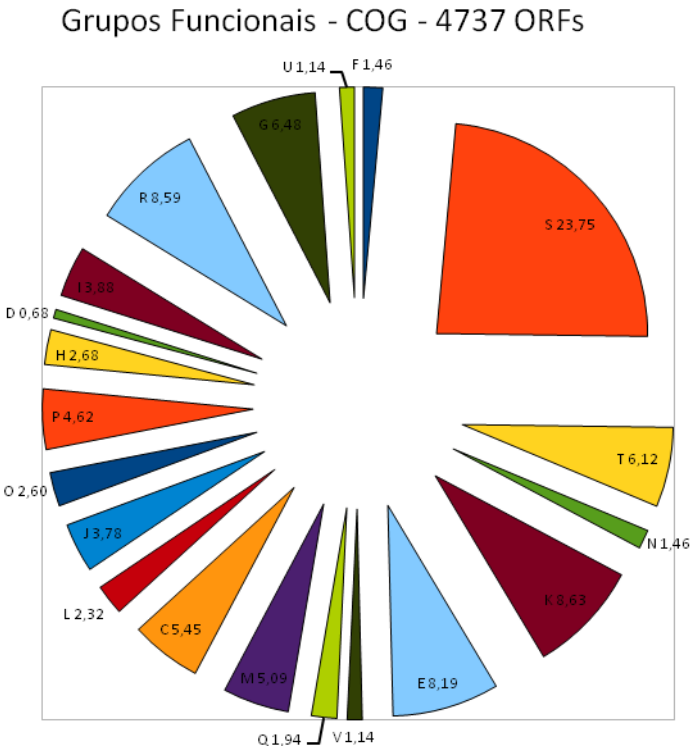
1- Em amarelo GCskew positivo e roxo negativo;

2- Em amarelo conteúdo de GC acima da média do genoma. Em roxo conteúdo GC abaixo da média do genoma;

3- Genes de *H. seropedicae.*

Os tRNAS estão em verde e os rRNAs em vermelho e em destaque na figura.

Também estão em evidência os genes *nif*, responsáveis pela fixação de nitrogênio, genes *nar* e *nir*, responsáveis, respectivamente, pelo metabolismo de nitrato e assimilação de nitrito, e *hrp/hrc*, envolvidos no sistema de secreção tipo III



	Processamento e armazenamento de informação	N°	%
J	Estrutura de tradução ribossomal e biogênese	179	3,78
K	Transcrição	409	8,63
L	Replicação, recombinação e reparo	110	2,32
	Processos celulares e sinalização		
D	Controle do ciclo celular divisão celular e particionamento do cromossomo	32	0,68
V	Mecanismos de Defesa	54	1,14
T	Mecanismo de transdução de sinal	209	6,12
M	Parede celular/membrana	241	5,09
N	Motilidade celular	69	1,46
U	Tráfego intracelular, secreção e transporte vesicular	54	1,14
O	Modificação pós-traducional, renovação de proteína, chaperonas	123	2,60
	Metabolismo		
C	Produção e conversão de energia	258	5,45
G	Transporte de carboidrato e metabolismo	307	6,48
E	Transporte e metabolismo de aminoácido	388	8,19
F	Transporte e metabolismo de nucleotídeo	69	1,46
H	Transporte e metabolismo de coenzimas	127	2,68
I	Transporte e metabolismo de lipídeos	184	3,88
P	Transporte e metabolismo de íons inorgânicos	218	4,62
Q	Biossíntese de metabolitos secundários, transporte e catabolismo	92	1,94
	Não Caracterizados		
R	Função geral predita	407	8,59
S	Função não conhecida	1125	23,75

FIGURA 10. Distribuição das ORF anotadas de *H. seropedicae* nas categorias funcionais COG.

5.2.8 Análise de uso de códon no genoma de *H. seropedicae*

Como resultado da análise de correspondência (COA) aplicada à preferência de uso de códons, realizada com o programa CodonW foi obtido um valor de inércia de 38,2% para o conjunto de genes. Foram identificados no genoma 55 tRNAS e seus anticódons são mostrados na Tabela 12. A partir das 4.737 ORF anotadas para o genoma completo de *H. seropedicae* foi construída uma tabela de preferência de uso de códons (Tabela 13 e na Figura 11). O genoma de *H. seropedicae* contem alto índice de GC e observou-se uma forte tendência ao uso de códons sinônimos contendo estes nucleotídeos na terceira base do códon (Figura 11). A distribuição de uso de aminoácidos no genoma, apresentado na Figura 12, mostra que alguns aminoácidos, como glicina, alanina e valina são usados mais frequentemente nas proteínas codificadas no genoma de *H. seropedicae*.

TABELA 12. tRNAS PRESENTES NO GENOMA DE *H. seropedicae* E SEUS ANTICÓDONS

tRNA	Anticodon	tRNA	Anticodon	tRNA	Anticodon	tRNA	Anticodon	tRNA	Anticodon	tRNA	Anticodon
Tyr	GTA	Ser	CGA	Pro	TGG	Asp	GTC	Ala	GGC	His	GTG
Gly	TCC	Gly	GCC	Arg	TCT	Thr	TGT	Glu	TTC	Leu	CAA
Thr	GGT	Cys	GCA	Leu	CAG	Ile	GAT	Arg	ACG	Lys	TTT
Trp	CCA	Gly	GCC	Thr	CGT	Ala	TGC	Val	TAC	Met	CAT
Ile	GAT	Gly	GCC	Gln	TTG	Val	CAC	Asp	GTC	Gly	CCC
Ala	TGC	Val	GAC	Arg	ACG	Met	CAT	Ser	GCT	Pro	CGG
Met	CAT	Val	GAC	Phe	GAA	Asn	GTT	Leu	TAA		
Ser	GGA	Pro	GGG	Ile	GAT	Asn	GTT	Leu	CAG		
Leu	GAG	Leu	TAG	Ala	TGC	Ala	GGC	Lys	CTT		
Ser	CGA	Ser	TGA	Arg	CCG	Glu	TTC	Lys	CTT		

TABELA 13. PREFERÊNCIA NO USO DE CÓDON PARA O GENOMA DE *H. seropedicae*

Codon	Número	RSCU*	Codon	Número	RSCU*	Codon	Número	RSCU*	Codon	Número	RSCU*
UUU (Phe/F)	8804	0,31	UCU (Ser/S)	3059	0,19	UAU (Tyr/Y)	16193	0,82	UGU (Cys/C)	1983	0,27
UUC (Phe/F)	48322	1,69	UCC (Ser/S)	22072	1,39	UAC (Tyr/Y)	23258	1,18	UGC (Cys/C)	12559	1,73
UUA (Leu/L)	1111	0,04	UCA (Ser/S)	3102	0,2	UAA (Stop)	1010	0,64	UGA (Stop)	3251	2,06
UUG (Leu/L)	8213	0,62	UCG (Ser/S)	27432	1,73	UAG (Stop)	477	0,3	UGG (Trp/W)	21467	1
CUU (Leu/L)	4288	0,15	CCU (Pro/P)	6283	0,32	CAU (His/H)	15779	0,87	CGU (Arg/R)	13858	0,79
CUC (Leu/L)	26888	0,91	CCC (Pro/P)	28455	1,43	CAC (His/H)	20607	1,13	CGC (Arg/R)	71513	4,06
CUA (Leu/L)	2070	0,07	CCA (Pro/P)	4734	0,24	CAA (Gln/Q)	14424	0,39	CGA (Arg/R)	2816	0,16
CUG (Leu/L)	23747	4,41	CCG (Pro/P)	40200	2,02	CAG (Gln/Q)	58750	1,61	CGG (Arg/R)	13028	0,74
AUU (Ile/I)	8666	0,32	ACU (Thr/T)	4329	0,21	AAU (Asn/N)	15711	0,66	AGU (Ser/S)	4615	0,29
AUC (Ile/I)	69798	2,61	ACC (Thr/T)	53339	2,64	AAC (Asn/N)	31603	1,34	AGC (Ser/S)	35030	2,21
AUA (Ile/I)	1677	0,06	ACA (Thr/T)	3178	0,16	AAA (Lys/K)	7204	0,25	AGA (Arg/R)	1447	0,08
AUG (Met/M)	41203	1	ACG (Thr/T)	20101	0,99	AAG (Lys/K)	50462	1,75	AGG (Arg/R)	3091	0,18
GUU (Val/V)	3898	0,13	GCU (Ala/A)	3016	0,26	GAU (Asp/D)	29788	0,71	GGU (Gly/G)	15729	0,49
GUC (Val/V)	41081	1,4	GCC (Ala/A)	14998	2,3	GAC (Asp/D)	53687	1,29	GGC (Gly/G)	93163	2,88
GUA (Val/V)	5131	0,17	GCA (Ala/A)	14307	0,29	GAA (Glu/E)	44165	1,03	GGA (Gly/G)	7747	0,24
GUG (Val/V)	67312	2,29	GCG (Ala/A)	57459	1,15	GAG (Glu/E)	41700	0,97	GGG (Gly/G)	12603	0,39

*Frequencia relativa de códons sinônimos

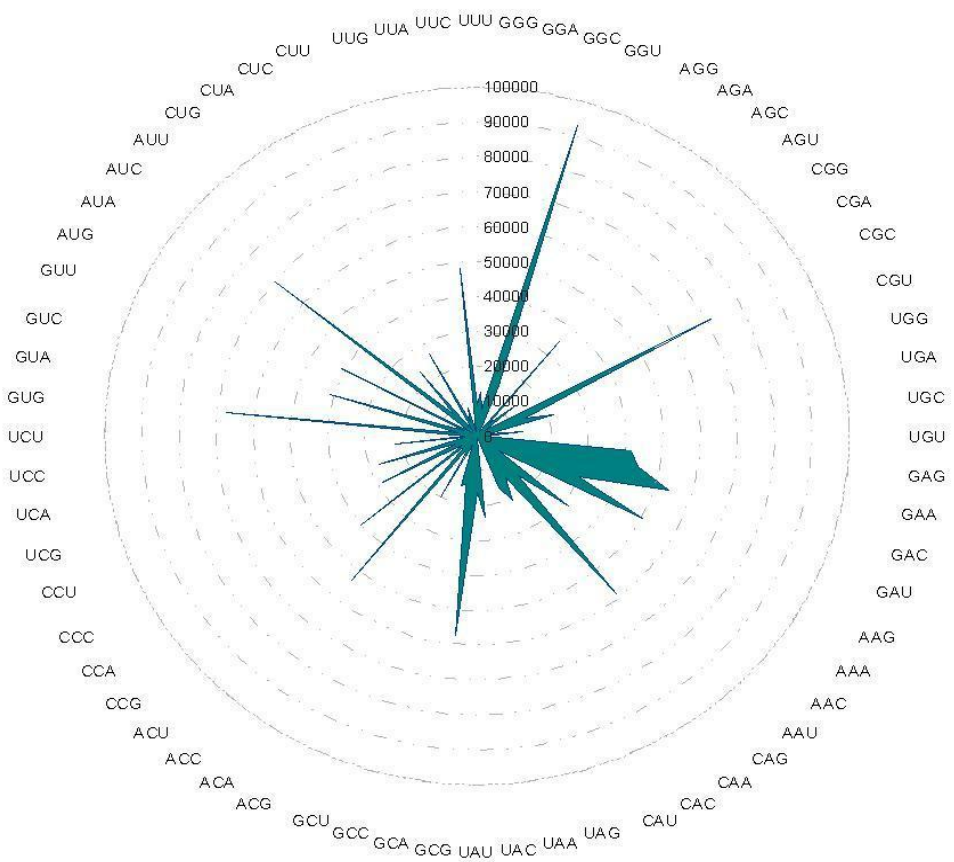
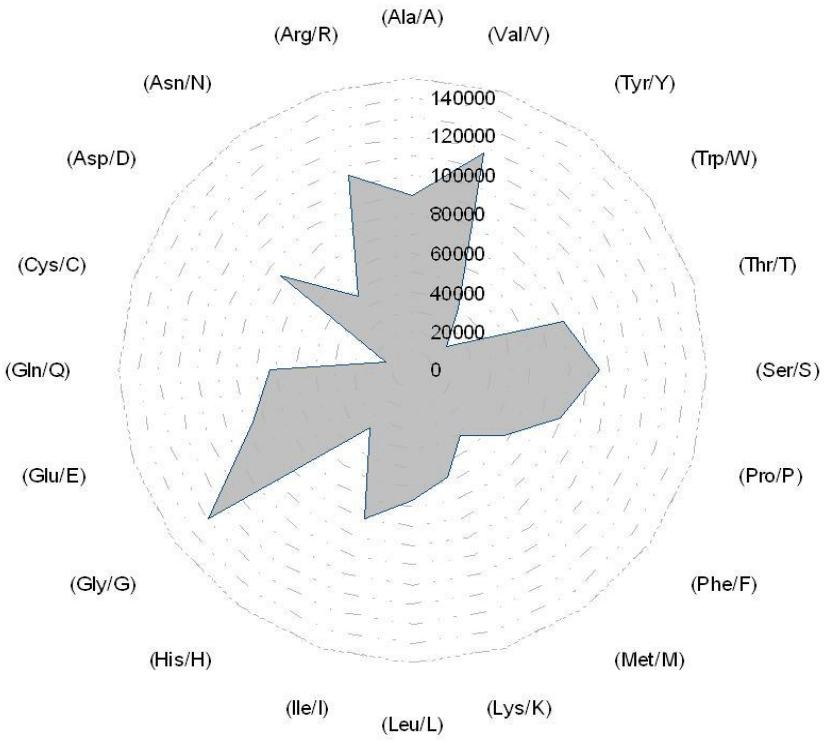


FIGURA 11. Distribuição do uso de códons no genoma de *H.seropedicae.*

Uso de códons correspondente ao resultado obtido para todas as *ORFs* de *H. seropedicae*. A tabela ao lado mostra o número de vezes que cada códon foi usado.

Codon	N°	%
UUU	8804	0,63
UUC	48322	3,43
UUA	1111	0,08
UUG	8213	0,58
CUU	4288	0,30
CUC	26888	1,91
CUA	2070	0,15
CUG	23747	1,69
AUU	8666	0,62
AUC	69798	4,96
AUA	1677	0,12
AUG	41203	2,93
GUU	3898	0,28
GUC	41081	2,92
GUA	5131	0,36
GUG	67312	4,78
UCU	3059	0,22
UCC	22072	1,57
UCA	3102	0,22
UCG	27432	1,95
CCU	6283	0,45
CCC	28455	2,02
CCA	4734	0,34
CCG	40200	2,86
ACU	4329	0,31
ACC	53339	3,79
ACA	3178	0,23
ACG	20101	1,43
GCU	3016	0,21
GCC	14998	1,07
GCA	14307	1,02
GCG	57459	4,08
UAU	16193	1,15
UAC	23258	1,65
UAA	1010	0,07
UAG	477	0,03
CAU	15779	1,12
CAC	20607	1,46
CAA	14424	1,03
CAG	58750	4,18
AAU	15711	1,12
AAC	31603	2,25
AAA	7204	0,51
AAG	50462	3,59
GAU	29788	2,12
GAC	53687	3,82
GAA	44165	3,14
GAG	41700	2,96
UGU	1983	0,14
UGC	12559	0,89
UGA	3251	0,23
UGG	21467	1,53
CGU	13858	0,98
CGC	71513	5,08
CGA	2816	0,20
CGG	13028	0,93
AGU	4615	0,33
AGC	35030	2,49
AGA	1447	0,10
AGG	3091	0,22
GGU	15729	1,12
GGC	93163	6,62
GGA	7747	0,55
GGG	12603	0,90



Aminoácido	Número	%
(Ala/A)	89780	6,40
(Arg/R)	105753	7,54
(Asn/N)	47314	3,37
(Asp/D)	83475	5,95
(Cys/C)	14542	1,04
(Gln/Q)	73174	5,22
(Glu/E)	85865	6,12
(Gly/G)	129242	9,22
(His/H)	36386	2,59
(Ile/I)	80141	5,72
(Leu/L)	66317	4,73
(Lys/K)	57666	4,11
(Met/M)	41203	2,94
(Phe/F)	57126	4,07
(Pro/P)	79672	5,68
(Ser/S)	95310	6,80
(Thr/T)	80947	5,77
(Trp/W)	21467	1,53
(Tyr/Y)	39451	2,81
(Val/V)	117422	8,37

FIGURA 12. Frequência do uso dos aminoácidos e proteínas codificadas pelo genoma de *H. seropedicae.*

5.3 O GENOMA *DE H. rubrisubalbicans*

Devido ao fato do projeto contar com apenas 24.757 leituras de sequências do método Sanger do genoma de *H. rubrisubalbicans,* apenas um mapa parcial do genoma deste organismo foi obtido e os possíveis genes identificados utilizando o genoma anotado de *H. seropedicae* como referência.

As sequências de *H. rubrisubalbicans* foram submetidas ao montador Phrap resultando em 2.703 contigs (Tabela 14). Pelo fato de não haver dados suficientes para o fechamento do genoma, a montagem foi encerrada neste estágio e foi anotada utilizando o genoma de *H. seropedicae* como referência.

TABELA 14. ESTATÍSTICAS PARA A MONTAGEM PARCIAL DA SEQUÊNCIA GENÔMICA DE *H. rubrisubalbicans* (MontagemHR).

	MontagemHR
Número de Bases	3291242 pb
Número de contigs	2703
Maior contig	7223 pb
Menor contig	56 pb
Tamanho Médio dos contigs	1217 pb
Número de *Scaffolds*	165
Leituras de sequências Totais	24757
Leituras de sequências na Montagem	13128
Singlets	7770
Leituras de sequências com vetor	4996
Total de Clones sequenciados	24757
Clones Inconsistentes na montagem	777

5.3.1 Anotação e análise de códons do genoma de *H. rublisubalbicans*

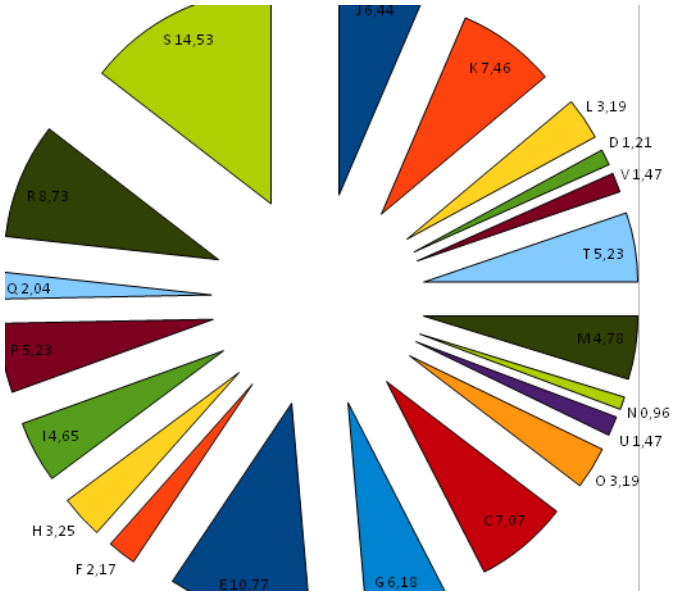
Foram realizadas duas abordagens para *a* anotação parcial do genoma de *H. rubrisubalbicans*. Na primeira foi realizada uma busca por similaridade utilizando o programa BLAST entre os contigs obtidos para a montagem do genoma de *H. rubrisubalbicans* e o genoma de *H. seropedicae*. A comparação foi realizada em nível de nucleotídeos e, desta forma, foi possível identificar as regiões de alta identidade entre os dois genomas. Em seguida foi realizada uma varredura no genoma de *H. seropedicae* e identificado os genes presentes nas regiões similares. Esta estratégia resultou na identificação de 543 genes, o que é pouco, comparado com o tamanho da sequência parcial de *H. rubrisubalbicans*.

Visando a identificação de um número maior de genes, utilizamos as leituras de sequências diretamente ao invés dos contigs. Para a comparação das leituras de

sequências com o genoma de *H. seropedicae* como referência, foi utilizado o programa SSAHA (NING *et al*., 2001), que possibilita o mapeamento das leituras de sequência em um genoma de referencia, considerando a distância entre os pares “b” e “g” do inserto e a qualidade Pherd em seus resultados. Este programa oferece um resultado melhor que o BLAST porque as leituras de sequências só serão ancoradas no genoma de referência se obedecerem a distância estipulada pelo usuário, eliminando assim a presença de pares inconsistentes. Foi estipulada uma distância mínima de 500 pb e máxima de 4.000 pb que corresponde ao tamanho máximo do inserto no plasmídeo. A identificação das ORFs foi realizada levando em consideração as ORFs anotadas nas regiões cobertas pelos pares de sequências “b” e “g” ancorados no genoma de *H. seropedicae*. Foram obtidas regiões de alta similaridade entre os genomas, totalizando 1.569 ORFs identificadas, correspondendo a 33% das ORFs de *H. seropedicae* (Tabela 15). Sequências parciais do genoma de *H. rubrisubalbicans* ficaram distribuídas ao longo de todo o genoma de *H. seropedicae,* mostrando que o sequenciamento aleatório resultou em um panorama geral do genoma de H *rubrisubalbicans*. (Figura 15). As categorias COG das ORF de *H. seropedicae* identificadas no genoma parcial de *H. rubrisubalbicans* são mostradas na Figura 14.

TABELA 15. CARACTERÍSTICAS DO GENOMA PARCIAL DE *H. rubrisubalbicans*

Tamanho do genoma parcial	3291242 pb
Número de contigs	2703
Conteúdo de G+C	61,3%
tRNAS	10
Operons rRNA 16S-23S-5S	1
Genes codificadores para proteínas	1569
Genes com função conhecida	1341
Tamanho médio dos genes	1029 pb



	Processamento e armazenamento de informação	N°	%
J	Estrutura de tradução ribossomal e biogênese	101	6,44
K	Transcrição	117	7,46
L	Replicação, recombinação e reparo	50	3,19
	Processos celulares e sinalização		
D	Controle do ciclo celular divisão celular e particionamento do cromossomo	19	1,21
V	Mecanismos de Defesa	23	1,47
T	Mecanismo de transdução de sinal	82	5,23
M	Parede celular/membrana	75	4,78
N	Motilidade celular	15	0,96
U	Tráfego intracelular, secreção e transporte vesicular	23	1,47
O	Modificação pós-traducional, renovação de proteína, chaperonas	50	3,19
	Metabolismo		
C	Produção e conversão de energia	111	7,07
G	Transporte de carboidrato e metabolismo	97	6,18
E	Transporte e metabolismo de aminoácido	169	10,8
F	Transporte e metabolismo de nucleotídeo	34	2,17
H	Transporte e metabolismo de coenzimas	51	3,25
I	Transporte e metabolismo de lipídeos	73	4,65
P	Transporte e metabolismo de íons inorgânicos	82	5,23
Q	Biossíntese de metabolitos secundários, transporte e catabolismo	32	2,04
	Não Caracterizados		
R	Função geral predita	137	8,73
S	Função não conhecida	228	14,53

FIGURA 13. Distribuição das ORFs identificadas no genoma parcial de *H. rubrisubalbicans* nas categorias funcionais COG.

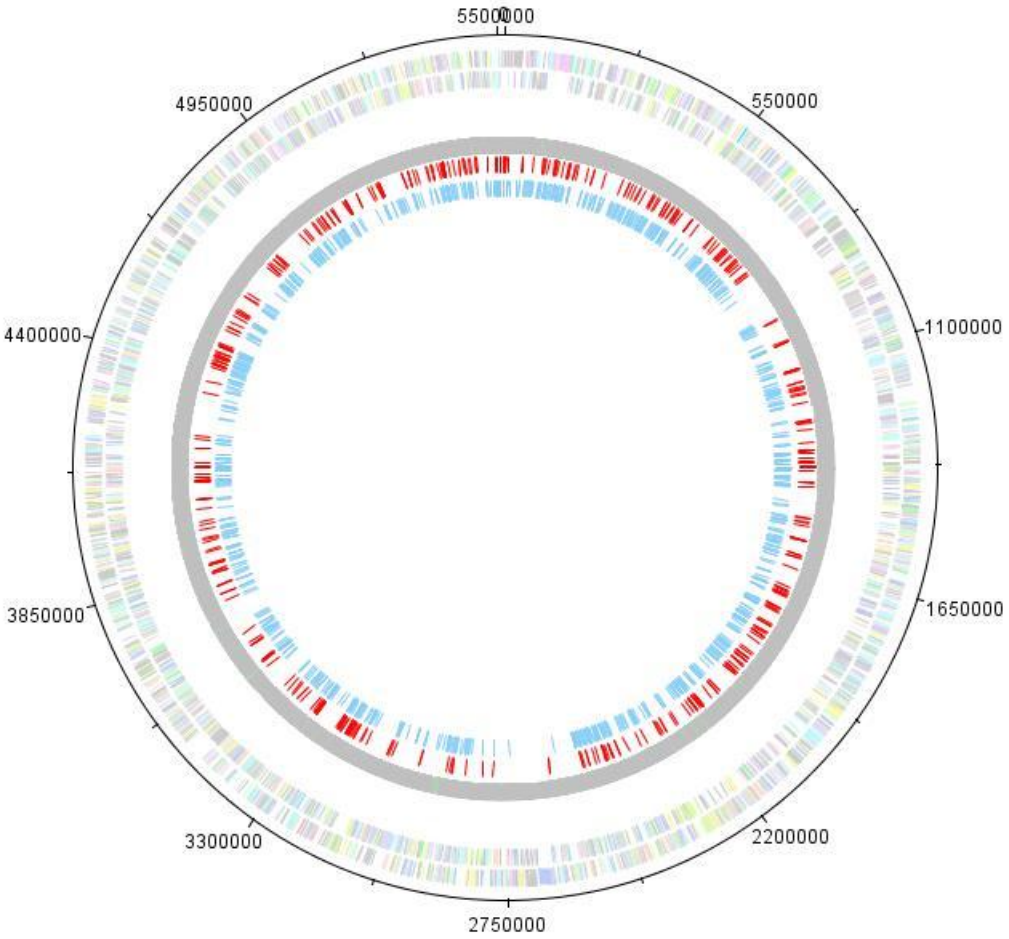


FIGURA 14. Sobreposição do genoma parcial de *H. rubrisubalbicans* no genoma de *H. seropedicae.*

Em azul está representada a comparação realizada pelo programa SSAHA mostrando as regiões onde as leituras de sequência de *H. rubrisubalbicans* ancoraram no mapa do genoma de *H. seropedicae.* Analisando a região foram anotadas 1569 ORFs. Em vermelho esta representada a comparação realizada com os contigs do genoma de *H. rubrisubalbicans* resultando em 543 ORFs anotadas.

6 CONCLUSÕES

H. seropedicae apresenta um genoma circular com 5.513.887pb e conteúdo GC de 63,4%. A anotação do genoma identificou 4737 ORFs, três operons RNA ribossomais (16SrRNA-23SrRNA-5SrRNA), 55 tRNAS. As regiões codificadoras de proteínas perfazem cerca de 88% do genoma.

O padrão de fragmentação do genoma do H. seropedicae in silico foi similar ao obtido experimentalmente e a soma das massas moleculares pelas duas metodologia foram praticamente idênticas. A análise de uso de códons identificou uma forte tendência ao uso GC na terceira base do códon

Glicina, alanina e valina apresentaram as maiores freqüências nas proteínas codificadas pelo genoma de *H. seropedicae.*

O sequenciamento parcial do genoma de *H rubrisubalbicans* resultou em uma montagem com 3.291.242 pb distribuídas em 2703 contigs, com conteúdo GC de 61,3%. A anotação do genoma de *H. rubrisubalbicans* a partir da comparação com o genoma fechado de *H. seropedicae* identificou 1569 ORFs, um operon ribossomal 16SrRNA-23SrRNA-5SrRNA e 10 tRNAS.

REFERÊNCIAS

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W., AND LIPMANL, D. J. Basic local alignment search tool. **J. Mol. Biol.**, p. 403–410, 1990.

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

BALDANI, J.I.; BALDANI V.L.D.; SELDIN, L.; DÖBEREINER, J. Identification and ecology of *Herbaspirillum seropedicae* and the closely related *Pseudomonas rubrisubalbicans*. **Symbiosais**, v.13, p. 65-73, 1992.

BALDANI , J.I.; CARUSO, L.; BALDANI, V.L.D.; GOI, S.R.; DÖBEREINER, J. 1997 Recent advanteages in BNF whith non-legume plants. **Soil Biol. Biochem**., v. 29, p. 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. Emended description of Herbaspirillum; inclusion of [*Pseudomonas*] *rubrisubalbicans*, a mild plant pathogen, as *Herbaspirrillum rubrisubalbicans* comb. nov.; and classification of a group of clinical isolates (EF Group 1) as species 3. **Int. J. Syst. Bacteriol.**, v. 46, p. 802-810, 1996.

BALDANI, V. L. D.; BALDANI, J.I.; DÖBEREINER, J. Inoculation of rice plants with the endophytic diazotrophs *Herbaspirillum seropedicae* and *Burkholderia spp*. **Biol. Fert. Soils**, v. 30, p. 485–491, 2000.

BASTIÁN, F.; COHEN, A.; PICCOLI, P.; LUNA, V.; BARALDI, R.; BOTTINI, R. Production of indole-3-acetic acid and gibberellins A1 and A3 by *Acetobacter diazotrophicus* and *Herbaspirillum seropedicae* in chemically-defined culture media. **Plant Growth Regul.**, v. 24, p. 7-11,1998.

BATZOGLOU S.; JAFFE D. B.; STANLEY K,; BUTLER J.; GNERRE S.; MAUCELI E.; BERGER B.; MESIROV J. P.; LANDER E. S. “ARACHNE: A Whole-Genome Shotgun Assembler. **Genome Res.,** v. 12, p. 177-189, 2002.

BAUM, B. R. PHYLIP: Phylogeny Inference Package. Version 3.2. (Software review). **Quart. Rev. Biol.,** v. 64, p. 539-541, 1989.

BAYAT, A. Science, medicine, and the future Bioinformatics. **BMJ**, v. 324 p. 1018–22, 2002.

BINNEWIES, T. T.; MOTRO, Y.; HALLIN, P. F.; LUND, O.; DUNN, D.; L.A.T.; HAMPSON, D. J.; BELLGARD, M.; WASSENAAR, T. M.; USSERY, D. W. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. **Funct. Integr. Genomics**, v. 6, p. 165-185, 2006.

BIOPERL (http://www.bioperl.org/wiki/Main_Page)

BLATTNER, F. R., PLUNKETT, G., III, BLOCH, C. A., PERNA, N. T., BURLAND, V., RILEY, M., COLLADO-VIDES, J., GLASNER, J. D., RODE, C. K., MAYHEW, G. F.; *et al*. The complete genome sequence of *Escherichia coli K-12*. **Science,** v. 277, p. 1453-1474, 1997.

BOUCK, J.; MILLER, W.; GORRELL, J. H.; MUZNY, D.; GIBBS, R. A. Analysis of the Quality and Utility of Random Shotgun Sequencing at Low Redundancies. **Genome Res.,** v.10, p.1074-84, 1998.

CHAMBERLIN, D. D.; ASTRAHAN, M. M.; BLASGEN, M. W.; GRAY, J. N.; KING, W. F.; LINDSAY, B. G.; LORIE, R.; MEHL, J. W.; PRICE, T. G.; PUTZOLU, F.; SELINGER, P. G.; SCHKOLNICK, M.; SLUTZ, D. R.; TRAIGER, I. L.; WADE, B. W.; YOST, R. A. A history and evaluation of System R. Commun. **ACM**, p. 632-646, 1981.

COCKING E. C. Endophytic colonization of plant roots by nitrogen-fixing bacteria. **Plant and Soil,** v. 252, p. 169–175, 2003.

COLE, S. T.; BROSCH, R.; PARKHILL, J.; GARNIER, T.; CHURCHER, C.; HARRIS, D.GORDON, S. V.; EIGLMEIER, K.; GAS, S.; BARRY, C. E.; *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. **Nature**, v. 393, p. 537-44, 1998.

CRUZ, L.M.; SOUZA, E.M.; WEBER, O.B.; BALDANI, J.I.; DÖBEREINER, J.; PEDROSA, F.O. 2001. 16S ribossomal DNA characterization of nitrogen-fixing bacteria isolated from banana (*Musa spp.*) and pineapple (*Ananas comosus* (L.) Merril). **Appl. Environ. Microbiol.,** v. 67, p. 2375-2379.

EDWARDS, A.; VOSS, H.; RICE, P.; CIVITELLO, A.; STEGEMANN, J.; SCHWAGER, C.; ZIMMERMAN, J.; ERFLE, H.; CASKEY, T.; ANSORGE, W. Automated DNA sequencing of the human HPRT locus. **Genomics,** v. 6, p. 593-608, 1990.

ELBELTAGY, A.; NISHIOKA, K.; SATO, T.; SUZUKI, H.; YE, B.; HAMADA, T.; ISAWA, T.; MITSUI, H.; MINAMISAWA, K. Endophytic colonization and in plant nitrogen fixation by a *Herbaspirillum sp* isolated from wild rice species. **App. Environ. Microbiol.,** v. 67, n. 11, p. 5284-5293, 2001.

EWING, B.; HILLIER, L.; WENDL, M.; GREEN, P. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. **Genome Res.,** v. 8, p. 175-185, 1998.

FLEISCHMANN, R.D.; ADAMS M.D.; WHITE, O.; CLAYTON, R.A.; KIRKNESS, E.F.; KERLAVAGE, A.R.; BULT, C.J.; TOMB, J.F.; DOUGHERTY, B.A.; MERRICK, J.M.; et al. Whole-genome random sequencing and assembly of *Haemophilus influenza Rd***. Science,** v. 269, n. 5223 p. 496-512.1995.

GARDNER, M. J.; HALL, N.; FUNG, E.; WHITE, O.; BERRIMAN, M.; HYMAN, R.

W.; CARLTON, J. M. ; PAIN, A. ; NELSON, K. E. ; BOWMAN, S. ; *et al*. Genome sequence of the human malaria parasite *Plasmodium falciparum*. **Nature**, v. 419, p. 498-511, 2002a.

GARDNER, M. J.; SHALLOM, S. J.; CARLTON, J. M.; SALZBERG, S. L.; NENE, V.; SHOAIBI, A.; CIECKO, A.; LYNN, J.; RIZZO, M.; WEAVER, B.; *et al*. Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14. **Nature,** v. 419, p. 531-534, 2002b.

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

GORDON, D.; ABAJIAN, C.; GREEN, P. Consed: a graphical tool for sequence finishing**. Genome Res.,** v. 8, p. 195-202, 1998.

GREGORY, S. G.; BARLOW, K. F.; MCLAY, K. E.; KAUL, R.; SWARBRECK, D.; DUNHAM, A.; SCOTT, C. E.; HOWE, K. L.; WOODFINE, K.; SPENCER, C. C. A.; *et al.* The DNA sequence and biological annotation of human chromosome 1. **Nature**, v. 441, 2006.

HALL, N.; PAIN, A.; BERRIMAN, M.; CHURCHER, C.; HARRIS, B.; HARRIS, D.; MUNGALL, K.; BOWMAN, S.; ATKIN, R.; BAKER, S.; *et al*. Sequence of *Plasmodium falciparum* chromosomes 1, 3-9 and 13. **Nature,** v. 419, p. 527-531, 2002.

HALL, N. Advanced sequencing technologies and their wider impact in microbiology. **J. Exp. Biol.,** v. 210, p. 1518-1525, 2007.

HIGGINS, D.G.; SHARP, P.M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. **Gene**, v. 73, p. 237-244, 1988.

HUANG, X.; WANG, J.; ALURU, S.; YANG, S.P.; HILLIER, L. PCAP: a whole-genome assembly program. **Genome Res.**, v. 13, p. 2164–2170, 2003.

HUGHEY, R.; KARPLUS, K. Bioinformatics:a new field in engineering education. **J. Engineering Educ.,** p.101–104, 2003.

HYMAN, R. W.; FUNG, E.; CONWAY, A.; KURDI, O.; MAO, J.; MIRANDA, M.; NAKAO, B.; ROWLEY, D.; TAMAKI, T.; WANG, F.; *et al*. Sequence of *Plasmodium falciparum* chromosome 12. **Nature,** v. 419, p. 534-537, 2002

JAMES, E. K.; OLIVARES, F. L.; BALDANI, J. I.; DÖBEREINER, J. *Herbaspirillum*, an endophytic diazotroph colonizing vascular tissue in the leaves of Sorghum bicolor L. Moench. **J. Exp. Bot.,** v. 48, p. 785-797, 1997.

JAMES, E. K.; GYANESHWAR, P.; MATHAN, N.; BARRAQUIO, W. L.; REDDY, P. M.; IANNETTA, P. P.; OLIVARES, F. L.; LADHA, J. K. Infection and colonization of rice seedlings by the plant growth-promoting bacterium *Herbaspirillum seropedicae* Z67. **Mol. Plant-Microbe Interact.**, v. 15, p. 894-906, 2002.

KLENK, H. P., CLAYTON, R. A., TOMB, J. F., WHITE, O., NELSON, K. E., KETCHUM, K. A., DODSON, R. J., GWINN, M., HICKEY, E. K., PETERSON, J. D. *et al*. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. **Nature**, v. 390, p. 364-370, 1997.

KUMAR, S.; NEI M.; DUDLEY J.; TAMURA K. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences**. Brief Bioinform.**, v. 11 p. 299-306, 2008.

KURTZ, S.; PHILLIPPY, A.; DELCHER, A.L.; SMOOT, M.; SHUMWAY, M.; ANTONESCU C.; SALZBERG S. L. Versatile and open software for comparing large genomes. **Genome Biol.**, v. 5:R12, 2004.

LANDER, E. S.; LINTON, L. M.; BIRREN, B.; NUSBAUM, C.; ZODY, M. C.; BALDWIN, J.; DEVON, K.; DEWAR, K.; DOYLE, M.; FITZHUGH, W.; *et al*. Initial sequencing and analysis of the human genome. **Nature,** v. 409, p. 860-921, 2001.

LUSCOMBE, N. M.; GREENBAUM, D.; GERSTEIN, M. What is bioinformatics? An introduction and overview, Yearbook of Medical Informatics, 2001.

MARDIS, E. R. The impact of next-generation sequencing technology on genetics. **Trends in Genetics**, v. 24, p. 133-141, 2008.

MARGULIES, M.; EGHOLM, M.; ALTMAN, W. E.; ATTIYA, S.; BADER, J. S.; BEMBEN, L. A.; BERKA, J.; BRAVERMAN, M. S.; CHEN, Y. J.; CHEN, Z.; *et al*. Genome sequencing micro fabricated high-density picolitre reactions. **Nature,** v.437, p. 326-327, 2005.

MCLNERNEY, J.O. GCUA: General codon usage analysis. **Bioinformatics**, v. 14, n. 4, p. 372-373, 1998.

MIKKELSEN, T.; HILLER, L. W.; EICHLER, E. E.; ZODY, M. C.; JAFFE, D. B.; YANG, S. P.; ENARD, W.; HELLMANNM, I.; LINBALD-TOH, K.; ALTHEIDE, T. K. Initial sequence of the chimpanzee genome and comparison with the human genome. **Nature,** v. 437, p. 69-87, 2005.

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

MOROZOVA, O.; MARRA, M.A. Applications of next-generation sequencing technologies in functional genomics. **Genomics**, v. 92, p.255-264, 2008.

MYERS, E .W.; SUTTON, G. G.; DELCHER, A.L., DEW, I.M.; FASULO, D.P.; FLANIGAN, M.J.; KRAVITZ, S,A.; MOBARRY, C,M.; REINERT, K.H.; REMINGTON, K.A.; ANSON, E.L.; BOLANOS, R.A.; CHOU, H.H.; JORDAN, C.M.; HALPERN, A.L.; LONARDI, S.; BEASLEY, E.M.; BRANDON, R.C.; CHEN, L.; DUNN, P.J.; LAI, Z.; LIANG, Y.; NUSSKERN, DR.; ZHAN, M.; ZHANG, Q.; ZHENG, X.; RUBIN, G.M.; ADAMS, M.D.; VENTER, J.C. A Whole-Genome Assembly of *Drosophila*. **Science**, v.

287, n. 5461, 2196-2204, 2000.

NIERMAN, C. W.; EISEN, J. A.; Fleischmann, R. D.; Fraser, C. M. Genome data: what do we learn? **Structural Biol.**, v.10, p. 343-348, 2000.

NING, Z.; COX, A. J.; MULLIKIN, J. C. SSAHA: a fast search method for large DNA databases. **Genome Res.,** v.11 (10), p. 1725-9, 2001.

OLIVEIRA, A. L. M.; URQUIAGA, S.; DÖBEREINER, J.; BALDANI, J. I. The effect of inoculating endophytic N2-fixing bacteria on micropropageted sugarcane plants. **Plant and Soil,** v. 242, p. 205–215, 2002.

PEARSON, W. R. Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. **Genomics,** v.11, p. 635-650, 1991.

PENDEN, J.F. Analysis of codon usage. **Tese de Doutorado**. Universidade de Nottingham, Inglaterra, 1999.

PROBER, J. M; TRAINOR, G. L; DAM, R. J; HOBBS, F. W; ROBERTSON, C. W; ZAGURSKY, R. J; COCUZZA, A. J; JENSEN, M. A; AND BAUMEISTER, K. A system for rapid DNA sequencing with fluorescent chainterminating dideoxynucleotides. **Science,** v.238, p. 336-341, 1987.

RAMOS, J. R. L. S. Análises moleculares comparativas de estirpes de *Herbaspirillum* por PFGE, RAPD, RFLP e sequenciamento do gene que codifica para 16s rRNA. **Tese de Doutorado**. Universidade Federal do Paraná, 2003.

RONCATO-MACCARI, L. D.; RAMOS, H. J. O.; ALQUINI, Y.; CHUBATSU, L. S.; YATES, M. G.; RIGO, L. U.; STEFFENS, M. B. R.; SOUZA, E. M. Endophytic *Herbaspirillum seropedicae* expresses *nif* gene in gramineous plants. **FEMS Microbiol. Ecol.,** v. 45, p. 39-47, 2003.

RUTHERFORD, K.; PARKHILL, J.; CROOK, J.; HORSNELL, T.; RICE, P.; RAJANDREAM, M. A.; BARRELL, B. Artemis: sequence visualization and annotation **Bioinformatics,** v.16, n.10, p. 944-945, 2000.

SALZBERG, S. L.; DELCHER, A. L.; KASIF, S.; WHITE, O. Microbial gene identification using interpolated Markov models. **Nucleic Acids Res.**, v. 26, n. 2, p. 544-548, 1998.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors (DNA polymerase/nucleotide sequences/bacteriophage 4X174). **Biochemistry**, v. 74, n. 12, p. 5463-5467, 1977.

SMITH, L. M.; SANDERS, R. J . K.; HUGHES, P.; DOOD, C.; CONNEL, C. R.; HEINER, C.; KENT, S. B. H.; HOOD, L. E. Fluorescence detection in automated DNA sequence analysis**.** **Nature**, v. 321, p. 674 – 679, 1986.

SUZEK, B. E.; ERMOLAEVA, M. D.; SCHREIBER, M. SALZBERG S. L. A

probabilistic method for identifying start codons in bacterial genomes, **Bioinformatics**, v. 17, p. 1123-1130, 2001.

TATUSOV, R. L.; GALPERIN, M. Y.; NATALE, D. A.; AND KOONIN, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. **Nucleic Acids Res.**, v. 28, p. 33–36, 2000.

THOMPSON J.D.; HIGGINS D.G.; GIBSON, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position- specific gap penalties and weight matrix choice. **Nucleic Acids Res.**, v. 22, n. 22, p. 4673-4680, 1994.

YOU, M.; NISHIGUCHI, T.; SAITO, A.; ISAWA, T.; MITSUI, H.; MINAMISAWA, K. Expression of the nifH gene of a *Herbaspirillum* endophyte in wild rice species: daily rhythm during the light-dark cycle. **Appl. Envir. Microbiol**, v. 71, p.8183-8190, 2005.

ANEXO I

Instruções para visualização da comparação das montagens de *H. seropediace*

Os resultados das análises com o BLAST estão separados em diretórios distintos no CD que acompanha a dissertação.

Para visualização da tabela com os dados de alinhamento basta acessar o diretório escolhido e abrir o arquivo index.php.

Este arquivo abrirá o navegador de sua preferência e mostrará uma tabela.

Além das informações dos alinhamentos existem duas colunas com links.

Na columa PI, o link mostrará o resultado do BLAST.

Na coluna Graf, o link acessará o gráfico de cobertura entre os alinhamentos

No campo Estatísticas são mostrados os alinhamentos com scores baixos, como alinhamentos repetitivos

No Campo Queries com hits nulos, são mostradas as sequências que não obtiveram alinhamento.

Diretório v2Xv1 - MontagemV2xMontagemV1
Diretório v3Xv1 - MontagemV3xMontagemV1
Diretório v4Xv1 - MontagemV4xMontagemV1

Universidade Federal do Paraná
Sistema de Bibliotecas

Weiss, Vinicius Almir
Estratégias de finalização da montagem do genoma da bactéria diazotrófica endofítica *Herbaspirillum seropedicae* SmR1. / Vinicius Almir Weiss. – Curitiba, 2010.
70 f.: il. ; 30cm.

Orientador: Leonardo Magalhães Cruz
Co-orientador: Roberto Tadeu Raittz

Dissertação (mestrado) - Universidade Federal do Paraná, Setor de Ciências Biológicas. Programa de Pós-Graduação em Bioquímica.

1. Bacterias nitrificantes 2. DNA 3. Mapeamento cromossômico I. Título II. Cruz, Leonardo Magalhães III. Raittz, Roberto Tadeu IV. Universidade Federal do Paraná. Setor de Ciências Biológicas. Programa de Pós-Graduação em Bioquímica.

CDD (20. ed.) 589.9