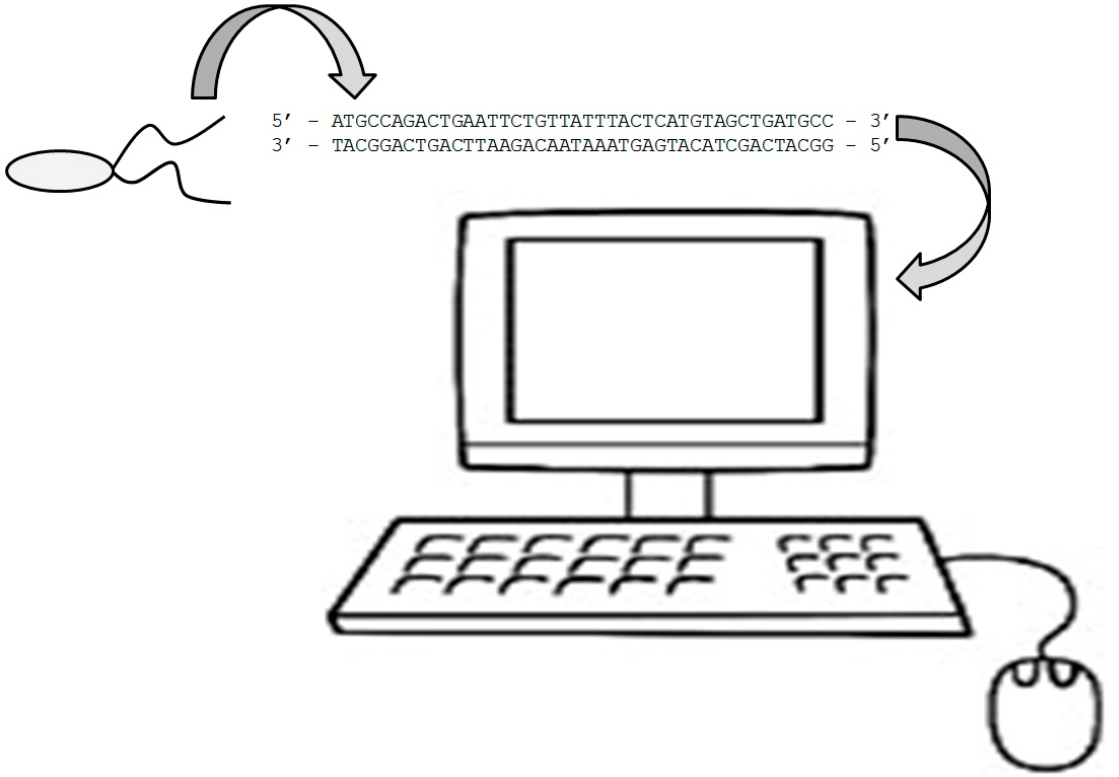
Documentos

ISSN 1517-8498
Dezembro, 2011

282

Base de dados genômica de estirpes que compõem o inoculante de cana-de-açúcar e milho



Embrapa

ISSN 1517-8498
Dezembro, 2011

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Agrobiologia
Ministério da Agricultura, Pecuária e Abastecimento

Documentos 282

Base de dados genômica de estirpes que compõem o inoculante de cana-de-açúcar e milho

José Ivo Baldani
Helma Ventura Guedes
Márcia Soares Vidal
Stefan Schwab
Kátia Regina dos Santos Teixeira
Leonardo Magalhães Cruz
Jean Luiz Simões de Araújo

Embrapa Agrobiologia
Seropédica, RJ
2011

Exemplares desta publicação podem ser adquiridos na:

Embrapa Agrobiologia
BR 465, km 7, CEP 23.851-970, Seropédica, RJ
Caixa Postal 74505
Fone: (21) 3441-1500
Fax: (21) 2682-1230
Home page: www.cnpab.embrapa.br
E-mail: sac@cnpab.embrapa.br

Supervisora editorial: Norma Gouvêa Rumjanek
Normalização bibliográfica: Carmelita do Espírito Santo
Tratamento de ilustrações: Maria Christine Saraiva Barbosa
Editoração eletrônica: Marta Maria Gonçalves Bahia
Foto da capa: Márcia Soares Vidal

1ª edição
1ª impressão (2011): 50 exemplares

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Agrobiologia

BASE DE DADOS genômica de estirpes que compõem o inoculante de cana-de-açúcar e milho. / José Ivo Baldani et al. Seropédica: Embrapa Agrobiologia, 2011. 32 p. Embrapa Agrobiologia. Documentos 282).

ISSN: 1517-8498

1. Informação genômica. 2. Bactéria diazotrófica. 3. Bioinformática. 4. Sequenciamento genético. I. Baldani, José Ivo. II. Guedes, Helma Ventura. III. Vidal, Márcia Soares. IV. Schwab, Stefan. V. Teixeira, Kátia Regina dos Santos. VI. Cruz, Leonardo Magalhães. VII. Araújo, Jean Luis Simões. VIII. Série. IX. Embrapa Agrobiologia.

CDD 632.32

Autores

José Ivo Baldani
Márcia Soares Vidal
Stefan Schwab
Kátia Regina dos Santos Teixeira
Jean Luiz Simões de Araújo
Pesquisadores da Embrapa Agrobiologia.
BR 465, km 7, CEP 23890-000 - Seropédica, RJ.
E-mails: ivo.baldani@embrapa.br, marcia.vidal@embrapa.br, stefan.schwab@embrapa.br, katia.teixeira@embrapa.br, jean.araujo@embrapa.br

Helma Ventura Guedes
Analista da Embrapa Cerrados. BR 020
Km 18, Planaltina, DF, CEP 73310-970.
E-mail: helma.guedes@embrapa.br

Leonardo Magalhães Cruz
Professor da Universidade Federal do Paraná.
Departamento de Bioquímica e Biologia Molecular
Caixa Postal 19046. Curitiba, PR, CEP 81531-980.
E-mail: leonardo@ufpr.br

Apresentação

As atitudes de usar com responsabilidade os recursos naturais (solo, água, ar, flora, fauna, energia), de preservar e conservar a natureza são cada vez mais necessárias para a sociedade moderna acarretando em uma busca constante por sistemas de produção agropecuários apoiados em princípios ecológicos e naturais.

Dentro desse cenário, a Embrapa Agrobiologia construiu o seu atual plano diretor de pesquisa, desenvolvimento e inovação com a seguinte missão "gerar conhecimentos e viabilizar tecnologias e inovação apoiados nos processos agrobiológicos, em benefício de uma agricultura sustentável para a sociedade brasileira".

A série documentos se constitui em uma linha de publicações que visa disponibilizar informações relevantes das mais diversas etapas dos processos de pesquisa científica e tecnológica. Podem disponibilizar revisões de literatura sobre temas relevantes, relatórios técnicos, um determinado procedimento metodológico, levantamentos de campo, entre outros tipos de conteúdo.

A presente publicação intitulada "Base de Dados genômica de estirpes que compõem o inoculante de cana-de-açúcar e milho" tem indicação

para todos aqueles interessados em conhecer mais sobre o assunto, portanto, boa leitura.

Eduardo Francia Carneiro Campello
Chefe Geral da Embrapa Agrobiologia

Sumário

Base de dados genômica de estirpes que compõem o inoculante de cana-de-açúcar e milho

José Ivo Baldani
Helma Ventura Guedes
Márcia Soares Vidal
Stefan Schwab
Kátia Regina dos Santos Teixeira
Leonardo Magalhães Cruz
Jean Luiz Simões de Araújo

Introdução

De acordo com Prosdocimi et al. (2002) um banco de dados pode ser considerado "uma coleção de dados inter-relacionados, projetado para suprir as necessidades de um grupo específico de aplicações e usuários". Sua função é estruturar e organizar as informações de forma que facilite consultas e atualizações de dados. Segundo Napoli (2003) a construção de bancos de dados para armazenamento de informações de sequências de DNA, proteínas e suas estruturas tridimensionais, bem como vários outros produtos da era genômica, tem sido extremamente importante e um grande desafio, principalmente devido à imensa quantidade de dados gerados em inúmeros laboratórios de todo o mundo.

Um dos principais banco de dados sobre informações genômicas é o *GenBank* do Centro Nacional para Informação Biotecnológica dos EUA (NCBI), disponível em: <www.ncbi.nlm.nih.gov>. Outros bancos de dados distribuídos por países da Europa e Japão funcionam de forma integrada com o NCBI, mantendo a constante troca de informações e dados, como fruto da parceria estabelecida com o *European Nucleotide Archive* (ENA) e com *DNA Data Bank of Japan* (DDBJ) (BENSON et al., 2011). Além desses parceiros, o Escritório Americano de Patentes e Marcas também contribui na alimentação do banco de dados, com sequências proveniente das patentes emitidas (BENSON et al., 2011).

O *GenBank* é acessado a partir do NCBI Entrez, sistema que integra dados de sequencias de DNA, RNA e proteínas relacionando-os juntamente com a taxonomia, genomas, estrutura e domínio de proteínas, bem como com a literatura através do Pubmed (BENSON et al., 2011). O *GenBank* reúne e organiza as informações em diferentes divisões, seja com base na taxonomia ou até mesmo na estratégia de sequenciamento utilizada para obtenção dos dados (BENSON et al., 2011). Em 2011, o *GenBank* continha mais de 1.200 genomas completos de bactérias e archea,(BENSON et al., 2011). De acordo com o Banco de Dados GOLD - Genomes OnLine Database (http://www.genomesonline.org/), atualmente existem 2572 projetos genomas completos, sendo 153 de arqueobactérias, 2268 de bactérias e 151 de eucariotos (GOLD, 2013). Além do *GenBank* outros bancos de dados atendem as necessidades de consultas específicas, como por exemplo: o Pfam (PUNTA et al., 2012), para comparação entre família de proteínas, o SwissProt (BAIROCH e APWEILER, 1996), que reúne informações como modificações pós-traducionais e domínios de proteínas, o KEGG (KANEHISA et al., 2004), para comparação de vias metabólicas, entre outros.

A ferramenta mais utilizada para comparação de sequências de DNA com os bancos de dados genômicos é o BLAST ou *Basic Local Alignment Search Tool* (ALTSCHUL, 1990). Através do algoritmo utilizado pelo BLAST, que agiliza o alinhamento de sequência local, podemos comparar uma sequência de DNA ou proteína qualquer com todas as sequências de um banco de dados específico. De acordo com Napoli (2003) o programa BLAST não conduz a uma comparação da extensão total das sequências, mas procura identificar, no banco de dados, a presença de uma pequena sequência similar com a sequência desconhecida de interesse ou "pergunta" (*Query*), descartando os resultados não produtivos. A partir da identificação desse pequeno fragmento similar, a comparação é então estendida para a vizinhança da região de similaridade identificada, aumentando o tamanho da região avaliada e possibilitando a comparação da sequência desconhecida com as sequências presentes no banco de dados. Essa estratégia aumenta sobremaneira a velocidade de análise do BLAST. O resultado desta busca tem como retorno aquelas sequências

(DNA ou proteínas) com maior similaridade presentes no banco de dados (*Subject*). Para facilitar a análise pode-se, por exemplo, utilizar na busca apenas uma das várias subdivisões do *GenBank*, e também limitar a pesquisa a sequências de um dado organismo de interesse. Através da análise de similaridade realizada no BLAST, várias regiões do DNA ou da proteína deduzida podem ser identificadas, o que auxilia na definição da provável função de qualquer segmento de DNA que apresente homologia significativa a outras sequências de DNA ou proteínas com função estabelecida e sequência previamente depositada no *GenBank* (SANTOS e ORTEGA, 2003). Para análise de similaridade, programas como o BLAST utilizam algoritmos e matrizes de pontuação que quantificam a similaridade entre sequências a partir de uma tabela de valores que descreve a probabilidade de um determinado aminoácido ou base de DNA ocorrer no alinhamento entre duas sequências. Na descrição das comparações de sequência, vários termos são bastante usados. De acordo com Napoli (2003): "A homologia de sequência é um termo mais geral indicando a relação evolutiva entre as sequências. Duas sequências são consideradas homólogas se derivarem da mesma sequência ancestral. *Similaridade* refere-se à presença de locais similares e idênticos em duas sequências, enquanto homologia indica que duas sequências possuem uma probabilidade alta de compartilharem o mesmo ancestral" (NAPOLI, 2003).

Na interpretação dos resultados do BLAST são avaliados os seguintes parâmetros: Valor-E (*E-value*) que corresponde à probabilidade do alinhamento ocorrer ao acaso; Identidade ou *Identities* que é o número de pareamentos idênticos entre a sequência desconhecida de interesse (*query*) ou "pergunta" e a sequência do banco de dados; Positivos ou *Positives* que referem-se a porcentagem de pareamento de semelhança química, no caso da comparação entre duas proteínas, representado pelo os sinais "+" (ALTSCHUL 1990; 1991; FORMIGHIERI, 2005). A pontuação (*Score*) corresponde à soma dos valores atribuídos a partir da matriz de pontuação para cada pareamento idêntico ou similar entre a sequência desconhecida de interesse ou pergunta e sequência presente no banco de dados (FORMIGHIERI, 2005).

Destaca-se que os parâmetros citados não devem ser analisados individualmente (ALCALDE, 2002). O principal parâmetro de análise é o *E-value*, valor estatístico da probabilidade que o alinhamento ocorra ao acaso (ALTSCHUL, 1991). Quanto menor o *E-value*, menor a chance de que o alinhamento tenha acontecido por acaso e consequentemente, mais semelhantes são as sequências alinhadas. Conforme exposto por Alcalde (2002) "o valor da similaridade de duas sequências é determinado pela presença de uma região alinhada de aminoácidos idênticos ou substituições conservadas, que recebem uma pontuação positiva, e de substituições não conservadas, que recebem uma pontuação negativa. Quando não existe similaridade numa região extensa o algoritmo cria uma lacuna (*gap*) em uma das sequências com o objetivo de encontrar o melhor alinhamento. Tanto para a criação de lacunas quanto para seu prolongamento, o algoritmo estabelece uma penalidade. Então a similaridade entre as duas sequências é determinada pela soma total das notas obtidas para cada um dos aminoácidos nas regiões alinhadas, descontadas as penalidades de abertura e prolongamento das lacunas" (ALCADE, 2002).

Há várias modalidades de BLAST, sendo o tBLASTx, de grande utilidade nos estudos de genômica e anotação. Nesta ferramenta, tanto a sequência de interesse ou pergunta submetida como a sequência comparada da base de dados (*Subject*) são de nucleotídeos; no entanto, antes de verificar a similaridade são geradas as seis sequências de aminoácidos possíveis a partir de uma dada sequência de nucleotídeos, ou seja, tanto a sequência de interesse ou pergunta quanto cada uma das presentes na base de dados. Cada uma dessas sequências de nucleotídeos decodificadas como proteínas correspondem a cada uma das três fases de leituras possíveis em cada uma das fitas do DNA a partir do nucleotídeo 1, 2 e 3 na direção 5' - 3', porém apenas uma das seis leituras possui significado biológico e os demais resultados são desprezados (SANTOS e ORTEGA, 2003). O tBLASTx permite uma comparação entre as proteínas deduzidas a partir de uma sequência de DNA, tornando análise mais robusta, uma vez que as proteínas de dois organismos podem ser mais parecidas entre si do que os

nucleotídeos que as codificam (SANTOS e ORTEGA, 2003). Além do tBLASTx, o BLAST possui outras modalidades, como o BLASTn, que busca similaridade entre sequências de nucleotídeos, BLASTp que busca similaridade entre sequências de proteínas e BLASTx que busca similaridade entre sequências de nucleotídeos e proteínas (ALTSCHUL, 1990, 1991; ALTSCHUL et al.,1997). A definição da modalidade de BLAST utilizada depende do tipo de sequência de interesse ou pergunta e do banco que será utilizado para a comparação.

Utilizando-se diferentes ferramentas da bioinformática, uma grande variedade de informações existentes nos bancos de dados pode ser acessada e analisada com múltiplas finalidades como, por exemplo, em pesquisas comparativas para estudos funcionais de genoma, identificação de genes relacionados a doenças e estudos evolutivos (SANTOS e ORTEGA, 2003). Foi demonstrado por genômica comparativa que entre os procariotos, na história evolutiva, vários segmentos de DNA foram trocados entre distintas espécies, num processo de transferência horizontal bastante intenso (SANTOS e ORTEGA, 2003).

Outro banco de dados de grande importância é o *Kyoto Encyclopedia of Genes and Genomes* - KEGG (KANEHISA et al., 2004) que surgiu como parte dos projetos de pesquisa dos laboratórios do Centro de Bioinformática da Universidade de Kyoto e do Centro de genoma humano da Universidade de Tokyo. O KEGG, disponível em <http://www.genome.ad.jp/kegg/> é um recurso da base de dados para compreender as funções da informação genômica (genes e proteínas) aplicadas a sistemas biológicos, tais como a célula ou organismos. Ele disponibiliza uma representação gráfica do sistema biológico, consistindo de fluxogramas que interrelacionam genes, funções e moléculas que constituem as vias metabólicas, permitindo uma visão completa do papel dos genes e proteínas presentes em um genoma com os processos celulares. Segundo Kanehisa et al. (2004) este sistema eventualmente poderá se tornar uma representação computacional em forma gráfica, não só de células e organismos, mas também da biosfera, permitindo a análise in silico de diferentes sistemas biológicos.

Anotação genômica e predição de genes

A partir da obtenção da sequência de DNA de um genoma ou parte dele é necessário identificar os genes e atribuir a cada um deles uma função biológica através da análise *in silico*, processo conhecido como anotação genômica (http://www.ncbi.nlm.nih.gov/genome/guide/build.shtml, acessado em 04/04/2012). Para a predição de genes, ou seja, para se fazer a busca por ORFs (*Open Reading Frames*), que correspondem à sequências com possíveis regiões codificadoras identificadas por um códon iniciador e um terminador, diferentes programas (Ex: ORF Finder, GenScan) utilizam estratégias como a identificação de sequências no DNA que possam funcionar como região promotoras seguidas por sequências que possam gerar proteínas, moléculas de RNA transportador, RNA ribossômico, ou que tenham similaridade com genes/proteínas hipotéticos ou desconhecidos, regiões repetitivas (fagos, transposon, etc), regiões regulatórias (peptídeo sinal, "enhancer", "clustered, regularly interspaced short palindromic repeats" - CRISPRs, etc) (SANTOS e ORTEGA, 2003; PRODOSCIMI et al., 2003; WHEELER et al., 2003, GRISSA et al., 2010).

Dessa forma, após a identificação dos genes é possível associá-los à prováveis funções e determinar as regiões codificadoras e não-codificadoras, construindo assim um mapa genômico estrutural e funcional do organismo. Nessa fase é necessária a anotação das vias metabólicas, na qual o objetivo é a identificação das rotas bioquímicas completas, incompletas e alternativas que o organismo possui (PRODOSCIMI et al., 2003). Segundo Prodoscimi e colaboradores nesse processo é "fundamental a participação de biólogos especialistas em diversas áreas para que se possa descobrir como o metabolismo do organismo pode influenciar seu modo de vida e seu comportamento. Esse é o momento onde é possível levantar várias hipóteses que relacionem o funcionamento dos organismos com dados de seu genoma. Tais hipóteses devem ser testadas experimentalmente, permitindo assim a caracterização do genoma funcional em resposta a condições diversas as quais o organismo pode ser submetido "in vitro" ou "in vivo" (PRODOSCIMI et al., 2003).

Na atualidade, grande parte dos genes e proteínas potenciais identificados no processo da anotação, não possuem homólogos conhecidos ou apresentam homologia com genes e proteínas de função biológica desconhecida, sendo, portanto considerados como hipotéticos ou desconhecidos. Estima-se que em *Escherichia coli*, *Arabidopsis thaliana* e em *Drosophila melanogaster*, em torno de 40 a 60% dos genes anotados não possuem produto gênico ou função biológica conhecida (SANTOS e ORTEGA, 2003).

De acordo com Formighieri (2005), a fim de se facilitar as análises futuras e para que se atinja o objetivo de um projeto genoma é importante que, desde o início, a anotação seja feita com minúcia e precisão, confirmando-se a identificação dos genes e sua função, pois, uma anotação mal feita pode gerar ou frustrar expectativas futuras e acima de tudo levará tempo para sua correção.

Genomas de bactérias diazotróficas sequenciadas

A grande maioria dos genomas de bactérias fixadoras de nitrogênio depositadas no banco de genes (NCBI) refere-se a bactérias simbióticas do gênero *Rhizobium*, *Bradyrhizobium*, *Sinorhizobium*, *Mesorhizobium* e *Azorhizobium*, que fixam nitrogênio em estrutura especializada formada durante associação com plantas leguminosas. Dentre bactérias que se associam às plantas não leguminosas, especificamente com aquelas da família Poaceae, já foram sequenciados os genomas de *Azoarcus* BH72 (KRAUSE et al., 2008), *Pseudomonas stutzeri* estirpe A1501 (YAN et al., 2008), *Klebsiela pneomoniae* estirpe 342 (FOUTS et al., 2008), *Gluconacetobacter diazotrophicus* estirpe PAL5T (BERTALAN et al., 2009) e *Herbaspirillum seropedicae* estirpe Z78 (PEDROSA et al, 2011). Em adição, os genomas das espécies *Azospirillum brasilense* estirpe Sp245, isolada pela equipe da Embrapa Agrobiologia, e de *Burkholderia unamae* estirpe MTI-641 também estão sendo sequenciados. Portanto, existe uma oportunidade única para o avanço de conhecimento e exploração do potencial biotecnológico dessas bactérias. Entretanto,

é essencial que os estudos sobre o genoma das bactérias como, por exemplo, da bactéria *G. diazotrophicus*, estejam associados aos conhecimentos acumulados ao longo dos anos sobre a própria bactéria especialmente aqueles aspectos que efetivamente contribuirão para o futuro uso prático da bactéria na agricultura.

Metodologia aplicada no sequenciamento e na montagem do genoma

O sequenciamento do DNA total das estirpes do inoculante para cana-de-açúcar (*Herbaspirillum seropedicae* HRC54, *H. rubrisubalbicans* HCC103, *Burkholderia tropica* PPe8 e *Azospirillum amazonense* CBAmC) e da estirpe do inoculante de milho (*H. seropedicae* ZAE94), listadas na Tab. 1, foi realizado utilizando o sequenciamento em larga escala na plataforma do Illumina/Solexa em parceria com a empresa Fasteris, Suíça. Cabe informar que a estirpe *G. diaoztrophicus* PAL5 não faz parte do presente documento uma vez que o genoma seqüenciado (BERTALAN et al., 2009) já está disponível na base de dados do Centro Nacional para Informação Biotecnológica dos EUA (NCBI) e que pode ser acessada sem restrições.

As informações sobre 03 genomas completos de bactérias diazotróficas e 01 de bactéria endofítica não fixadora de nitrogênio (Tab. 2), previamente depositados no banco de genes do NCBI, foram usadas como referência para a montagem do genoma de cada estirpe do inoculante. A validação do sequenciamento foi feita através do software MAQ (*Mapping and Assembly with Qualities*, version 0.7.1).

As sequências obtidas foram montadas utilizando o programa VELVET (realizado pela empresa FASTERIS SA), que é um software para montagem de genomas, concebido para utilizar sequências curtas (pequenos *reads*), gerados pelas tecnologias de em larga escala, e capazes de montar *contigs* com alta qualidade (ZERBINO e BIRNEY, 2008).

Tabela 1. Bactérias diazotróficas endofíticas que compõem os inoculantes de cana-de-açúcar e milho.

Espécies	Estirpe*
Inoculante para cana-de-açúcar	
Herbaspirillum seropedicae	BR 11335 = HRC54
Herbaspirillum rubrisubalbicans	BR 11504 = HCC103
Azospirillum amazonense	BR 11145 = CBAmC
Burkholderia tropica	BR 11366^{T} = PPe8
Gluconacetobacter diazotrophicus¤	BR 11281^{T} = PAL5
Inoculante para milho	
Herbaspirillum seropedicae	BR 11417 = ZAE94

*Todas as estirpes estão depositadas na coleção de culturas da Embrapa Agrobiologia. ¤ Genoma previamente seqüenciado e disponível na base de dados do Centro Nacional para Informação Biotecnológica dos EUA (NCBI), podendo ser acessado sem restrições.

Tabela 2. Bactérias diazotróficas e endofíticas utilizadas como genoma de referência no processo de sequenciamento e montagem dos genomas das estirpes dos inoculantes.

Espécie/estirpe sequenciada	Genoma de referência	Informações adicionais
Herbaspirillum seropedicae estirpes HRC54 e ZAE 94	*H. seropedicae* estirpe Z78, SMR1	NC_014323.1; tamanho 5,6 Mb; Pedrosa *et al.*, 2011
Herbaspirillum rubrisubalbicans estirpe HCC103	*H. seropedicae* estirpe Z78, SMR1	
Burkholderia tropica estirpe PPe8	*B. phytofirmans* estirpe PsJN*	NC_10681 e NC_010676; possui 01 plasmídio; Weilharter *et al.*, 2011
Azospirillum amazonense estirpe CBAmC	*Azospirillum* sp B510	NC. 013854.1; tamanho 4,83 Mb; possui 06 plasmídeos; Kaneko *et al.*, 2010

*Bactéria endofítica e não-diazotrófica

A ferramenta computacional RAST foi utilizada com o objetivo de localizar de forma rápida os potenciais genes presentes no genoma. O programa é automatizado e foi desenvolvido para a anotação de genomas bacterianos e archaea. O mesmo identifica as ORFs (*Open Read Frames*), rRNA e tRNA e atribui funções aos genes (AZIZ et al. 2008).

Informações geradas pelo sequenciamento das estirpes e anotação automática no sistema RAST

O sequenciamento do genoma das cinco (05) estirpes do inoculante para cana e milho, tomando como base os genomas de referência listados na Tab. 2, mostrou uma cobertura de mais de 95% do total de *reads* mapeados (Tab. 3). Quando analisados os *reads* mapeados nas duas pontas, os valores foram menores, já que foi obtido um percentual de 1,6 a 3% de *reads* quando mapeados somente em uma ponta.

O emprego do programa MAQ (*Mapping e Assembly with Qualities*) possibilitou estimar o tamanho do genoma das estirpes que compõem o inoculante de cana e milho (Tab. 4). As duas estirpes da espécie *Herbaspirillum seropedicae* apresentaram tamanho muito similares: 5.530.106 pb (estirpe HRC54), 5.653.622 pb (estirpe ZAE94). De maneira muito similar, ocorreu com a estirpe HCC103 pertencente a espécies *H. rubrisubalbicans*: 5.628. 704 pb. Por outro lado, o tamanho do genoma da estirpe *A. amazonense* CBAmc foi estimado em 7. 256.180 pb, considerado menor do que de outras espécies de *Azospirillum* já depositadas no banco de dados. O tamanho do genoma da estirpe de *B. tropica* PPe8, foi de 8.720.735 pb, considerado normal para genomas pertencentes ao gênero *Burkholderia*. A montagem dos genomas gerou um número bastante variável de *contigs* (567 a 4832) e *scaffolds* (176 a 879) e que foi dependente da estirpe analisada (Tab. 4). O tamanho médio de cada *contig* apresentou-se na faixa de 1558 a 9846 pb enquanto que esse valor variou de 8255 a 31421 para os *scaffolds*. O *contig* de maior tamanho foi calculado possuir 76.355 pb

Tabela 3. Percentual de *reads* mapeados nas estirpes do inoculante para cana e milho com base nas sequências completas dos genomas das bactérias referências listadas acima.

Estirpes mapeadas	% de *reads* mapeadas	% de *reads* mapeados nas duas pontas	% de *reads* mapeados em uma ponta
H. seropedicae ZAE94	97.0	94.9	2.0
H. seropedicae HRC54	97.3	95.7	1.6
H. rubrisubalbicans HCC103	97.9	96.2	1.6
B. tropica PPe8	96.0	93.2	2.9
A. amazonense CBAmC	95.8	93.0	3.0

Tabela 4. Informações genômicas das cinco estirpes diazotróficas geradas pelo sequenciamento na plataforma Illumina/Solexa.

Organismo	*Scaffolds*			*Contigs*		
	Nº	Tamanho médio (pb)	Total (pb)	Nº	Tamanho médio (pb)	Tamanho máximo (pb)
H. seropedicae ZAE94	213	26.542	5.653.622	1.459	3.820	51.288
H. seropedicae HRC54	176	31.421	5.530.106	1.188	4.602	60.470
H. rubrisubalbicans HCC103	228	24.687	5.628.704	567	9.846	76.335
B. tropica PPe8	681	12.805	8.720.735	4.832	1.760	15.133
A. amazonense CBAmC	879	8.255	7.256.180	4.560	1.558	28.919

para a estirpe HCC103 pb, enquanto o *scaffold* de maior tamanho foi de 521.241 pb na estirpe ZAE94 (Tab. 4).

A anotação automática gerada pelo programa RAST mostrou a ocorrência de praticamente todas as categorias de genes encontrados em outras bactérias já sequenciadas e depositadas nos bancos de dados, incluindo as diazotróficas (Fig. 1). Algumas categorias apresentaram um número de sequências codantes superior a 200: carboidratos; aminoácidos e derivados; metabolismo de proteínas;

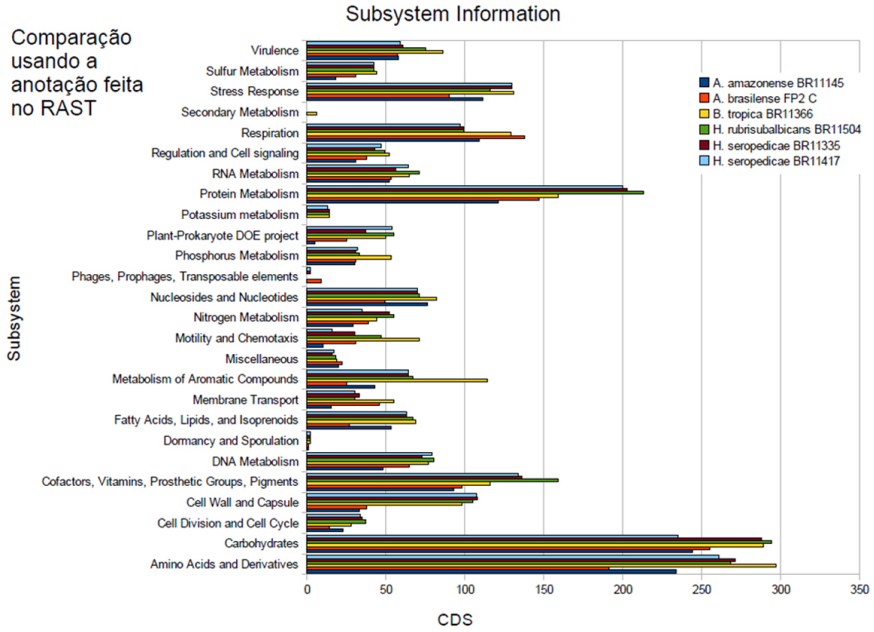


Fig. 1. Categoria de genes identificados nos genomas das estirpes que compõem o inoculante de cana-de-açúcar e milho gerados por meio da anotação do sistema RAST.

enquanto as categorias resposta a estresse, respiração, cofatores, vitaminas, grupos prostéticos e pigmentos apresentaram mais de 100 sequencias codificadoras. Outras categorias essencialmente envolvidas em atividades metabólicas em bactérias diazotróficas, mas em menor número de genes foram: virulência, metabolismo de S, K e P etc.

É sabido que somente com a anotação automática não será possível identificar todos os genes e as respectivas funções nos genomas das cinco estirpes. Para tanto, faz-se necessário estabelecer uma base de dados que permita a avaliação individual dos genes previamente anotados pelo sistema automático RAST. Além disso, deve permitir a exploração do potencial biotecnológico dos genes presentes nos genomas, em especial aqueles envolvidos no processo de fixação biológica de nitrogênio, na interação planta-bactéria, na promoção

Tabela 5. Subcategorias de genes presentes nos genomas das estirpes sequenciadas e que estão envolvidos na formação da parede celular e capsular.

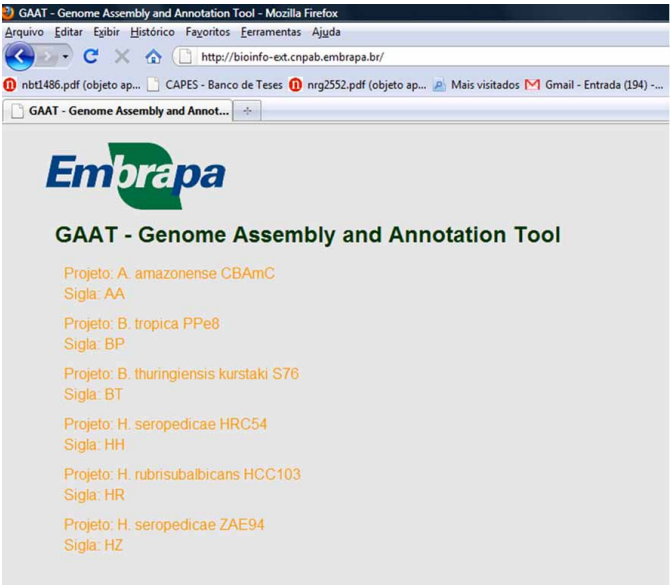
Estirpes	Genes identificados no genoma		
	Parede celular e capsular	Polissacarídeos extracelulares e capsulares	Componentes de parede celular
H. seropedicae ZAE94	128	37	34
H. seropedicae HRC54	132	33	34
H. rubrisubalbicans HCC103	125	38	33
B. tropica PPe8	140	68	38
A. amazonense CBAmC	140	17	0

de crescimento e também daqueles 30 a 40% de genes identificados na categoria de função desconhecida. A Tab. 5 exemplifica algumas subcategorias de genes que estão envolvidos com o processo de formação de parede celular e capsular encontradas nas cinco estirpes.

Base de dados genômica das estirpes do inoculante para cana e milho

A base de dados contendo as informações genômicas (sequências) das 05 estirpes encontra-se hospedada nos servidores da Embrapa Agrobiologia, Seropédica, RJ. O acesso ao genoma é feito através do programa GAAT (*Genome Assembly and Annotation Tool*), gentilmente cedido pela UFPR, no endereço http://bioinfo-ext.cnpab.embrapa.br conforme ilustrado nas Fig. 2 e 3. Para a prospecção do conteúdo genômico da estirpe de interesse é necessário apresentar proposta de projeto visando o desenvolvimento de pesquisas em conjunto. Após a aprovação da proposta o usuário recebe a senha designada pelo coordenador do projeto.

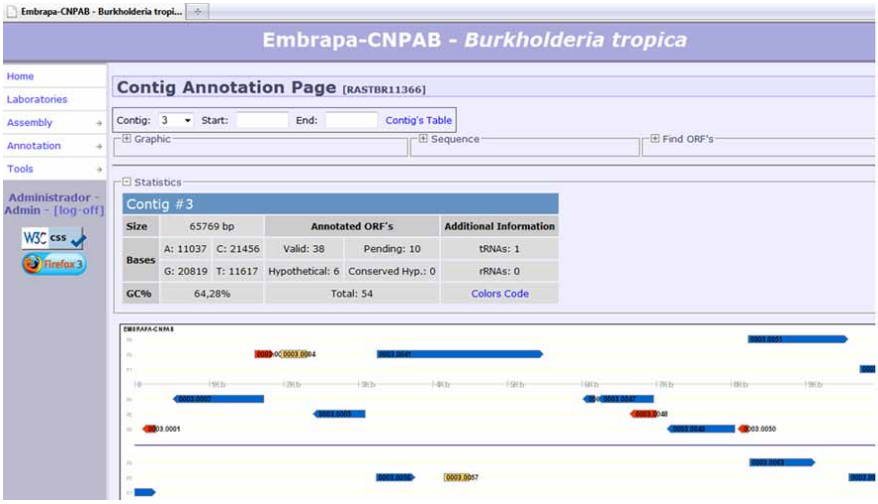
A



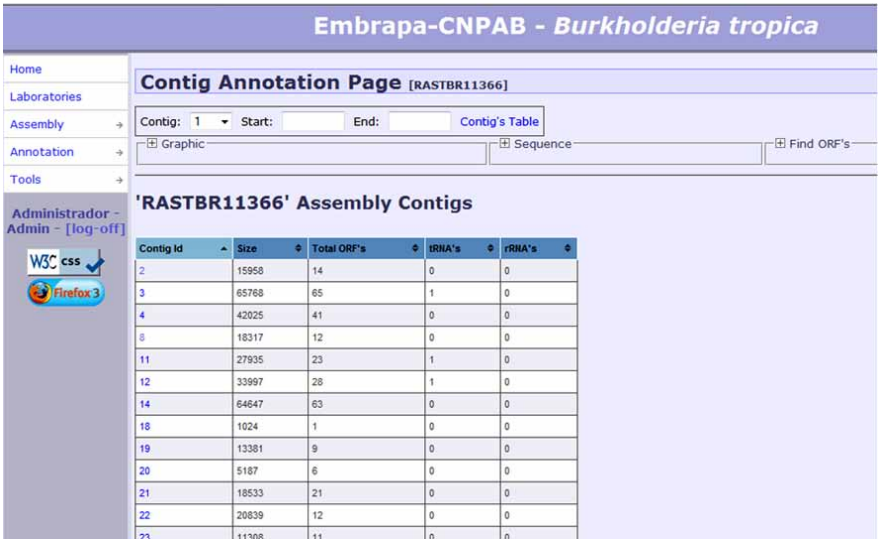
B

Fig. 2. Página do website da Embrapa Agrobiologia com o endereço de acesso ao programa GAAT (**Fig. 2A**) e com a interface da base de dados do genoma da estirpe *Burkholderia tropica* PPe8 (**Fig. 2B**).

A



B



Contig Id	Size	Total ORF's	tRNA's	rRNA's
2	15958	14	0	0
3	65768	65	1	0
4	42025	41	0	0
8	18317	12	0	0
11	27935	23	1	0
12	33997	28	1	0
14	64647	63	0	0
18	1024	1	0	0
19	13381	9	0	0
20	5187	6	0	0
21	18533	21	0	0
22	20839	12	0	0
23	11208	11	0	0

Fig. 3. Interface da base de dados do genoma da estirpe *Burkholderia tropica* PPe8 com as informações referentes ao número de *contigs* (**Fig. 3A**) e o detalhamento do número as ORFs presentes no *contig* # 3 (**Fig. 3B**).

Perspectivas

O desenvolvimento biotecnológico de inoculantes é uma das prioridades nas pesquisas da Embrapa Agrobiologia, o que necessariamente requer um constante avanço do conhecimento científico para atender as demandas resultantes de lançamento de novas variedades e cultivares de culturas de importância comercial aliado ao impacto das mudanças climáticas na agricultura. Neste contexto, é necessário determinar, a nível biológico, como a comunicação molecular entre planta e bactéria ocorre e como influencia sua regulação genética durante processos como a fixação e transferência do nitrogênio (N) para a planta e, em outras atividades que estimulam o crescimento vegetal durante a associação entre os organismos. Adicionalmente, estudos comparativos do genoma funcional de mutantes e da estirpe selvagem do modelo endofítico *G. diazotrophicus* e a caracterização funcional de promotores e proteínas permitirão elucidar os diversos processos que exercem papel importante para a sobrevivência, o potencial biotecnológico e o caráter endofítico dessa bactéria, tais como adesão, colonização e estabelecimento em plantas de cana-de-açúcar e outros plantas da família das Poaceas.

O conhecimento de menos 90% das sequências do genoma de cada estirpe permitirá que sejam identificadas aquelas que se encontram normalmente conservadas nestas espécies de bactérias fixadoras de nitrogênio, aquelas que são exclusivas de cada espécie comparada e possivelmente genes com função desconhecidos conforme já observado para outros genomas descritos na literatura (KRAUSE et al., 2006; FOUTS et al., 2008). É possível que genes com atividade de interesse para outras áreas de pesquisa (farmacêutica e industrial) sejam encontrados conforme já observado para a *Gluconacetobacter diazotrophicus* (BERTALAN et al., 2009) e *Gluconobacter oxydans* (PRUST et al., 2005), *Herbaspirillum seropedicae* SMR1 (PEDROSA et al., 2011) sequenciados recentemente. As possibilidades de patenteamento de genes na área de FBN são ainda muito incipientes, porém existe potencial para o caso dos promotores de transcrição

que estão sendo selecionados. A prospecção de genes de interesse presentes nos genomas sequenciados será continuada e, portanto não deverá ser descartado o desenvolvimento de novos compostos e suas aplicações derivado do conhecimento acumulado.

Na área de Agrobiotecnologia, os ganhos advindos do sequenciamento dessas bactérias virão com a aplicação do conhecimento gerado com a interrupção da função dos genes de interesse e avaliação na interação com a planta, assim como da super-expressão de genes alvos sob o controle de promotores de transcrição especificamente selecionados. Não se pode também descartar os futuros programas de transgenia através da clonagem e expressão destes genes em bactérias diazotróficas endofíticas visando o controle de pragas e doenças.

Em conclusão, podemos inferir que o sequenciamento genômico das estirpes do inoculante para cana e milho, abre novas perspectivas para a prospecção de genes de interesse Agrobiotecnológico e a consequente exploração comercial com a possibilidade de depósito de patentes. Se considerarmos que existem cerca de 4.000 genes no genoma da cada estirpe e que 1% possuem potencial biotecnológico, teríamos então 40 pedidos de patentes para depósito junto ao INPI (Instituto Nacional de Propriedade Industrial). Obviamente, haverá necessidade de esforço contínuo da pesquisa junto ao banco de dados das estirpes de bactérias diazotróficas para que o potencial biotecnológico seja efetivamente explorado e forneça à sociedade brasileira informações que agregam valor ao agronegócio nacional.

Agradecimentos

Os autores agradecem o apoio financeiro recebido do projeto INCT-FBN/ CNPq/MCT, FAPERJ/Cientista do Nosso Estado e da Embrapa.

Referências Bibliográficas

BALCALDE, F. S. C. **Elucidação do destino metabólico de glicose no fungo filamentoso *Trichoderma reesei* por análise EST (Espressed Sequences Tags) e microarrays de cDNA**. 2002. 108 f. Tese (Doutorado em Ciências: Bioquímica) - Instituto de Química, Departamento de Bioquímica, Universidade de São Paulo, São Paulo, 2002.

ALTSCHUL, S. F. Amino acid substitution matrices from an information theoretic perspective. **Journal of Molecular Biology**, v. 219, p. 555-565, 1991.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E.W.; LIPMAN, D. J. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, p. 403-410, 1990.

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

AZIZ, R.; BARTELS, D.; BESTE, A. A.; DEJONGH, M.; DISZ, T.; EDWARDS, R.; FORMSMA, K.; GERDES, S.; GLASS, E.; KUBAL, M.; MEYER, F.; OLSEN, G. J.; OLSON, R.; OSTERMAN, A. L.; OVERBEEK, R. A.; MCNEIL, L. K.; PAARMANN, D.; PACZIAN, T.; PARRELLO, B.; PUSCH, G. D.; REICH, C.; STEVEN, R.; VASSIEVA, O.; VONSTEIN, V.; WILKES, A.; ZAGNITKO, O. The RAST Server: Rapid Annotations using subsystems Technology. **BMC Genomics**, v. 9, n. 1, p. 75, 2008.

BENSON, D. A.; KARSCH-MIZRACHI, I.; LIPMAN, D. J.; OSTELL, J.; SAYERS, E. W. GenBank. **Nucleic Acids Res.**, 39: D32-D37, 2011.

BAIROCH, A.; APWEILER, R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. **Nucleic Acids Research**, v. 24, p. 21-25, 1996.

BERTALAN, M.; ALBANO, R. M.; PÁDUA, V. L. M.; ROUWS, L. F. M.; ROJAS, C.; HEMERLY, A.; TEIXEIRA, K. R. dos S.; SCHWAB, S.; ARAUJO, J. L. S. de; OLIVEIRA, A.; FRANÇA, L.; MAGALHÃES, V.; ALQUÉRES, S.; CARDOSO, A.; ALMEIDA, W.; LOUREIRO, M. M.; NOGUEIRA, E.; CIDADE, D.; OLIVERIA, D.; SIMÃO, T.; MACEDO, J.; VALADÃO, A.; DRESCHSEL, M.; FREITAS, F.; VIDAL, M. S.; GUEDES, H.; RODRIGUES, E.; MENESES, C. H. S. G.; BRIOSO, P. S. T.; POZZER, L.; FIGUEIREDO, D.; MONTANO, H.; CUNHA JÚNIOR, J. de O.; SOUZA FILHO, G. de; FLORES, V. M. Q.; FERREIRA, B.; BRANCO, A.; GONZALEZ, P.; GUILLOBEL, H.; LEMOS, M.; SEIBEL, L.; MACEDO, J.; FERREIRA, A. M.; MARTINS, G. S.; COELHO, A.; SANTOS, E.; AMARAL, G.; NEVES, A.; PACHECO, B. A.; CARVALHO, D.; LERY, L.; BISCH, P.; ROSSLE C, S.; URMÉNYI, T.; PEREIRA, R. A.; SILVA, R.; RONDINELLI, E.; KRUGER, W. V.; MARTINS, O.; BALDANI, J. I.; FERREIRA, P. C. G. Complete genome sequence of the sugarcane nitrogen-fixing endophyte *Gluconacetobacter diazotrophicus* Pal5. **BMC Genomics**, v. 10, n. 450, p. 1-17, 23 set., 2009.

EBERSBERGER, I.; METZLER, D.; SCHWARZ, C.; PAABO, S. Genomewide comparison of DNA sequences between humans and chimpanzees. **American Journal of Human Genetics**, v. 70, p. 1490-1497, 2002.

FORMIGHIERI, E. F. **Manual de anotação café**. Disponível em <http:// www.lge.ibi.unicamp.br/ manuais/manual_ anota_cafe.htm> Acesso em: 22/11/2005).

FOUTS, D. E.; TYLER, H. L.; DEBOY, R. T.; DAUGHERTY, S.; REN, Q.; BADGER, J. H.; DURKIN, A. S.; HUOT, H.; SHRIVASTAVA, S.; KOTHARIS, S.; DODSON, R. J.; MOHAMOUD, Y.; KHOURI, H.; ROESCH, L. F.; KROGFELT, K. A.; STRUVE, C.; TRIPLETT, E. W. Complete genome sequence of the N_2-fixing broad host range endophyte *Klebsiella pneumoniae* 342 and virulence predictions verified in mice. **PLoS Genetic**s, v. 4, n. 7, 2008. Doi: 10.1371/journal.pgen.1000141.

GOLD. **Genomes Online Database**. Disponível em: <http://www. genomesonline.org/> Acesso em: 20/08/2013.

GRISSA, I.; POURCEL, V. G. The CRISRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. **BMC Bioinformatics**, V. 8, p. 172, 2007. 8:172. [PubMed:1752143]

KANEHISA, M.; GOTO, S.; KAWASHIMA, S.; OKUNO, Y.; HATTORI, M. The KEGG resource for deciphering the genome. **Nucleic Acids Research**, v. 32, p. 277-280, 2004.

KANEKO, T.; MINAMISAWA, K.; ISAWA,T.; NAKATSUKASA, H.; MITSUI, H.; KAWAHARADA, Y.; NAKAMURA, Y.; WATANABE, A.; KAWASHIMA, K.; ONO, A.; SHIMIZU, Y.; TAKAHASHI, I. C., MINAMI, C.; FUJISHIRO, T.; KOHARA, M.; KATOH, M.; NAKAZAKI, N.; NAKAYAMA, S.; YAMADA, M.; TABATA, S. I.; SATO, S. Complete genomic structure of the cultivated rice endophyte *Azospirillum* sp. B510. **DNA Research**, v. 317, n. 1, p. 37-50, 2010.

KRAUSE, A.; RAMAKUMAR, A.; BARTELS, D.; BATTISTONI, F.; BEKEL, T.; BOCH, J.; BOEHM, M.; FRIEDRICH, F.; HUREK, T. Complete genome of the mutualistic, N_2-fixing grass endophyte *Azoarcus* sp. strain BH72. **Nature Biotechnology**, v. 24, p. 1385-1391, 2008.

LANDER, E. S.; LINTON, L. M.; BIRREN, B.; NUSBAUM, C.; ZODY, M. C.; BALDWIN, J. Inital sequencing and analysis of the human genome. **Nature**, v. 409, p. 860-921, 2001.

MEYERSON, M.; COUNTER, C. M.; EATON, E. N.; ELLISEN, L. W.; STEINER, P.; CADDLE, S. D.; ZIAUGRA, L.; BEIJERSBERGEN, R. L.; DAVIDOFF, M. J.; LUI, Q.; BACCHETTI, S.; HABER, D. A.; WEINBURG, R. A hEST2, the putative human telomerase catalytic subunit gene, is up-regulated in tumor cells and during immortalization. **Cell**, v. 90, p. 785-795, 1997.

NAPOLI, W. F. M. **Introdução à Bioinformática**. 2003. 161 f. Projeto Final (Graduação em Engenharia da computação: Bioinformática) Escola de Engenharia Elétrica e de Computação, Universidade Federal de Goiás, 2003.

PEDROSA, F. O.; MONTEIRO, R. A.; WASSEM, R.; CRUZ, L. M.; AYUB, R. A.; COLAUTO, N. B.; FERNANDEZ, M. A.; FUNGARO, M. H. P.; GRISARD, E. C.; HUGRIA, M.; MADEIRA, H. M. F.; NODARI, R. O.; OSAKU, C. A.; PETZL-ERLER, M. A.; TERENZI, H.; VIEIRA, L. G. E.; STEFFENS, M. B. R.; WEISS, V. A.; PEREIRA, L. F. P.; ALMEIDA, M. I. M.; ALVES, L. R.; MARIN, A.; ARAÚJO, L. M.; BALSANELLI, E.; BAURA, V. A.; CHUBATSU, L. S.; FAORO, H.; FAVETTI, A.; FRIEDEERMANN, G.; GLIENKE, C.; KARP, S.; KAVA-CORDEIRO, V.; RAITTZ, R. T.; RAMOS, H. J. O.; RIBEIRO, E. M. S. F.; RIGO, L. U; ROCHA, S. N.; SCHWAB, S.; SILVA, A. G.; SOUZA, E. M.; TADRA-SFEIR, M.; TORRES, R. A.; DABUL, A. N. G.; SOARES, M. A. M.; GASQUES, L. S.; GIMENES, C. C. T.; VALLE, J. S.; CIFERRI, R. R.; CORREA, L. C.; MURACE, N. K.; PAMPHILE, J. A.; PATUSSU, E. V.; PRIOLI, A. J.; PRIOLI, S. M.; ROCHA, C. L. M. S. C.; ARANTES, O. M.

N.; FURLANETO, M. C.; GODOY, L. P.; OLIVEIRA, C. E. C.; SATORI, D.; VILAS-BOAS, L. A.; WATANABES, M. A. E.; DAMBROS, B. P.; GUERRA, M. P.; MATHIONIS, S. M.; SANTOS, K. L.; STEINDEL, M.; VERNAL, J.; BARCELLOS, F. G.; CAMPO, R. J.; CHUEIRE, L. M. O.; NICOLÁS, M. F.; FERRARI, L. P.; SILVA, J. L. C.; GIOPPO, N. M. R.; MARGARIDO, V. P.; MENCK-SOARES, M. A.; PINTO, F. G. S.; SIMÕES, R. C.; TAKAHASHI, E. K.; YATES, M. G.; SOUZA, E. M. Genome of *Herbaspirillum seropedicae* Strain SmR1, a specialized diazotrophic endophyte of tropical grasses. **PLoS Genetics**, v. 7, p. 1-10, 2011.

PERNA, N. T.; BURLAND, V.; MAU, B.; GLASNER, J. D.; ROSE, D. J.; MAYHEW, G. F.; EVANS, P. S.; GREGOR, J.; KIRKPATRICK, H. A.; POSFAI, G.; HACKETT, J.; KLINK, S.; BOUTIN, A.; SHAO, Y.; MILLER, L.; GROTBECK, E. J.; DAVIS, N. W.; LIM, A.; DIMALANTA, E. T.; POTAMOUSIS, K. D.; APODACA, J.; ANANTHARAMAN, T. S.; LIN, J.; YEN, G.; SCHWARTZ, D. C.; WELCH, R. A.; BLATTNER, F. R. Genome sequence of enterhaemorrhagic *E. coli* O157:H7. **Nature**, v. 409: p. 529-533, 2001.

PUNTA, M.; COGGILL, P. C.; EBERHARDT, R.; MISTRY, J.; TATE, J.; BOURSNELL, C.; PANG, N.; FORSLUND, K.; CERIC, G.; CLEMENTS, J.; HEGER, A.; HOLM, L.; SONNHAMMER, E. L. L.; EDDY, S. R.; BATEMAN, A. R. D. The Pfam protein families database. **Finn Nucleic Acids Research Database Issue**, v. 40, p. 290-301, 2012.

PROSDOCIMI, F.; CERQUEIRA, G. C.; BINNECK E.; SILVA, A. F.; REIS, A. N. dos; JUNQUEIRA, A. C. M.; SANTOS, A. C. F. dos; NBANI JÚNIOR, A.; WUST, C. I.; CAMARGO FILHO, F.; KESSEDJIAN, J. L.; PETRETSKI, J. H; CAMARGO, L. P.; FERREIRA, R. de G. M.; LIMA, R. P.; PEREIRA, R. M.; JARDIM, S.; SAMPAIO V. de S.; FLATSCHART, A. V. F. Bioinformática: manual do usuário. **Revista Biotecnologia Ciência & Desenvolvimento**, v. 5, n. 29, nov./dez., 2002.

PRUST, C.; HOFFMEISTER, M.; LIESEGANG, H.; WIEZER, A.; FRICKE, W. F.; EHRENREICH, A.; GOTTSCHALK, G.; DEPPENMEIER, U. Complete genome sequence of the acetic acid bacterium *Gluconobacter oxydans*. **Nature Biotechnology**, v. 23, n. 2, p.195-200, 2005.

SANTOS, F. R.; ORTEGA, J. M. **Bioinformática aplicada à genômica**. 2003. Manuscrito para capítulo do Biowork IV. 21 páginas,

VENTER, J. C.; ADAMS, M. D.; MYERS, E. W.; LI, P. W.; MURAL, R. J.; SUTTON, G. G.; SMITH, H. O.; YANDELL, M. The sequence of the human genome. **Science**, v. 291, p. 1304-1351, 2001.

WHEELER, D. L.; CHURCH, D. M.; FEDERHEN, S.; LASH, A. E.; MADDEN, T. L.; PONTIUS, J. U.; SCHULER, G. D.; SCHRIML, L. M.; SEQUEIRA, E.; TATUSOVA, T. A.; WAGNER, L. Database resources of the National Center for Biotechnology. **Nucleic Acids Research**, v. 31, n. 1, p. 28-33, 2003.

WEILHARTER, A.; MITTER, B.; SHIN, M. V.; CHAIN, P. S. G.; NOWAK, J.; SESSITSCH, A. Complete genome sequence of the plant growth-promoting endophyte *Burkholderia phytofirmans* Strain PsJN. **Journal of Bacteriology**, v. 193, n. 13, p. 3383, 2011.

YAN, Y.; YANG, J.; DOU, Y.; CHEN, M.; PIN, S.; PENG, J.; LU, W.; ZHANG, W.; YAO, Z.; LI, H.; LIU, W.; HE, S.; GENG, L.; ZHANG, X.; YANG, F.; YU, H.; ZHAN, Y.; LI, D.; LIN, Z.; WANG, Y.; ELMERICH, C.; LIN, M.; JIN, Q. 2008. Nitrogen fixation island and rhizosphere competence traits in the genome of root-associated *Pseudomonas stutzeri* A1501. **Proceedings of the National Academy of Sciences**. USA, v. 105, p. 7564-7569, 2008.

ZERBINO, D. R.; BIRNEY, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. **Genome Research**, v. 18, p. 821-829, 2008.

Glossário de termos

ANOTAÇÃO - Atribuição de características funcionais, estruturais primárias através da busca por homologia das sequências de DNA obtidas por sequenciamento ou secundárias e terciárias por predição de sequencia de aminoácidos codificados. Descrição de funções e identificação de genes e seus produtos.

CONTIG - Segmento contíguo de sequências de DNA, resultante da montagem de *reads* obtidos como parte de um projeto de sequenciamento de genoma.

ORF - "Open Reading Frames" ou fase aberta de leitura do DNA, que correspondem a sequências com possíveis regiões codificadoras, identificadas por um códon iniciador e um terminador.

READS - Sequências de DNA obtidos a partir de uma reação de sequenciamento.

SCAFFOLD - Conjunto de *contigs* organizados em sequência e representam grandes porções de um genoma.

Embrapa

Agrobiologia

Ministério da
Agricultura, Pecuária
e Abastecimento

GOVERNO FEDERAL
BRASIL
PAÍS RICO É PAÍS SEM POBREZA