Laboratório Nacional de Computação Científica
Programa de Pós-Graduação em Modelagem Computacional
Curso Modelagem Computacional com Ênfase em Bioinformática e Biologia Computacional

GINGA – *GRAPHICAL INTERFACE FOR COMPARATIVE GENOME ANALYSIS*: O DESENVOLVIMENTO DE UM SISTEMA COMPUTACIONAL DE VISUALIZAÇÃO GRÁFICA PARA A ANÁLISE COMPARATIVA DE GENOMAS DE BACTÉRIAS

Por
Alexandre Rossi Paschoal

sob orientação de
Claudia de Barros Monteiro-Vitorello

e co-orientação de
Ana Tereza Ribeiro de Vasconcelos

Março de 2007
Petrópolis, RJ – Brasil

GINGA – *GRAPHICAL INTERFACE FOR COMPARATIVE GENOME ANALYSIS*: O DESENVOLVIMENTO DE UM SISTEMA COMPUTACIONAL DE VISUALIZAÇÃO GRÁFICA PARA A ANÁLISE COMPARATIVA DE GENOMAS DE BACTÉRIAS

Alexandre Rossi Paschoal

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DE FORMAÇÃO DE RECURSOS HUMANOS DO LABORATÓRIO NACIONAL DE COMPUTAÇÃO CIENTÍFICA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM MODELAGEM COMPUTACIONAL COM ÊNFASE EM BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL.

Avaliado por:

_____
Claudia Barros Monteiro Vitorello
Orientadora

_____
Ana Tereza Ribeiro Vasconcelos
Co-Orientadora

_____
Marcio Alves-Ferreira
UFRJ

_____
Luis Eduardo Aranha Camargo
ESALQ-USP

Março de 2007
Petrópolis, RJ – Brasil

PASCHOAL, ALEXANDRE ROSSI
GINGA - *Graphical INterface for comparative Genome Analysis*: o desenvolvimento de um sistema computacional de visualização gráfica para a análise comparativa de genomas de bactérias
Petrópolis 2007
XXII, 78 p. 29,7 cm (MCT/LNCC, M.Sc., Modelagem Computacional, 2007)
Dissertação - Laboratório Nacional de Computação Científica, LNCC
1. Bioinformática
2. Genômica comparativa
3. Bactérias fitopatógenas
I. MCT/LNCC II. Título (Série)

Agradecimentos

À Dra. Claudia B. Monteiro Vitorello, amiga, mestra e orientadora que acreditou em mim desde o início do meu mestrado, sendo sempre paciente, compreensiva e mostrando enorme sabedoria em me orientar.

À Dra. Ana Tereza R. de Vasconcelos, pela oportunidade de trabalhar no LABINFO, e apoio em todas as etapas do meu trabalho.

Aos colegas do LABINFO, que ajudaram com sugestões, críticas e idéias em intensidades e momentos diferentes. Em especial: Oberdan, Fabíola, Luciane, Marisa, Luiz Gonzaga, Zuleta, Jorge, Márcia, Rangel, Vicente, Fernanda e Alex.
Aos demais colegas pelos momentos inesquecíveis de convivência.

Às Instituições que apoiaram este trabalho, cada qual de sua forma:

- Labinfo – Laboratório de bioinformática;
- LNCC – Laboratório Nacional de Computação Científica;
- CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superio;
- MCT - Ministério da Ciência e Tecnologia.

Agradeço, ainda, àqueles que desde o início estiveram comigo e sempre estarão, sendo a minha razão de viver. À minha família: Mary, Antonio, Odila, João Paulo, Egle, Enzo, Vi, Nádia, Felipe, Lila, Nina e o “Meio Quilo”.

Por fim, termino com uma frase que me inspira: “*Nunca, jamais, desanimeis, embora venham ventos contrários.*” (Santa Paulina)

Resumo da Dissertação apresentada ao MCT/LNCC como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

GINGA – *GRAPHICAL INTERFACE FOR COMPARATIVE GENOME ANALYSIS*: O DESENVOLVIMENTO DE UM SISTEMA COMPUTACIONAL DE VISUALIZAÇÃO GRÁFICA PARA A ANÁLISE COMPARATIVA DE GENOMAS DE BACTÉRIAS

Alexandre Rossi Paschoal

23 de março de 2007

Orientadoras: Claudia de Barros Monteiro Vitorello
Ana Tereza Ribeiro de Vasconcelos

Modelagem Computacional com Ênfase em Bioinformática e Biologia Computacional

Esta dissertação resultou de um sistema computacional voltado para a visualização gráfica de análises comparativas entre genomas de procariotos. O sistema denominado de GINGA – *Graphical Interface for comparative Genome Analysis* – foi desenvolvido basicamente para analisar genomas parcialmente seqüenciados por meio da comparação com genomas completos. O sistema mostra a representação do alinhamento entre seqüências de *reads*, *contigs* e *scaffolds* do genoma parcial com a seqüência completa do outro genoma, permitindo a identificação de blocos comuns, regiões específicas e rearranjos. GINGA é um sistema *web-based* que foi desenvolvido em linguagem PERL para acessar um banco de dados MySQL, onde estão armazenadas as informações obtidas nas análises comparativas. O módulo de interface da biblioteca gráfica GD da linguagem PERL foi utilizado para a construção da ferramenta de visualização. A representação gráfica criada permite a navegação com opções de *zoom in*/*out*, disponibilizando as

informações de montagem, anotação das seqüências codificadoras e da organização das seqüências entre os genomas. Relatórios são ainda disponibilizados como fonte complementar da apresentação dos resultados.

O sistema GINGA foi utilizado para analisar de maneira comparativa o genoma das bactérias *Leifsonia xyli* subsp. *cynodontis* (Lxc – genoma parcialmente seqüenciado) e *Leifsonia xyli* subsp. *xyli* (Lxx – genoma completamente seqüenciado). Lxx provoca o raquitismo da soqueria em cana-de-açúcar, enquanto Lxc é capaz de colonizar cana-de-açúcar sem provocar sintomas de doença. O objetivo foi revelar, ainda durante o processo de seqüenciamento do genoma de Lxc, diferenças genéticas existentes entre os genomas dessas duas bactérias. Fizeram parte das análises comparativas um total de 9.754 *reads* do genoma de Lxc que formaram 1.064 *contigs* e 317 *scaffolds,* totalizando 1.470.731 de bases não redundantes. GINGA permitiu a identificação de 206.320 bases (~19%) em seqüências de *contigs* específicos (*contigs* que não apresentaram alinhamento algum com o genoma completo de Lxx) e 19 *scaffolds* (5,9%) que totalizaram 56.884 bases específicas ao genoma de Lxc, além de aproximadamente 1 milhão de nucleotídeos alinhados ao genoma de Lxx e pelo menos 6 grandes rearranjos.

Estes resultados foram disponibilizados em uma interface gráfica e relatórios, permitindo orientar o andamento do projeto de seqüenciamento do genoma de Lxc quanto à seleção das regiões a serem seqüenciadas e, simultaneamente, oferecendo informações para a formalização de hipóteses relevantes à biologia destes microorganismos.

Dissertation presented to fulfill the requirements of MCT/LNCC to obtain the Master´s Degree in Science (M.Sc.)

GINGA – GRAPHICAL INTERFACE FOR COMPARATIVE GENOME ANALYSIS: DEVELOPMENT OF A COMPUTATIONAL SYSTEM TO VISUALIZE THE COMPARATIVE OF BACTERIAL GENOMES IN A GRAPHICAL VIEW

Alexandre Rossi Paschoal

23th March 2007

Advisors: Claudia de Barros Monteiro-Vitorello
Ana Tereza Ribeiro de Vasconcelos

This study aimed to develop a computational system applied to the comparative analysis of prokaryotic genomes in a graphical view. The system named GINGA – *Graphical Interface for comparative Genome Analysis* – was developed to analyse a draft genome sequence in comparison to a complete genome. The system shows the alignment between sequence of *reads*, *contigs* and *scaffolds* from partial sequenced genomes and the complete sequence of another genome and allows the identification shared and unique regions as well as rearrangements. GINGA is a *web-based* system developed using the PERL language to access a MySQL database where all the information regard to the comparative analysis is stored. The module of the interface to GD (Graphics Library) was used to help the construction of the graphical tool. The graphical view allows zoom in/out on the information on assembly, annotation and the organization of the sequences. Supplementary information can be accessed in the form of reports.

GINGA system was used to compare the genomes of *Leifsonia xyli* subsp. *cynodontis* (Lxc – draft genome sequence) and *Leifsonia xyli* subsp. *xyli* (Lxx – complete genome sequence). The mail goal was to identify genetic differences that may help to understand the pathogeniciy of *Lxx* towards sugarcane. A total of 9.754 reads assembled in 1.064 contigs and 317 scaffolds produced 1.470.731 of no redundant bases of *Lxc* genome and were used in the analysis. GINGA allowed the identification of 206.320 bp (~20%) of *Lxc* specific sequences organized in contigs and 56.884 bp organized in 19 scaffolds (5,9%), around 1 milion bp aligned to Lxx genome and at least 6 large scale genomic rearrangements. These results were presented in a graphical interface and allowed to guide the partial genome sequencing, helping to decide which regions should be further sequenced and at the same time allowing the formulation of hypothesis related to important biological aspects of these microorganisms.

Sumário

Lista de Figuras

Lista de Tabelas

1 INTRODUÇÃO

Nos últimos anos, diversos organismos tiveram o seu genoma completamente seqüenciado e as informações obtidas encontram-se disponibilizadas em bancos de dados públicos. Geralmente, os organismos são estrategicamente selecionados e abrangem representantes de espécies que habitam os mais diversos nichos ecológicos. Os projetos de seqüenciamento de genomas têm a característica de gerar um grande volume de dados, e o desenvolvimento de métodos computacionais para organizar, armazenar e analisar a informação disponível é fundamental para a pesquisa em biologia e biotecnologia.

Hoje existem (acesso em janeiro de 2007) 444 genomas de bactérias e arqueobactérias completamente seqüenciados (*http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html*) e outros 1.092 projetos de seqüenciamento em andamento (*http://www.genomesonline.org/*). A comparação entre as seqüências de genomas de duas ou mais espécies pode ser utilizada para: (a) reconhecer regiões de similaridade em conteúdo e organização de genes; (b) estabelecer associações evolutivas; (c) ajudar a inferir a função biológica dos genes; e (d) identificar regiões contendo genes específicos a cada grupo de organismos que possam estar relacionados com seu estilo de vida [1,2].

Análises comparativas realizadas entre genomas de bactérias filogeneticamente próximas ou genomas de linhagens de uma mesma espécie revelaram que é possível detectar inserções, deleções e rearranjos entre longas

regiões colineares dos genomas [2]. Além de permitir identificar estas diferenças, as análises revelaram ainda que regiões específicas associadas à cada espécie ou linhagem são freqüentemente encontradas agrupadas nos genomas e, em muitos casos, não contêm mais do que 20% de genes específicos [1,3,4,5,6]. Estratégias de seqüenciamento parcial e a análise comparativa a um genoma totalmente seqüenciado de uma espécie próxima estão sendo desenvolvidas com o objetivo de maximizar o aproveitamento dos dados disponíveis em bancos de dados e, assim, reduzir os custos na obtenção da seqüência completa de um novo genoma.

Vários programas estão disponíveis para análises comparativas entre seqüências completas de genomas [7,8,9,10,11,12]. No entanto, somente o programa BACCardl [7] e o sistema SABIA [13] permitem visualização gráfica comparativa entre um genoma parcialmente seqüenciado e um genoma completo. No caso do BACardI, o uso deste programa fica restrito a projetos que foram estruturados em bibliotecas genômicas de insertos grandes [7]. O sistema GINGA está associado ao sistema SABIA.

O objetivo deste trabalho foi o de criar um sistema de visualização (representação) gráfica que permita o acompanhamento da montagem de um genoma parcialmente seqüenciado, de maneira comparativa a um genoma completo. O sistema, denominado de GINGA (***G****raphical* ***IN****terface for comparative* ***G****enome* ***A****nalysis*), foi desenvolvido com a finalidade de auxiliar na obtenção de informações sobre o conteúdo e a organização comparativa de genes entre dois genomas com ênfase em diferenças na estrutura do genoma (conteúdo de genes e organização do genoma) entre microrganismos filogeneticamente próximos.

As seqüências do DNA genomico das bactérias *Leifsonia xyli* subsp. *cynodontis* (Lxc) e *Leifsonia xyli* subsp. *xyli* (Lxx) foram utilizadas para testar e validar o sistema

computacional desenvolvido neste trabalho. Lxx é responsável pela doença raquitismo da soqueira (RSD - *Ratoon Stunting Disease*) na cultura de cana-de-açúcar. A doença pode ser encontrada em todas as áreas de cultivo da cana-de-açúcar, tendo sido responsável por perdas econômicas significativas no setor agrícola nos últimos anos [14]. Considerando a importância econômica do combate à esta doença [15], o genoma da Lxx foi completamente seqüenciado [16]. Lxc é encontrada colonizando gramíneas do gênero *Cynodon* mas, embora seja capaz de habitar o xilema, não provoca sintomas de doença em cana-de-açúcar. Poucas são as informações disponíveis sobre o genoma ou a biologia desta subespécie**.** A análise comparativa entre os genomas de Lxc e Lxx pode contribuir para o estudo sobre o comportamento diferencial destes organismos com relação ao hospedeiro, além de ser um excelente modelo biológico para validar o sistema computacional desenvolvido neste trabalho.

2 REVISÃO BIBLIOGRÁFICA

Desde a descrição da estrutura molecular do DNA por Watson e Crick, em 1953 [17], vários foram os avanços científicos e tecnológicos que permitiram a análise genômica de diversos organismos. Dentre esses, a metodologia descrita por Sanger e colaboradores permitiu o seqüenciamento de fragmentos de DNA tendo como base estratégias de polimerização *in vitro* do DNA [18]. Durante os anos seguintes, vírus, plasmídeos e fragmentos de DNA do genoma de diversos organismos tiveram a sua seqüência determinada. A automatização da técnica e a metodologia conhecida como *shotgun* (descrita na seção 2.1) [19], permitiram o seqüenciamento de DNA em grande escala e, em 1995, foi obtido o primeiro seqüenciamento completo do genoma de um organismo, o da bactéria *Haemophilus influenzae* [19].

Recentemente, novas metodologias de seqüênciamento têm surgido e permitido o sequenciamento de um genoma em poucas horas. Destacam-se duas técnicas em particular: (a) o método conhecido como *pyrosequencing,* que permite o seqüenciamento de pequenos fragmentos de DNA, sem a necessidade de clonagem ou construção de bibliotecas [20]; e (b) outro método, o qual usa um laser microscópico confocal, que foi capaz de "reseqüênciar" o genoma da E. coli em menos de um dia [21].

O seqüenciamento de genomas permite revelar todo o conteúdo de genes de um dado organismo. As espécies de bactérias escolhidas para projetos de seqüenciamento habitam os mais diferentes nichos, sendo que as de maior interesse

são em geral patógenos causadores de doenças em humanos e animais, bem como bactérias que vivem em condições extremas (pH, temperatura, radiação etc.) (*http://www.ncbi.nlm.nih.gov/genomes/static/eub.html*). Além disso, projetos de seqüenciamento em grande escala são de interesse dos mais diversos grupos de pesquisa em redor do mundo (*http://www.ncbi.nlm.nih.gov/genomes/static/lcenters.html*). No Brasil, o primeiro genoma de uma bactéria completamente seqüenciado foi o da *Xylella fastidiosa* [22]. O seqüenciamento do genoma desta bactéria marcou o início do estabelecimento de uma competência em genômica no Brasil que desencadeou o desenvolvimento de outros projetos-genoma tanto de bactérias como de outros organismos (Tabela 1).

Essa revisão de literatura tem como propósito apresentar informações sobre: (a) alguns aspectos do processo de seqüenciamento de genomas de bactérias relevantes a este projeto; (b) apresentar resultados encontrados em análises comparativas de genomas na procura de variações genéticas que possam estar associadas a fenótipos diferentes; (c) as ferramentas computacionais disponíveis para estas análises; e (d) as informações sobre a biologia do modelo biológico utilizado para validar o sistema desenvolvido neste trabalho.

Tabela 1: Relação dos projetos-genoma desenvolvidos no Brasil.

Projeto	*Site*	*Ref.*	*Consórcio*
Xanthomonas campestris pv. *campestris* e *Xanthomonas axonopodis* pv. *citri*	http://genoma4.iq.usp.br/xanthomonas	[23]	ONSA[2]
ESTs[1] de cana-de-açúcar	http://sucest.lad.dcc.unicamp.br/en	[24]	ONSA[2]
Leifsonia xyli subsp. *Xyli*	http://www.leifsonia.lncc.br	[16]	AEG[3]/ONSA[2]
Leptospira interrogans serovar copenhageni		[25,26]	AEG[3]/ONSA[2]
ESTs[1] de *Eucaliptos*	https://forests.esalq.usp.br/		Forests[19]/ONSA[2]
Coffea arabica[20]	http://www.lge.ibi.unicamp.br/cafe/	[27]	AEG[3]/ONSA[2]
ESTs[1] de *Schistosoma mansoni*	http://verjo18.iq.usp.br/schisto	[28]	ONSA[2]
Projeto Genoma Humano do Câncer[18]	http://watson.fapesp.br/cancer/outros.htm		ONSA[2]/Ludwig[4]
Chromobacterium violaceum	http://www.brgene.lncc.br/cviolaceum/	[29]	BRGENE[5]
Mycoplasma synoviae	http://www.brgene.lncc.br/finalMS/	[30]	BRGENE[5]
Mycoplasma hyopneumoniae J	http://www.genesul.lncc.br/finalMH/	[30]	BRGENE[5]
Mycoplasma hyopneumoniae 7448	http://www.genesul.lncc.br/finalMP/	[30]	GeneSul[6]
Mycoplasma hyopneumoniae 7422	http://www.genesul.lncc.br/	parcial	GeneSul[6]
EST[1] do fungo de *Paracoccidioides brasiliensis*			Rede do Centro-Oeste[7]
Herbaspirillum seropedicae	http://www.genopar.org/		GenoPar[8]
Trypanosoma cruzi			Programa Genoma do *Trypanosoma cruzi*[9]
Gluconacetobacter diazotrophicus	http://www.riogene.lncc.br/		RioGene[10]
Schistosoma mansoni			Rede Genoma do Estado de Minas Gerais
Leishmania chagasi	http://biolab.cin.ufpe.br/leishmania/leishmania.html		ProGeNe[11]
Crinipellis perniciosa	http://www.lge.ibi.unicamp.br/vassoura/		Rede Genoma do Estado da Bahia
EST[1] de *Anopheles darlingi*	http://www.darlingi.lncc.br/		Labinfo/LNCC[12], INPA[13], UFAM[14] e UNB[15]
Biological Nitrogen Fixation	http://www.bnf.lncc.br/		Labinfo/LNCC[12], Embrapa[16], UFPR[17]

[1] *Expressed Sequence Tags;* [2] *ONSA – Organization for Nucleotide Sequencing and Analysis;* [3] *AEG Project - Agronomical & Environmental Genomes;* [4] *Instituto Ludwig de Pesquisa sobre o Câncer;* [5] *BRGENE - Virtual Institute of Genome Research;* [6] *GeneSul - Southern Genome Investigation Program;* [7] *Projeto em Rede do Centro-Oeste;* [8] *GenoPar - Programa Genoma do Estado do Paraná;* [9] *Implantação no Instituto de Biologia Molecular do Paraná;* [10] *RioGene - Programa da Rede Genoma do Estado do Rio de Janeiro;* [11] *ProGeNe - Programa Genoma do Nordeste;* [12] *Laboratório de Bioniformática / Laboratório Nacional de Computação Científica;* [13] *Instituto Nacional de Pesquisas da Amazônia;* [14] *Universidade Federal do Amazonas;* [15] *Universidade de Brasília;* [16] *Centro Nacional de Pesquisa de Soja;* [17] *Universidade Federal do Paraná;* [18] *Human Cancer Genome Projec;* [19] *Eucaliptus Genome Sequencing Project Consortim;* [20] *The Brazilian Coffee Genome Project*

2.1 Seqüenciamento, montagem e anotação de genomas

O seqüenciamento completo de genomas foi possível depois de avanços tecnológicos que incluem a metodologia conhecida como *shotgun* [19] (Figura 1). A técnica de seqüenciamento descrita por Sanger e colaboradores [18] permite obter a seqüência de pequenos segmentos de DNA de até 800 nucleotídeos. Para o seqüenciamento de um genoma completo utilizando a estratégia de *shotgun*, é necessária a fragmentação do DNA e o seqüenciamento dos pequenos segmentos de maneira aleatória. Em seguida, programas computacionais são utilizados para fazer a sobreposição de seqüências (*reads*), num processo chamado de montagem (*assembly*), para a obtenção de uma seqüência consenso de bases contíguas ou *contigs* (conectadas sem quebras) (Figura 1). Os *contigs*, por sua vez, podem ser agrupados formando os *scaffolds* ou *super-contigs*, como descrito a seguir.

Assim, todo o processo de seqüenciamento de um genoma completo inclui as seguintes etapas: (a) extração do DNA das células de um organismo; (b) fragmentação do DNA extraído em pequenos pedaços de maneira aleatória; (c) construção de bibliotecas genômicas; (d) identificação da seqüência dos insertos clonados por seqüenciamento; e (e) a montagem das seqüências por programas de computador.

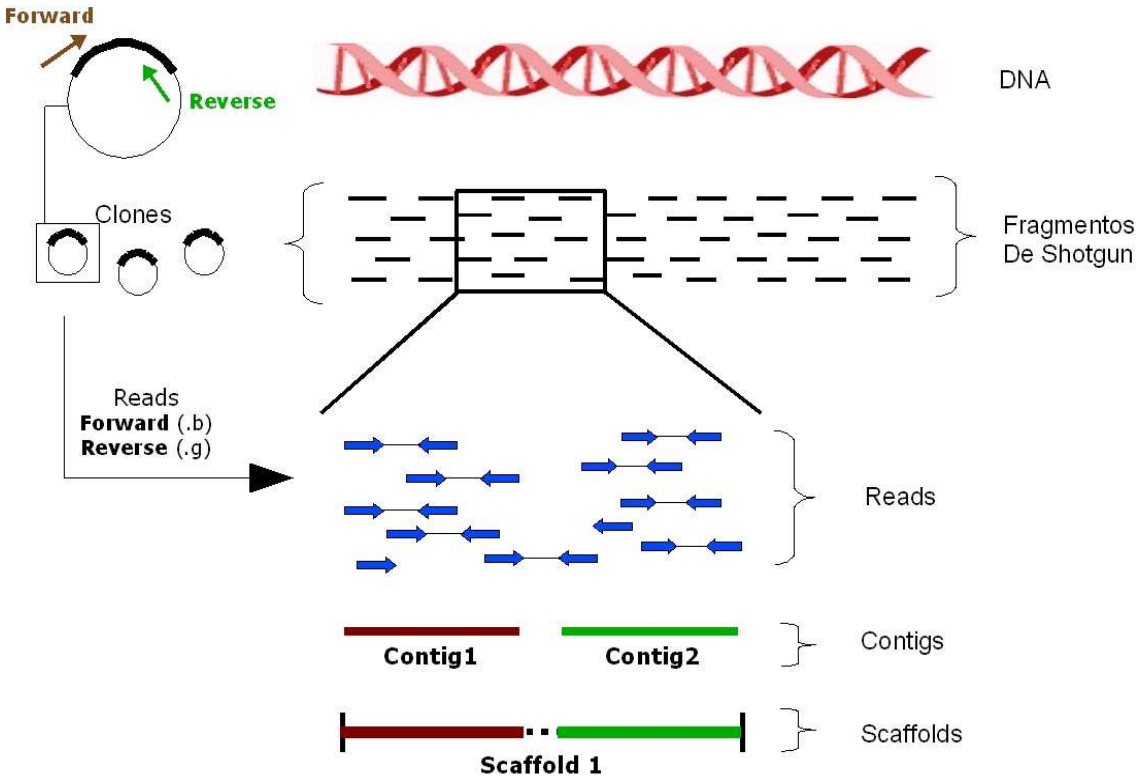


Figura 1: Representação esquemática da estratégia de Shotgun utilizada em projetos de seqüenciamento completo de genomas.

O tipo de biblioteca a ser utilizada em um projeto depende dos objetivos e do tamanho do genoma do organismo em questão. É comum a utilização de mais de um tipo de biblioteca para o mesmo projeto, que varia quanto ao tamanho dos fragmentos a serem clonados (Tabela 2) [31]. Para cada inserto clonado são geralmente obtidas duas seqüências (*read*s), uma de cada fita. Neste trabalho, os *reads* serão identificados como b (direto, da palavra em inglês *forward*) e g (reverso, da palavra em inglês *reverse*) dependendo do iniciador (*primer*) utilizado durante o processo de seqüenciamento. A correta identificação dessas seqüências é essencial durante o processo de montagem de *scaffolds* ou *super-contigs*. Os *contigs* são ligados "virtualmente" formando *scaffolds*, considerando as informações das seqüências .b e .g de um mesmo inserto (*reads* casados) (Figura 1).

Tabela 2: Tipos de bibliotecas genômicas que podem ser construídas em um projeto de seqüenciamento de genomas.

Vetor da biblioteca	***Tamanho do inserto clonado (em média)***	***Limite máximo no tamanho dos fragmentos clonados***
Plasmídeo	0.5 – 2 kb	~10 kb
Bacteriófago	7-10 kb	~20 kb
Cosmídeo ou fosmídeo	35-40 kb	~45 kb
BAC (*Bacteria artificial chromosome*)	80-120 kb	~200 kb
YAC (*Yeast artificial chromosome*)	200-800 kb	~1.5 Mb

Existem diversas ferramentas que realizam o processo de montagem: Phrap [32], CAP3 [33], TIGR Assembler [34,35], FAK [36,37], Staden [38] e STROLL [39,40] dentre outras. Cada uma utiliza algoritmos diferentes para obter a seqüência contígua de DNA. De maneira geral, os programas seguem o modelo que utiliza um algoritmo de programação dinâmica, como Smith-Waterman [41], para fazer a alinhamento das seqüências comuns. Os programas Phrap e CAP3 foram utilizados durante o desenvolvimento deste projeto e a função de cada um deles será discutida a seguir.

O Phrap (*Phragment Assembly Program*) está entre os programas mais utilizados para montagem de genomas e faz parte de um pacote de programas, denominado de Phred/Phrap/Consed [32,42,43,44], distribuído sem custos para fins acadêmicos. As principais etapas de cada um desses programas é descrita a seguir: (Figura 2):

(a) O programa **Phred** faz a leitura dos arquivos cromatogramas e atribui qualidade às bases. O cálculo da qualidade atribuída a cada base é dado pela fórmula $Q = -10 \log_{10} (Pe)$, onde Q e Pe são, respectivamente, o valor da qualidade e a probabilidade da base ter sido nomeada erroneamente. Por exemplo, Q = 20 significa que se tem a probabilidade de 1 (uma) base em 100 (cem) ser nomeada erroneamente $p = 10^{-2}$ (p = 0.01). Outro exemplo, Q = 40 significa que se tem a probabilidade de 1 (um) erro em 10.000 (dez mil) bases $p = 10^{-4}$ (p=0.0001);

(b) o programa **cross_match** identifica e filtra a seqüência de vetores e repetições;

(c) o programa **Phrap** faz a montagem da seqüência utilizando, dentre outros fatores, o critério de qualidade de base gerado pelo Phred. Através da sobreposição das bases, gera a seqüência de consenso de um *contig;*

(d) e o programa **Consed** permite a visualização da montagem e a edição das bases.

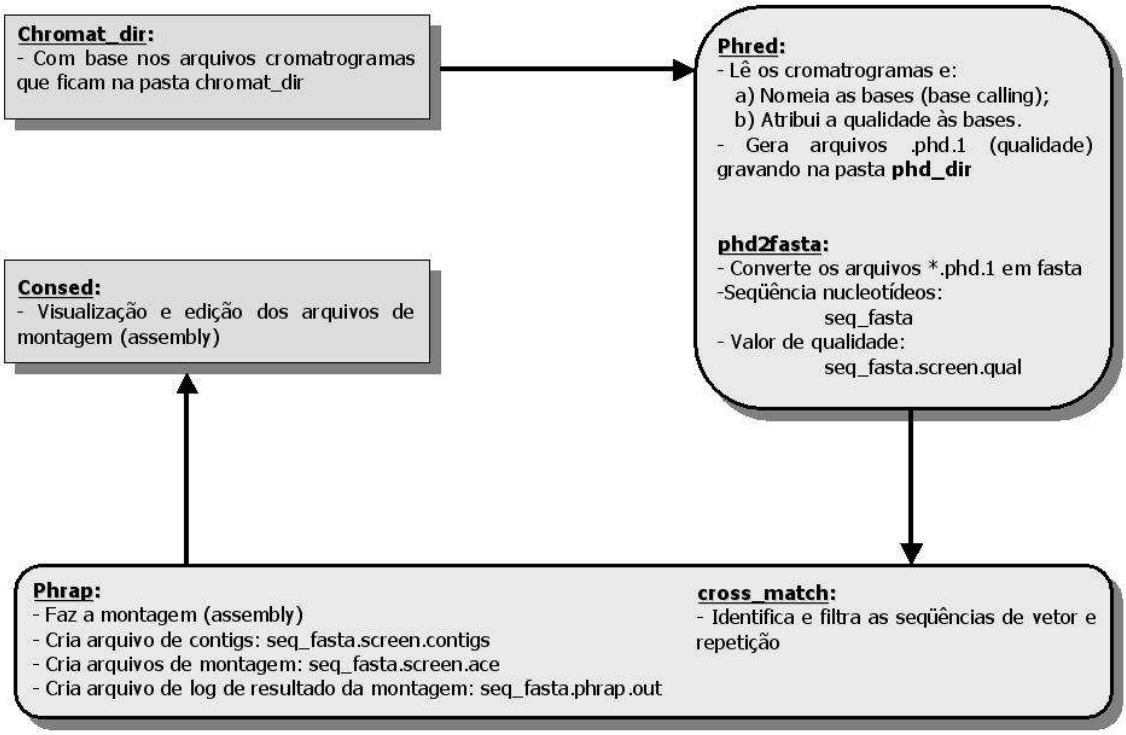


Figura 2: *Pipeline* de execução do pacote *Phred/Phrap/Consed*.

O programa CAP3 é um derivado do programa CAP (*Contig Assembly Program*) [45] e exerce função semelhante àquelas descritas pelo programa Phrap. Esse programa pode ser utilizado em associação com os programas Phred e Consed. A diferença entre os programas CAP3 e Phrap está no algoritmo utilizado para realizar a alinhamento das seqüências. O programa CAP3 considera a distância entre os *reads* .b e .g para posicionar as extremidades do inserto em relação à seqüência de consenso do *contig* [45].

2.2 A anotação de genomas de procariotos

A anotação de um genoma é o processo que compreende a predição e a localização (coordenadas) dos genes na seqüência de DNA e a associação de informações biológicas à seqüência desses genes quando disponíveis, como será discutido a seguir.

Os genes contêm seqüências de nucleotídeos que são compostas por regiões reguladoras e uma região codificadora (CDS – *coding sequence*). A busca pela CDS em uma seqüência desconhecida é feita por meio da verificação das possíveis ORFs (*Open Read Frames*). Uma ORF é um trecho de DNA que possui *codons* de iniciação, terminação e com uma seqüência de nucleotídeos do tamanho múltiplo de três. As ORFs identificadas por programas de predição podem representar a região codificadora de uma possível proteína . Existem ferramentas computacionais que tem como objetivo localizar possíveis genes, utilizando métodos matemáticos e estatísticos, tendo como base algoritmos de *Hidden Markov Model* (HMM) [46], heurística ou uma combinação desses com outros métodos. Entre os programas mais utilizados estão GLIMMER [47], GeneMark [48,49] e EasyGene [50].

A etapa seguinte à predição dos genes na seqüência de DNA é a comparação da seqüência de nucleotídeos e da seqüência de aminoácidos das proteínas codificadas por esses genes com seqüências depositadas em bancos de dados. A busca por similaridade tem como principal objetivo identificar seqüências similares cuja função já tenha sido descrita experimentalmente.

Existem vários bancos de dados de acesso livre que agrupam informações que atendem as mais diversas questões biológicas. Os bancos de dados mais completos que armazenam seqüências de DNA e de proteínas são: o GenBank [51,52]

(*http://www.ncbi.nlm.nih.gov/Genbank/*), o DDBJ (*DNA DataBank of Japan*) [53,54], e o EMBL (*European Molecular Biology Laboratory*) [55]. Esses bancos de dados são integrados (*http://www.insdc.org/*) e acumulam hoje (dado de Janeiro de 2007) mais de 100 gigabases contidos em seqüências depositadas nas divisões tradicionais e de WGS (*whole genome shotgun sequence)*. Outros bancos de dados estão disponíveis, como, por exemplo, KEGG (*Kyoto Encyclopedia of Genes and Genomes – http://www.genome.jp/kegg/*) [56,57,58] que contém as informações de seqüências e organiza as proteínas em vias metabólicas. Os bancos de dados ProDom (*Database of Protein Domain Families – http://prodom.prabi.fr/*) [59], Pfam (*Protein Families Database - http://www.sanger.ac.uk/Software/Pfam/*) [60] e SMART (*Simple Modular Architecture Research Tool - http://smart.embl.de/*) [61] disponibilizam informações de domínios e de outras regiões conservadas nas proteínas, agrupando-as em famílias e de acordo com a sua função biológica. Existem, ainda, outros bancos que utilizam a anotação manual para acurar as informações biológicas inferidas às proteínas por anotação automática, como é o caso do banco de dados do Swiss-Prot/UniProt (*Swiss-Prot Protein Knowledgebase* / *Universal Protein Resource - http://www.ebi.ac.uk/swissprot/*) [62].

Para buscar as informações biológicas disponíveis, as seqüências de interesse são comparadas com aquelas depositadas nos bancos de dados, utilizando os programas BLAST (*Basic Local Alignment Search Tool*) [63], FASTA ou variantes [64]. O programa BLAST [63] é uma ferramenta computacional criada em 1990 [63] e muito utilizada devido à sua rapidez na obtenção dos resultados. O BLAST utiliza matrizes de valores (*score*), como a matriz BLOSUM62, para procurar pelo mais alto valor (grau de *score*) no alinhamento da seqüência de interesse (*query)* contra as seqüências de banco de dados (*subject*).

Existem sistemas que auxiliam no processo de montagem e anotação para

projetos de seqüenciamento de genomas [65]. Um desses sistemas é o SABIA (*System for Automated Bacterial Integrated Annotation*) [13], que é capaz de montar e anotar genomas de bactérias, ESTs e eucariotos. Além de ser utilizado em projetos de seqüenciamento de genomas, o SABIA pode ser usado em projetos de re-anotação de genomas [13]. Os programas que constituem este sistema fornecem, por intermédio de interfaces gráficas, várias informações e estatísticas sobre a montagem (seqüências e sua qualidade, *contigs*, *scaffolds*) que auxiliam no fechamento de *gaps.*[1] As informações da anotação são disponibilizadas para o usuário e fazem referências cruzadas com vários dos mais importantes bancos de dados biológicos. Uma categorização funcional automática é gerada, bem como as regiões regulatórias e de possíveis *operons*. Entre todos os sistemas desenvolvidos, o SABIA é o único que integra dados de montagem, anotação e comparação de genomas. O sistema descrito neste trabalho faz uso do sistema SABIA no que se refere às informações de montagem e anotação, como será descrito em Resultados e Discussões.

2.3 Análise comparativa de genomas de procariotos

Existem hoje (acesso em Janeiro de 2007), disponíveis publicamente, a seqüência completa de 414 genomas de bactérias (*http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html*), e ainda, mais de 1.034 outros projetos (*http://www.genomesonline.org/*) de seqüenciamento em andamento. Na maioria dos casos são genomas completos de espécies patogênicas de humanos e animais. Além de novas espécies, várias linhagens de uma mesma espécie de bactéria apresentam o genoma completamente seqüenciado, revelando uma diversidade pouco conhecida anteriormente [1]. Dentre as análises comparativas

[1] Espaços gerados por seqüências não conhecidas

descritas na literatura, as análises comparativas entre genomas de variantes patogênicos e não patogênicos de uma mesma espécie são de grande interesse científico. São vários os exemplos onde as diferenças nestes casos são devidas à presença e/ou ausência de genes agrupados em um ou outro genoma [2]. Em *Staphylococcus aureus*, por exemplo, ORFs específicas que potencialmente codificam fatores de patogenicidade representam 6% dos genes encontrados no genoma, enquanto que para *S. epidermidis* e *S. haemolyticus* constitui 2% [66]. Em *Pseudomonas*, dos 298 genes (3,3%) identificados como associados a patogenicidade, 96 (1,7%) são específicos de *Pseudomonas syringae* pv. *tomato* DC3000 (PTO) referente aos encontrados nas espécies *Pseudomonas aeruginosa* e *Pseudomonas putida* [67], indicando que possam codificar proteínas que sejam necessárias a interação entre PTO e o seu hospedeiro [68].

Dentre as linhagens de *Escherichia coli* cujo genoma foi completamente seqüenciado, as diferenças encontradas foram maiores. Por exemplo, existe um total de 1.34 Mb específicos à linhagem patogênica EDL933 de *E.coli* O157:H7 e 0.53 Mb específicos à linhagem não patogênica MG1655 (K12) que estão agrupadas em 411 regiões com mais do que 50 Kbp (177 regiões específicas de O157:H7 e 234 específicas de K12) [69]. A análise destas regiões revelou genes diretamente ligados a patogenicidade da linhagem O157:H7. Outras 18 linhagens de *E. coli* tiveram seu genoma completamente ou parcialmente seqüenciados, revelando que em cada linhagem existem pelo menos 100 genes específicos [1].

Entre duas linhagens (J99 e 26695) de *Helicobacter pylori,* 6 e 7,5% dos genes foram identificados como específicos a cada uma, sendo que metade deles estão agrupados em uma única região no genoma [3]. As linhagens (CDC1551 e H37Rv) de *Mycobacterium tuberculosis* mostraram uma variação de cerca de 10% [6]. A análise comparativa entre os genomas das linhagens J e 7448 de *Mycoplasma*

hyopneumoniae revelaram uma região de 22,3 kb específica da linhagem 7488 [30], que é similar a um elemento conjugativo (*ICEF - integrative conjugal element*) de *Mycoplasma fermentans* [70]. O ICEF é um elemento conjugativo sem função definida que está presente no genoma de *M. fermentans* [70]. Uma inversão de 243.104 pb foi identificada no genoma da linhagem 232 de *M. hyopneumoniae* em relação às linhagens J e 748 [30]. Entre os genomas das estirpes de *Xylella fastidiosa*, XF-PD [71] e XF-9a5c, foram encontrados 51 genes específicos (2,47%) em XF-PD e 152 em XF-9a5c (6,78%) [72]. As bactérias *X. fastidiosa* pv. *almond* (XF-Dixon) e *X. fastidiosa* pv. *oleander* (XF-Ann-1) são agentes causadores de doença na amendoeira e no loureiro rosa*,* respectivamente [73,74]. Na comparação entre esses dois genomas parcialmente seqüenciados com o genoma completo da XF-9a5c, foram encontradas 133 ORFs específicas de *X. fastidiosa* pv*. almond* (XF-Dixon) e 188 ORFs específicas de *X. fastidiosa* pv. *oleander* (XF-Ann-1) [72,73]. Entre as duas linhagens de *Xanthomonas campestris* pv. *campestris* 8004 (XCC-8004) e XCC-ATCC-33913 [23] foram encontrados 108 e 62 genes específicos, respectivamente [75].

Em algumas espécies do gênero *Bordetella* foram encontrados 114 genes específicos de *B. pertussis* quando comparado aos genomas de *B. bronchiseptica* e *B. parapertussis*, e somente 50 genes específicos de B*. parapertussis* comparado às genomas de *B. bronchiseptica* e *B. pertussis* [76]. Nos genomas de *Bacillus halodurans* e *Bacillus subtilis* [5], cerca de 35% dos genes são específicos à cada genoma. Entre *Neisseria meningitidis* e *Neisseria gonorrheae,* oito ilhas genômicas foram identificadas, que variam, em tamanho, entre 1,8 kbp até 40 kbp. A definição dessas ilhas permite distinguir as duas espécies em relação à sua patogenicidade [77]. Outros exemplos ainda trazem os profagos, DNA viral integrado ao genoma de bactérias, em regiões de até 40 Kbp como associados à virulência, como é o caso das toxinas produzidas por *Vibrio cholerae* [78] e *Pseudomonas aeruginosa* [79].

Em muitos dos casos citados anteriormente, os genes específicos estão agrupados em regiões do genoma conhecidas como ilhas genômicas [1,2,80]. Estas foram primeiramente descritas como regiões contendo fatores de virulência nos cromossomos da bactéria *E. coli* uropatogênica [81,82].

As ilhas genômicas são muitas vezes encontradas no cromossomo de algumas linhagens de bactérias, estando ausentes em linhagens ou organismos de uma mesma espécie ou de espécies próximas. Essas regiões são consideradas importantes para os processos de adaptação e evolução das bactérias, podendo ser responsáveis por alterações significativas nos seus fenótipos (por exemplo, a mudança na patogenicidade) [83]. Elas podem ser identificadas por uma variação no conteúdo de bases GC, que é diferente da média definida para aquele genoma, ou por viés na utilização dos códons. Muitas vezes estas ilhas estão integradas dentro ou próximas a um tRNA, podendo conter elementos de inserção (seqüências de inserção, *transposons*, *integrases*, *recombinases*), seqüências repetidas e invertidas nas extremidades [84]. Normalmente, variam entre os tamanhos de 10 e 500 Kb. Quando contam genes associados à virulência são denominadas de ilhas de patogenicidade [81] e já foram encontradas em uma grande variedade de bactérias patogênicas de plantas e animais [84,85,86].

Essas regiões podem ser adquiridas em eventos de transferência horizontal (HGT – *Horizontal Gene Transfer*) com o auxílio de elementos de inserção, e potencialmente, podem manter a sua capacidade de transferência. No entanto, muitas vezes os genes associados à integração são perdidos, e essas regiões tornam-se permanentemente ancoradas no genoma. Ilhas genômicas podem conter outros genes que não estão associados à virulência. Como exemplos, tem-se *Shigella flexinery,* em cujo genoma foram encontrados genes de resistência a antibióticos agrupados nessas

regiões [87] e *Mezorhizobium loti,* onde foi descrita a presença da ilha de simbiose contendo genes associados ao processo de fixação de nitrogênio [88].

Entre os elementos de inserção, que estão envolvidos em eventos de HGT, estão os bacteriófagos, os *transposons* e as seqüências de inserção [83]. Bacteriófagos são vírus que infectam células de bactérias podendo inserir o seu material genético no cromossomo bacteriano. Este tipo de evento de HGT é denominado de transdução, e um exemplo importante desta é a aquisição de genes que codificam toxinas associadas à patogenicidade de *Vibrio cholerae* [78] e *Pseudomonas aeruginosa* [79]. As seqüências de inserção (IS) e *transposons* são pequenos segmentos de DNA transponíveis, podendo estar presentes em muitas cópias no genoma de bactérias. Devido à sua presença no genoma em um grande número de cópias, estão associados a eventos de rearranjos, servindo como sítios para recombinação por homologia [89,90]. Por exemplo, em *Leptospira interrogans*, as diferenças encontradas entre os genomas de dois *serovares* (*lai* e *conpenhagni*) são devidas principalmente à variação encontrada nesses elementos, onde a localização de IS coincidem com as regiões do genoma que sofreram grandes rearranjos [25,26].

Existe uma variação genética bastante significativa entre os genomas de bactérias, que pode ser encontrada utilizando ferramentas computacionais em análises comparativas. Vários são os exemplos de ferramentas que vêm sendo utilizados para análise de genomas e que serão discutidos a seguir.

2.4 Ferramentas para análise comparativa de genomas

Diversos são os métodos disponíveis para analisar genomas de procariotos. A escolha dos sistemas e dos bancos de dados que serão utilizados em cada projeto

está diretamente ligada aos seus objetivos. Entre as ferramentas disponíveis [65] para as análises de genomas de procariotos, existem aquelas que permitem o acompanhamento do processo de montagem (SABIA [13]) e os que permitem analisar dados de anotação e apresentam uma interface para visualização gráfica (Artemis [91], BASys [92], GenDB [93], MAGPie [94], SABIA [13]). Outras ferramentas ainda permitem o alinhamento local ou global de seqüências de genomas (BLAST[63], cross_match – sem publicação, MUMmer [95,96,97], AVID [98]). Por fim, outras foram desenhadas especificamente para analisar comparativamente genomas de bactérias, permitindo a análise das seqüências e de anotação [7,8,9,10,11,12]. Esses últimos métodos são os que têm objetivos comuns com o sistema desenvolvido por este trabalho (Tabela 3).

Tabela 3: As principais características das ferramentas para análise comparativa de genomas.

Nome	***Programa de Alinhamento***	***Informações de montagem***	***Informações de anotação***	***Tipo***[1]	***Tecnologia***	***Uso do programa***	***Ref.***
BACCardl	BLAST	X		gc/gp	Perl	Local	[7]
COMBO	BLAST, PatternHunter		X	gc	Java	Local/via Web	[8]
ACGT	BLAST, MSPcrunch		X	gc	Java, BioJava	Local	[9]
ACT/WebACT	BLAST MUMmer		X	gc	Java	Local/via Web	[10,11]
GenAlyzer	Vmatch		X	gc	C	Local	[12]

[1]gc – permite a análise comparativa de genomas completetamente seqüências dos (gc); ou parcialametne seqüenciados (gp)

2.4.1 BACCardl

O sistema BACCardl [7] é uma ferramenta que faz o mapeamento virtual das seqüências dos fragmentos de DNA clonados com base na seqüência completa de um genoma de uma espécie próxima. Os principais objetivos são permitir a validação da montagem de um genoma, orientar a ordenação dos *contigs* na montagem e

disponibilizar uma comparação com o genoma da espécie próxima. A análise é feita utilizando insertos grandes (mais de 20 Kbp), clonados em vetores do tipo cosmídeos/fosmídeos ou BACs (*Bacterial Artificial Chromosome*). O programa disponibiliza as informações de *gaps* e repetições que podem ser visualizadas na comparação entre as seqüências. A visualização gráfica do mapeamento virtual dos clones é disponível em dois formatos: (a) visualização circular do genoma e (b) visualização linear. Os *contigs* são ordenados utilizando uma variação do algoritmo *greedy path-merging* [99] para construção de *scaffolds*. Arquivo do tipo *ace* (gerado por programas de montagem como o Phrap [32] e o CAP3 [33]) contendo informações de montagem (*reads* e *contigs*) é o requerimento para a entrada no sistema BACCardl.

2.4.2 COMBO

O COMBO [8] é um programa integrante ao sistema *Argo Genome Browser* (*http://www.broad.mit.edu/annotation/argo/*) que mostra o alinhamento entre as seqüências de genomas com as informações de anotação. O objetivo do COMBO é mostrar, de forma gráfica, as informações de alinhamento local entre as seqüências de genomas completos em duas formas de visualização: a perpendicular e a paralela. A visualização perpendicular, também conhecida como *dot plot*, tem um visual semelhante ao programa o MUMmer [95,96,97]. Essa visualização permite ter uma visão global do alinhamento entre os genomas, marcando as regiões de alinhamento como pontos no gráfico. Ao final, tem-se uma linha ou linhas que demarcam as regiões de alinhamento, permitindo uma visualização global da comparação. A visualização paralela apresenta os genomas como linhas paralelas (horizontais) e as regiões alinhadas são representadas por linhas verticais. O programa aceita dados de seqüência em arquivos no formato FASTA e de anotação em arquivos no formato GFF (*http://www.sanger.ac.uk/Software/formats/GFF/*).

2.4.3 ACGT

O sistema ACGT [9] – *a comparative genomics tool* – mostra uma visualização global da comparação de genomas entre seqüências de até 2 milhões de bases. Esse sistema utiliza a visualização paralela para mostrar *clusters* de genes ortólogos, o que ajuda na compreensão da organização do genoma. O arquivo de entrada para a utilização nesse sistema pode estar nos formatos Genbank, EMBL ou FASTA. O programa aceita arquivo CMP que contém informações de anotação. O arquivo no formato CMP (*http://www.sanger.ac.uk/Software/Alfresco/manual/#cmp*) foi desenvolvido para ser capaz de importar informações do banco de dados EMBL [55].

2.4.4 ACT/WebACT

ACT [10] - *the Artemis comparison tool* – é um sistema para visualização gráfica da comparação entre a seqüência completa de genomas com informações de anotação, com o objetivo de mostrar as regiões de similaridades, rearranjos e inserções, e os alinhamentos entre os pares de bases. O sistema aceita genomas de organismos procariotos e eucariotos com tamanho de aproximadamente cinco milhões de pares de base. O ACT usa componentes do programa Artemis [91] para o sistema de anotação disponibilizando informações sobre os genes como, por exemplo, a orientação da transcrição. O ACT tem opção de *zoom-in* e *zoom-out* em sua visualização gráfica e disponibiliza outras informações, como porcentagem de bases guanina (G) e citosina (C), assinatura dinucleotídica e *codon bias*. O sistema permite salvar os alinhamentos em formato de imagem PNG ou JPEG para uso em produções científicas. Os arquivos de entrada para o sistema podem estar no formato do EMBL, Genbank, GFF ou FASTA.

2.4.5 GenAlyzer

GenAlyzer [12] é uma ferramenta construída para visualização de similaridade entre seqüências. O objetivo é mostrar a visualização do alinhamento (*match*), exato ou aproximado, entre dois tipos de seqüências (DNA ou proteína). GenAlyzer é uma versão aprimorada do programa REPuter [100,101] e seu visualizador REPvis [102]. REPuter é um programa para busca por repetições entre seqüências de DNA. GenAlyzer consegue analisar o alinhamento de seqüências com até dez milhões de pares de bases [12]. O tamanho do alinhamento (*match*) é dado pela cor da linha que liga as partes alinhadas apresentadas na vertical ou diagonal e as seqüências dispostas na forma horizontal ou paralela. O arquivo de entrada de seqüências pode estar no formato EMBL, Genbank, ou FASTA. O sistema GenAlyzer gera um arquivo texto de alinhamento da seqüência (*match file*) a partir do programa Vmatch (*http://www.vmatch.de*). Este tem as opções para o alinhamento de DNA ou de proteína, podendo ser utilizado com informações de genomas de organismos procariotos ou eucariotos. Permite que usuário especifique informações de anotação (cauda poli A, região promotora etc) da seqüência por intermédio de símbolos. Essas informações em símbolos são lidas a partir de um arquivo texto que pode ser criado pelo próprio usuário. GenAlyzer também aceita arquivos com o resultado dos programas GENSCAN [103], RepeatMasker (*sem publicação – http://www.repeatmasker.org/*) como informação de anotação.

2.5 Modelo biológico utilizado como estudo de caso

As bactérias da espécie *Leifsonia xyli* estão divididas em duas subespécies: *Leifsonia xyli* subsp*. xyli* (Lxx) e a *Leifsonia xyli* subsp*. cynodontis* (Lxc). Lxx causa raquitismo da soqueira ou RSD (*Ratoon Stunting Disease*) em cana-de-açúcar e Lxc

retarda o crescimento meristemático em gramíneas do gênero *Cynodon* (capim Bermuda ou grama seda) [104]. As duas subespécies colonizam os vasos xilemáticos de cana-de-açúcar, no entanto somente Lxx é capaz de provocar sintomas de doença [104,105,106].

Inicialmente reconhecidas como subespécies dentro da espécie *Clavibacter xyli* [104], foram re-classificadas como pertencentes ao gênero *Leifsonia*, juntamente com *L. poae*, encontrada em raízes infectadas de plantas *Poa annua*, e *L. aquatica*, uma bactéria de vida livre [104,107,108,109].

No Brasil, a cana-de-açúcar é uma das principais culturas no agronegócio, responsável por 2,4% do PIB nacional [110]. Do total produzido, 242,16 milhões de toneladas (50,9%) destinam-se à fabricação de açúcar, 183,82 milhões (38,6%) à produção de álcool e o restante, 49,74 milhões (10,5%), à fabricação de cachaça, alimentação animal, sementes, fabricação de rapadura, açúcar mascavo e outros fins [110]. O Brasil, além de ser o maior produtor mundial de açúcar, também é o maior produtor de etanol para a sua utilização como combustível [15].

O raquitismo da soqueira é encontrado em todas as áreas de cultivo da cana-de-açúcar, causando prejuízos anuais de 5 a 15% em plantios. Os sintomas de raquitismo aparecem com o encurtamento dos colmos, diminuindo a produtividade da cultura ao longo dos anos com os cortes sucessivos das socas. Perdas de aproximadamente US$ 36 milhões nas safras de 1988-89 na Flórida (EUA) [111] e perdas anuais de US$ 11 milhões na Austrália [112] foram descritas. No Brasil, acredita-se que o Estado de São Paulo, que produz US$ 8 bilhões, tenha perdido US$ 2 bilhões nos últimos 30 anos [113].

Em 2004, o genoma do isolado CTCB07 de *Leifsonia xyli* subsp. *xyli* foi

totalmente seqüenciado por laboratórios de rede AEG/ONSA/Fapesp [16]. Lxx é uma bactéria *gram*-positiva que contém um único cromossomo circular com 2.584.158 pares de base, conteúdo GC de 68%, onde foram preditos 2.044 genes. Quatro regiões de ilhas genômicas foram definidas com base nas diferenças encontradas no conteúdo de bases GC. Essas regiões contém genes potencialmente associados à patogenicidade, vários elementos transponíveis, profagos e genes normalmente encontrados em plasmídeos.

Leifsonia xyli subsp. *cynodontis* é encontrada nos vasos xilemáticos em gramíneas do gênero *Cynodon* (grama seda). Pouco se sabe sobre as características do genoma dessa bactéria, a não ser pela presença de um plasmídio criptico de 51 pb [114]. Além disso, faltam estudos sobre a população desta bactéria na cana-de-açúcar. A análise comparativa dos genomas de Lxc e Lxx tem como objetivo a busca de genes e regiões específicas a cada genoma que possam ajudar a entender o comportamento diferencial dessas espécies com relação à hospedeira cana-de-açúcar.

O genoma de Lxc está sendo parcialmente seqüenciado pelo mesmo grupo de pesquisa e colaboradores ao qual esse projeto pertence. As bibliotecas genômicas utilizadas nesse processo foram estrategicamente escolhidas, com o objetivo de realçar as diferenças entre os dois genomas. Foram construídas três bibliotecas de *shotgun* com insertos que variam entre 1-2 Kbp e 2-4 Kbp, uma biblioteca utilizando a tecnologia de hibridização subtrativa (*Suppression Subtractive Hybridization*) [115] e uma biblioteca genômica de insertos grandes clonados em BAC. A biblioteca de subtração foi construída utilizando sistema *PCR-Select Bacterial Genome Subtraction* (*Clontech*). Após a subtração, o resultado foi uma biblioteca de produtos de PCR, enriquecida de segmentos específicos de Lxc. As seqüências

obtidas foram alinhadas utilizando os programas Phred/Phrap e as informações de montagem foram utilizadas como entrada para o sistema de análise comparativa desenhada neste projeto.

3 OBJETIVOS

3.1 Objetivo geral

O objetivo geral deste trabalho foi o de criar um sistema de visualização (representação) gráfica para permitir o acompanhamento da montagem de um genoma parcialmente seqüenciado, de maneira comparativa a um genoma completo. O sistema foi criado principalmente para entender as diferenças entre os genomas de *Leifsonia xyli* subsp. *cynodontis* e *Leifsonia xyli* subsp. *xyli*. Contudo, o sistema aceita o uso de outros genomas.

3.2 Objetivos específicos

Os objetivos específicos para o desenvolvimento deste sistema foram os seguintes:

A. construir um sistema de análise e visualização gráfica para a comparação entre a seqüência de genomas, sendo um completamente seqüenciado e outro em processo de seqüenciamento;
B. integrar dados de montagem e visualização gráfica das seqüências que compõem cada *contig* e de *contigs* que compõem cada *scaffold*, de maneira comparativa ao genoma completo, sendo que não existam limitações quanto ao tipo e número de bibliotecas genômicas utilizadas;

C. fornecer subsídios para o seqüenciamento parcial de um genoma tendo como base a comparação com um genoma completo, possibilitando decisões sobre fragmentos de interesse que devem ser seqüenciados;
D. disponibilizar a representação gráfica de regiões comuns, assim como rearranjos, deleções, inserções e repetições;
E. disponibilizar a representação gráfica de regiões específicas ao genoma sendo seqüenciado;
F. disponibilizar as informações da integração com dados de anotação; e
G. apresentar as características especiais do genoma, como, por exemplo, conteúdo de bases GC, ilhas genômicas, presença de IS e introns.

4 METODOLOGIA

4.1 Ferramentas utilizadas na implementação do sistema

O sistema GINGA foi primeiramente e principalmente desenhado para ser utilizado em associação com os módulos de montagem (*Assembly*) e anotação (*Annotation*) do sistema SABIA [13]. O sistema GINGA utiliza as informações (montagem e anotação) organizadas pelo SABIA e um programa de alinhamento (cross_match) para fazer a comparação entre os genomas e a representação gráfica. O desenvolvimento desse sistema foi feito com a linguagem de programação PERL (*http://www.perl.org*) versão 5.6.1 e com o banco de dados MySQL (*http://www.mysql.com*) versão 3.23.46, sendo que o uso dessas tecnologias facilitou a integração com o sistema SABIA. O sistema é executado em um servidor SunOS 5.8 com o servidor *web Apache* 2.0. A representação gráfica foi construída utilizando a biblioteca gráfica GD da linguagem PERL, extraída do repositório de módulos desta linguagem denominado CPAN (*Comprehensive Perl Archive Network – http://www.cpan.org*). Todo o relatório de resultado de BLAST [63] dos *contigs* da montagem foi feito utilizando módulos do projeto Bioperl [116]. O acesso ao sistema GINGA foi centralizado em um portal denominado Portal GINGA. Esse portal unifica as funções do sistema por meio de uma *interface web*. No caso do modelo biológico, a anotação do genoma parcial de Lxc foi feita por meio do uso do SABIA, e a montagem com os programas Phrap [32] do pacote Phred/Phrap/Consed e CAP3 [33] também por intermédio do sistema SABIA, utilizando os parâmetros padrões desses programas. Para a construção dos *scaffolds* utilizou-se o programa *genscaff* [117].

4.2 Ferramentas utilizadas no alinhamento das seqüências

O programa *cross_match* é uma implementação do algoritmo de Smith–Waterman–Gotoh [41,118] desenvolvido por Phil Green, e foi utilizado para fazer o alinhamento entre as seqüências. Foram utilizados os parâmetros padrões do *cross_match*, com exceção do *masklevel,* sendo de 101, o que faz com que todos os alinhamentos (*match*) sejam apresentados (*http://bozeman.genome.washington.edu/phrap.docs/phrap.html*). De acordo com os parâmetros do *cross_match*, uma região alinhada tem que conter no mínimo 14 nucleotídeos contínuos, independentemente da quantidade de inserções/deleções/substituições do resto da seqüência.

Os resultados dos alinhamentos foram analisados por *scripts* (programas) do GINGA, que recuperam as informações de blocos comuns, regiões específicas, repetições e sobreposições entre as seqüências dos genomas. O reconhecimento dessas regiões a partir dos resultados de alinhamento formou a base do sistema de visualização do GINGA. O alinhamento sempre foi feito entre seqüências do genoma parcial (*scaffolds* e *contigs* isolados) contra a seqüência do genoma completo.

5 RESULTADO E DISCUSSÃO

5.1 Implementação do sistema

Fazem parte da implementação do sistema: (a) a montagem e a anotação dos genomas, por intermédio da integração ao sistema SABIA (item 5.1.1); (b) a análise e o tratamento das informações resultantes do alinhamento das seqüências dos genomas (item 5.1.2); (c) o armazenamento dessas informações em um banco de dados (item 5.1.3); (d) a criação de um Portal de acesso ao sistema e (e) o desenvolvimento de uma representação gráfica e a apresentação de relatórios (ambos no item 5.1.4). Esses serão os itens apresentados a seguir.

5.1.1 Montagem e anotação de genomas

5.1.1.1 Integração com o sistema SABIA

Para entender o uso do sistema GINGA com o suporte do sistema SABIA é necessário entender a divisão dos módulos do SABIA.

O sistema SABIA é dividido em dois módulos: (a) *Assembly* e (b) *Annotation*. Aquele oferece um suporte computacional para fornecer informações de seqüenciamento e montagem de um genoma por meio da integração e do uso de dois grupos de programas: o pacote Phred/Phrap/Consed [32,42,43,44] ou CAP3 [33] (esse

último somente para montagem). O SABIA possui um sistema completo para manipular as informações de todo o processo de montagem, desde o recebimento dos cromatrogramas até as informações de *reads (*qualidade Phred e seqüências) e formação de *contigs* e *scaffolds* de cada montagem de um genoma. Já o módulo *Annotation* contém as informações de anotação, tendo suporte de diversos programas e referências a bancos de dados que armazenam informações biológicas e funcionais, conforme descrito na revisão. Este módulo disponibiliza informações de predição de genes, anotação e resultados: BLAST [63], KEGG [56,57,58], COG [119,120] e InterPro [121].

A integração do sistema GINGA foi feita com base na mesma infraestrutura computacional em que o sistema SABIA foi desenvolvido [13] (Informações em: *http://www.sabia.lncc.br*) (Figura 3). O banco de dados e o conjunto de *scripts* (código fonte dos programas) do GINGA foram integrados àqueles do SABIA (Figura 3).

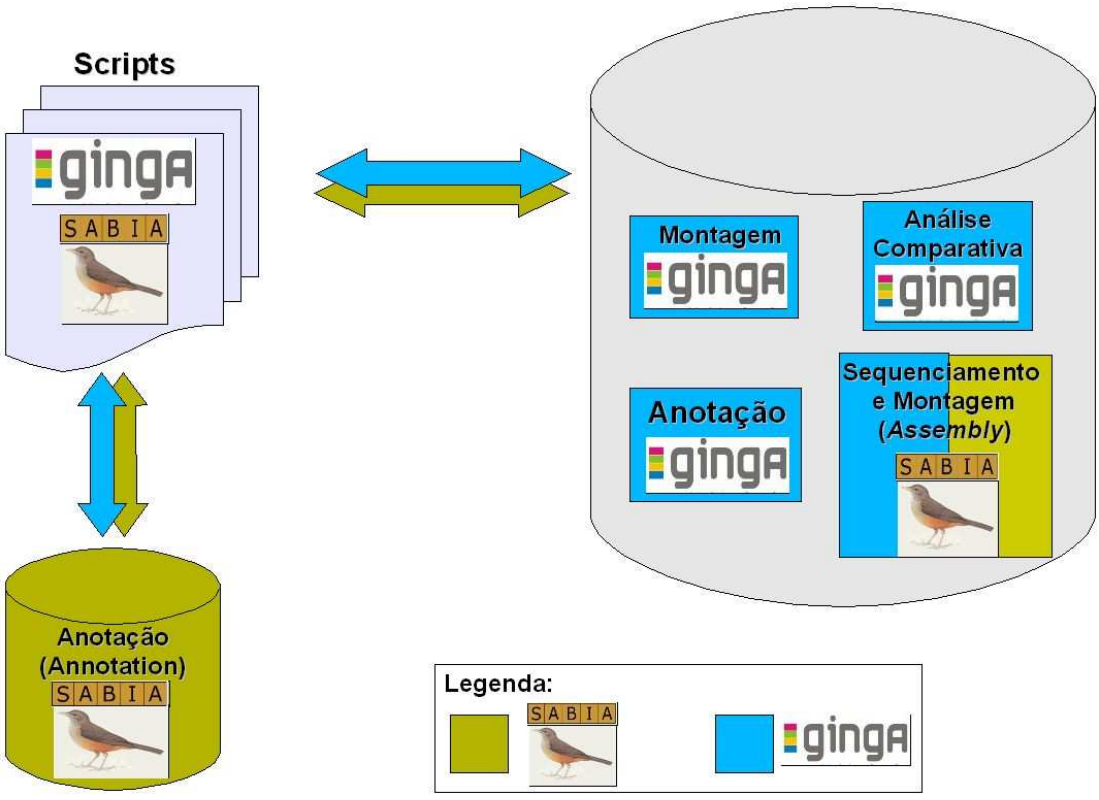


Figura 3: Esquema representativo da integração dos sistemas GINGA e SABIA.

5.1.1.2 Informações do genoma parcial

Foram criadas duas maneiras para a incorporação de informações do genoma parcial no sistema GINGA: (a) usando o sistema SABIA; e (b) sem o uso desse sistema.

5.1.1.2.1 Com o uso do sistema SABIA

O sistema GINGA pode utilizar as informações geradas pelo módulo de montagem (*Assembly*) do sistema SABIA. Essa integração permitiu disponibilizar informações de: (a) seqüência e qualidade Phred das seqüências de *reads*, *contigs* e *scaffolds; (b)* a composição de *reads* de cada *contig* e dos *singlets* (*reads* que não fizeram parte da composição de *contigs*)*; (c)* a composição de *contigs* em cada *scaffold*, e os *contigs* isolados (*contigs* que não fazem parte de *scaffolds*). As informações de anotação são obtidas, como descrito anteriormente, diretamente do módulo *Annotation* do SABIA.

5.1.1.2.2 Sem o uso do sistema SABIA

O GINGA pode ser utilizado sem as informações de montagem geradas pelo SABIA. Neste caso, um arquivo contendo seqüências em formato FASTA de *reads*, *contigs, scaffolds*, ou qualquer outra seqüência de interesse pode ser integrado ao sistema. A incorporação desse arquivo pode ser feita via Portal GINGA, conforme será explicado no item 5.1.4.1. Os resultados, neste caso, são limitados a informações de alinhamento.

5.1.1.3 Informações do genoma completo

A seqüência e as informações de anotação do genoma completo podem ser obtidas dos arquivos de seqüência e ppt, do *Genbank* [51,52] e inseridas via Portal *web* do sistema GINGA. O arquivo ptt é um arquivo texto que contém as informações (posição, tamanho, gene, produto, COG) das ORFs anotadas de um genoma.

O módulo de anotação do SABIA [13], neste caso, é opcional, e pode ser eventualmente utilizado para complementar as informações de anotação obtidas do NCBI (*http://www.ncbi.nlm.nih.gov/*). Quando da utilização do módulo de anotação do SABIA, a seqüência completa do genoma passa pelas etapas de predição e de atribuição da função das ORFs novamente. No caso do uso sem o sistema SABIA as informações do Genbank são incorporadas diretamente numa tabela de anotações exclusiva ao sistema GINGA (Figura 3).

5.1.2 Extração da informação de alinhamento

O alinhamento entre dois genomas se refere ao alinhamento de seqüências obtidas ao acaso do genoma parcial comparada à seqüência completa de um outro genoma. As seqüências utilizadas na comparação são as seqüências de *scaffolds* e *contigs* isolados resultantes da montagem. Nesta seção, são apresentados os passos do algoritmo de extração, para posterior armazenamento (item 5.1.3) e os resultados obtidos do alinhamento entre as seqüências utilizando o programa *cross_match* com os parâmetros descritos na metodologia:

1. Criação de um arquivo contendo a seqüência em formato FASTA do *scaffold*

ou *contig* isolado. A seqüência do arquivo de *scaffold* é formada pelos *contigs* que o compõem na mesma ordem e orientação em que foram construídos pelo programa genscaff [117]. O arquivo FASTA do *scaffold* é formado pela junção de todos os *contigs* que o compõem.

1a. Se o arquivo criado no *passo 1* for o arquivo de *scaffold*, um segundo arquivo é criado, contendo as seqüências em formato FASTA dos *contigs* pertencentes a esse *scaffold*.

2. Criação de um arquivo contendo a seqüência em formato FASTA do genoma completo.

3. Execução do programa *cross_match* para o alinhamento entre a seqüência do genoma parcial (arquivos gerados nos passos 1 e 1a) contra a seqüência do genoma completo (arquivo gerado no passo 2).

4. Leitura, identificação e extração de três tipos de regiões (A, B e C) a partir dos alinhamentos resultantes do programa *cross_match*:

- Regiões que alinharam:
    A. Blocos comuns (*Blocks*): são regiões do genoma parcial que foram alinhadas a uma única região do genoma completo.
    B. Repetições (R*epeats*): regiões do genoma parcial que foram alinhadas a mais de uma região no genoma completo, ou seja, evento de duplicação.
- Regiões que não alinharam:
    C. Específicas (*Specific regions*): regiões do genoma parcial que não foram alinhadas ao genoma completo, ou seja, são particulares a esse genoma.

5. Relacionamento das informações de regiões alinhadas (A, B) e não alinhadas (C) com as informações de montagem descritas anteriormente.

6. Armazenamento nas tabelas do banco de dados de toda a informação gerada.

O programa *cross_match* gera como resultado um arquivo em formato texto (Figura 4 e Figuras 5 – (a) ou (b)), que mostra todas as possibilidades de alinhamento entre as seqüências analisadas. Cada linha (Figura 4) desse arquivo representa o resultado do alinhamento entre as seqüências e que são descritos nos passos 4 a 6. Em todos os casos em que são utilizados *scaffolds* em alinhamentos, sempre foi necessário alinhar os *contigs* que pertencem a cada *scaffold* (Figura 5 – (a)). Esta foi a maneira utilizada para identificar o posicionamento de cada *contig* dentro de um determinado *scaffold* num determinado alinhamento.

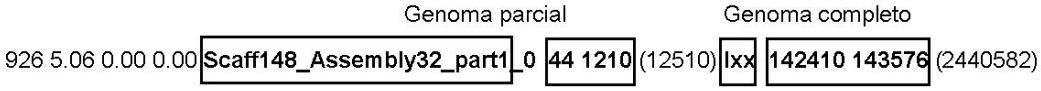


Figura 4: Informações utilizadas pelo sistema GINGA com base no resultado do alinhamento entre as seqüências dos genomas pelo programa cross_match. Os dois primeiros retângulos contêm as informações do genoma parcial com o número do *scaffold* e da montagem e as posições das regiões alinhadas. Os dois próximos retângulos contêm as mesmas informações sobre o genoma completo.

A definição dos três tipos de alinhamentos (grupos A, B e C) foi necessária à construção da representação gráfica. Os blocos comuns representam as regiões que alinharam em uma única posição no genoma completo. A partir dos blocos comuns pode-se identificar grandes eventos de reorganização. As regiões de repetição são seqüências que apresentam mais do que um alinhamento ao genoma completo. O resultado do alinhamento apresentado na figura 5b mostra um exemplo de cada um dos tipos de alinhamento e como as informações foram extraídas. A região em verde de 44 a 1.210 representa um bloco comum, a região em vermelho de 12.344 a 12.402 (que aparece duas vezes) representa uma repetição, e o intervalo da região em azul de 10.188 a 11.454 uma região específica (Figura 5).

O objetivo foi identificar todos os alinhamentos possíveis e armazenar as informações no banco de dados. Desta forma, é permitido ao usuário avaliar cada alinhamento por meio do sistema GINGA. Com isso, o sistema contém a opção do usuário para definir o tipo de região que deseja visualizar para uma análise. O resultado do alinhamento entre os genomas é parte central do sistema e integra todas as outras informações disponíveis no banco de dados.

(a)

926 5.06 0.00 0.00 Contig615 44 1210 (43) lxx 142410 143576 (2440582)
3568 5.74 0.10 0.76 Contig1061 1 4979 (7488) lxx 143823 148768 (2435390)
32 6.52 0.00 0.00 Contig1061 5973 6018 (6449) lxx 148840 148885 (2435273)
2378 4.86 0.03 0.38 Contig1061 6035 8935 (3532) C lxx (2224247) 359911 357021
1187 6.00 0.00 0.89 Contig1061 6044 7610 (4857) C lxx (2536093) 48065 46513 *
707 6.26 0.00 0.11 Contig1061 6135 7045 (5422) lxx 148884 149793 (2434365) *
64 10.00 1.00 0.00 Contig1061 10201 10300 (2167) lxx 2450262 2450362 (133796)
58 12.63 0.00 0.00 Contig1061 10207 10301 (2166) C lxx (1686475) 897683 897589 *
41 7.41 0.00 0.00 Contig1061 10249 10302 (2165) lxx 2451091 2451144 (133014) *
134 6.63 0.00 0.00 Contig1061 10841 11021 (1446) lxx 2450910 2451090 (133068)
49 5.08 0.00 0.00 Contig1061 11091 11149 (1318) lxx 2451829 2451887 (132271)
49 5.08 0.00 0.00 Contig1061 11091 11149 (1318) C lxx (1687318) 896840 896782 *
48 4.48 0.00 2.99 Contig1061 12058 12124 (343) lxx 2453377 2453441 (130717)

(b)

926 5.06 0.00 0.00 Scaff148_Assembly32_part1_0 44 1210 (12510) lxx 142410 143576 (2440582)
3568 5.74 0.10 0.76 Scaff148_Assembly32_part1_0 1254 6232 (7488) lxx 143823 148768 (2435390)
32 6.52 0.00 0.00 Scaff148_Assembly32_part1_0 7226 7271 (6449) lxx 148840 148885 (2435273)
2378 4.86 0.03 0.38 Scaff148_Assembly32_part1_0 7288 **10188** (3532) C lxx (2224247) 359911 357021
1187 6.00 0.00 0.89 Scaff148_Assembly32_part1_0 7297 8863 (4857) C lxx (2536093) 48065 46513 *
707 6.26 0.00 0.11 Scaff148_Assembly32_part1_0 7388 8298 (5422) lxx 148884 149793 (2434365) *
64 10.00 1.00 0.00 Scaff148_Assembly32_part1_0 **11454** 11553 (2167) lxx 2450262 2450362 (133796)
58 12.63 0.00 0.00 Scaff148_Assembly32_part1_0 11460 11554 (2166) C lxx (1686475) 897683 897589 *
41 7.41 0.00 0.00 Scaff148_Assembly32_part1_0 11502 11555 (2165) lxx 2451091 2451144 (133014) *
134 6.63 0.00 0.00 Scaff148 Assembly32 part1 0 12094 12274 (1446) lxx 2450910 2451090 (133068)
49 5.08 0.00 0.00 Scaff148_Assembly32_part1_0 12344 12402 (1318) lxx 2451829 2451887 (132271)
49 5.08 0.00 0.00 Scaff148_Assembly32_part1_0 12344 12402 (1318) C lxx (1687318) 896840 896782 *
48 4.48 0.00 2.99 Scaff148_Assembly32_part1_0 13311 13377 (343) lxx 2453377 2453441 (130717)

Figura 5: Exemplos dos resultados de alinhamento entre seqüências realizado pelo programa *cross_match*: (a) mostra o resultado do alinhamento entre a seqüência dos *contigs* que compõem um *scaffold* e a seqüência do genoma completo; e (b) mostra o resultado do alinhamento entre a seqüência de um *scaffold*, composto por *contigs* em (a), e a seqüência do genoma completo. A letra "C" representa que o alinhamento aconteceu de forma complementar, ou seja, uma seqüência está orientada de maneira invertida a outra. A região identificada como bloco comum está representada em cor verde, a região específica está representada em cor azul, e as regiões de repetição estão representadas em cor vermelha.

5.1.3 Estrutura do banco de dados

As informações extraídas de cada um dos genomas no alinhamento entre as seqüências são armazenadas em um banco de dados composto por 17 tabelas (Tabela 4 e Figura 6). Uma descrição de cada uma das tabelas e como elas se relacionam é feita a seguir (Tabela 4 e Figura 6):

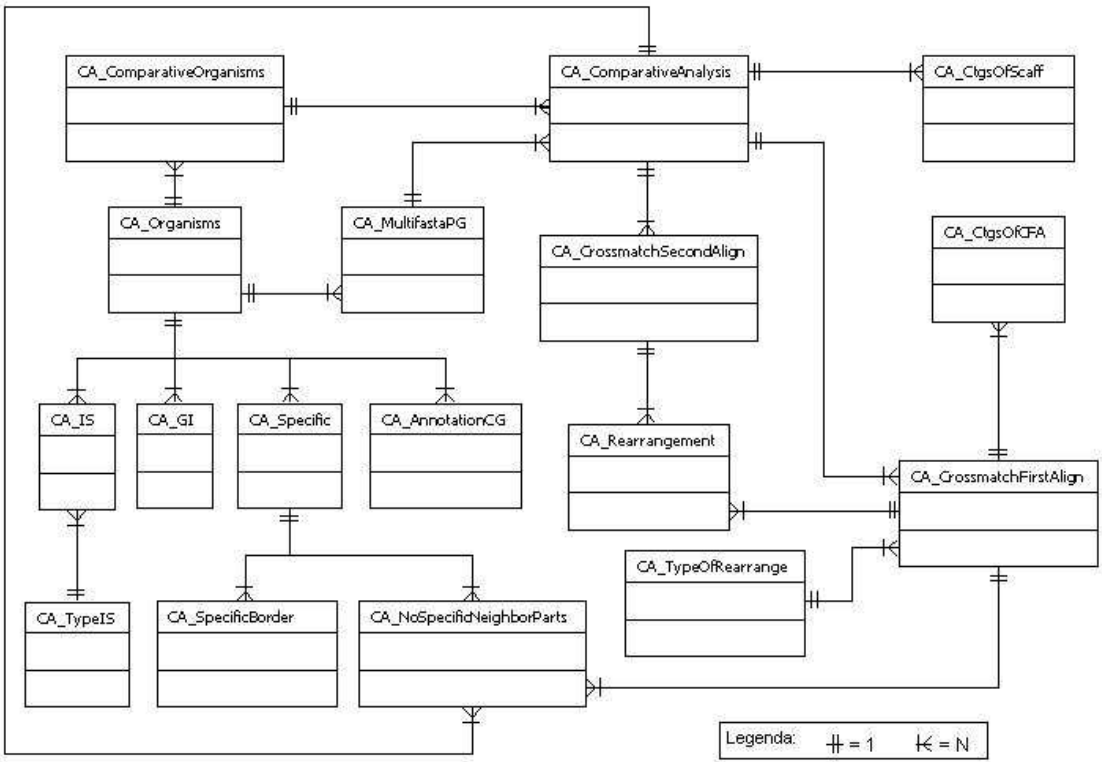


Figura 6: Representação dos relacionamentos entre as 17 tabelas (descritas na Tabela 4) do sistema GINGA. Na legenda destaca-se a notação do tipo de relacionamento que pode existir entre duas tabelas sendo: que 1 significa um registro e N muitos registros. Assim, pode-se ter três tipos de relacionamentos entre duas tabelas: (a) um para um (notação 1:1) – cada um registro de uma tabela relaciona-se com um registro da outra; (b) um para muitos (notação 1:N) – um registro de uma tabela relaciona-se com muitos registros de outra tabela; e (c) muitos registros de uma tabela relacionam-se com muitos de outra tabela (notação N:N). No relacionamento N:N deve-se utilizar uma tabela auxiliar tornando um relacionamento de (1:N). Exemplo: tabela *CA_Rearrangement*. A descrição de cada tabela é apresentada na Tabela 4.

Tabela 4: Descrição de cada uma das tabelas do banco de dados do sistema GINGA.

Tabela	*Descrição*
CA_Organisms	Armazena a informação de cada organismos (genoma), sendo: nome científico, código, tipo de genoma (completo ou parcial), a fonte da informação de anotação (ambos genomas) e de montagem (genoma parcial), seqüência do genoma completo, tamanho do genoma e porcentagem GC (genoma completo)
CA_MultiFastaPG	Armazena as seqüências de *reads*, *contigs*, *scaffolds* ou qualquer tipo de seqüência do genoma parcial de interesse do usuário a ser comparado
CA_ComparativeOrganisms	Contém a lista dos organismos escolhidos para comparação e identificação do genoma completo e genoma parcial
CA_ComparativeAnalysis	Contém uma lista dos segmentos (*scaffolds* e *contigs* isolados) comparados e identificação dos genomas (informação da tabela *CA_ComparativeOrganisms*). Armazena informações da montagem do genoma parcial: identificação, tamanho, seqüência e qualidade de seqüência dos segmentos comparados
CA_CrossmatchFirstAlign	Contém as informações de alinhamento dos *contigs* isolados e *scaffolds* (informação presente no relacionamento com *CA_ComparativeAnalysis*) do genoma parcial: coordenadas de início e fim, seqüência, tipo de arranjo
CA_CrossmatchSecondAlign	Contém as informações de alinhamento do genoma completo: coordenadas de início e fim, tamanho, seqüência e orientação. Através do relacionamento com a tabela *CA_ComparativeAnalysis sabe-se a qual* segmento (*contigs* isolados ou *scaffolds*) está ligada essa região alinhada do genoma completo
CA_TypeOfRearrange	Lista dos tipos de rearranjos (tipo de repetições ou *overlaps*)
CA_Rearrangement	Relaciona a informação de qual região do genoma parcial (informação da tabela *CA_CrossmatchFirstAlign*) alinhou com qual região do genoma completo (informação da tabela *CA_CrossmatchSecondAlign*)
CA_Specific	Informações sobre as regiões específicas do genoma parcial, como: *contig*, coordenadas de início e fim, tamanho, seqüência e segmento comparado (informação essa do relacionamento com a tabela *CA_ComparativeAnalysis*)
CA_SpecificBorder	Se a região específica da tabela *CA_Specific* estiver em mais de um *contig*, cada subregião de cada *contig* é armazenada nesta tabela. As informações são: *contig*, coordenadas de início e fim do pedaço, tamanho e seqüência
CA_NoSpecificNeighborParts	Informações (identificação e posição) das regiões flanqueadoras de cada uma das regiões específicas (informação das regiões flanqueadoras obtida através do relacionamento da tabela *CA_CrossmatchFirstAlign* e região específica da tabela *CA_Specific*)
CA_CtgsOfCFA	Contém as regiões (relacionamento com a tabela *CA_CrossmatchFirstAlign*) dos *contigs* do genoma parcial que foram alinhadas
CA_CtgsOfScaff	Contém a lista de *contigs* dentro de cada *scaffold* (motivo esse do relacionamento com a tabela *CA_ComparativeAnalysis* somente quando for comparado *scaffold*)
CA_GI	Lista das ilhas genômicas do genoma completo (nome, coordenadas de início e fim, tamanho e % GC)
CA_IS	Lista das seqüências de inserção (IS) do genoma completo (tipo, coordenadas de início e fim e orientação). O tipo é vinculado com a tabela *CA_TypeIS* que contém o nome das IS
CA_TypeIS	Lista dos nomes das seqüências de inserção (IS)
CA_AnnotationCG	Contém as informações de anotação do genoma completo recuperadas do Genbank (arquivo em formato ptt)

Para ilustrar os relacionamentos entre as tabelas, utlizamos o exemplo da tabela *CA_Organisms,* que contém as informações centrais sobre os organismos e alimenta outras tabelas como:

- A tabela *CA_MultiFastaPG* só será utilizada na ausência do uso do módulo de montagem do sistema SABIA. Assim, o relacionamento com a tabela *CA_Organisms* associa com as seqüências do genoma parcial.
- O relacionamento das tabelas *CA_ComparativeOrganisms* e *CA_Organisms* associa as informações dos dois genomas (parcial e completo) selecionados para a análise comparativa.
- O relacionamento com a tabela *CA_Specific* identifica o organismo que contém determinada região específica.
- As tabelas *CA_GI*, *CA_IS* e *CA_AnnotationCG* estão relacionadas tabela *CA_Organisms* para a identificação do genoma completo.

As tabelas *CA_ComparativeAnalysis*, *CA_CtgsOfScaff* e *CA_CtgsOfCFA*, *CA_Specific e CA_SpecificBorder* contém relacionamentos com as tabelas do sistema SABIA. Contudo, as tabelas do sistema SABIA não fazem parte dessa descrição (Informações em: *http://www.sabia.lncc.br*) [13].

As informações sobre as ilhas genômicas e as seqüências de inserção (IS) de um genoma completo são caractéristicas de cada genoma e resultado de uma análise detalhada e específica de cada projeto. Essas informações foram incorporadas ao sistema GINGA por participarem na organização diferencial dos cromossomos muitas vezes encontrada entre genomas de espécies próximas.

5.1.4 Portal GINGA – Portal de acesso ao sistema

Com o objetivo de unificar as funções do sistema e ainda buscar ser um ambiente amigável ao usuário, foi criado o Portal GINGA (*http://www.ginga.lncc.br*), que permite o acesso ao sistema via *interface web*. Nesse portal é permitido desde a inclusão das informações necessárias de cada um dos genomas analisados, definir os programas com as opções de configuração, iniciar o processo de alinhamento, extração das informações, armazenamento e a visualização dos resultados na forma de representação gráfica e em relatórios complementares.

5.1.4.1 Explorando as funções do Portal

Para o acesso ao portal é necessário primeiro realizar uma validação de usuário que restringe o acesso a usuários autorizados. Após a validação, o acesso ao portal está disponível e o usuário será direcionado às opções do sistema dispostas em um menu. Cada item do menu está ligado a uma função do sistema (Figura 7).

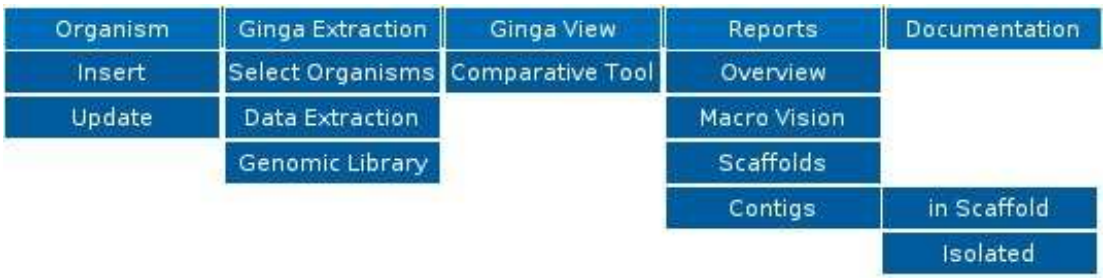


Figura 7: Tela que apresenta a lista de opções do menu do Portal GINGA.

- ***Organism***

Esta opção do menu é referente ao cadastro (inserção e atualização) das informações dos organismos em estudo, tanto do organismo com seqüenciamento do genoma em andamento (genoma parcial) quanto aquele cuja seqüência foi totalmente determinada (genoma completo). Foram criados dois sub-itens nesta etapa: *Insert Organism* e *Update Organism* (Tabela 5 e Figura 8). O subitem *Update* contém as mesmas opções disponíveis para subitem *Insert Organism*, porém apenas para atualização de dados.

Tabela 5: Descrição das opções de cadastro sobre os organismos que serão analisados.

Opção	***Descrição***
Organism – Scientific name	Nome científico do organismo
Organism – Code	Código para o organismo seguindo um padrão de três letras
Genome Sequence	Genoma completo ou genoma parcial
% GC	Porcentagem do conteúdo GC do genoma completo. Informação apenas para o genoma completo.
Select annotation of CG[1]	Item que informa de onde se deve extrair as informações de anotação do genoma completo: ***None –*** não mostrar as informações de anotação; ***Genbank –*** anotação obtida do Genbank (arquivo no formato .ptt); ***SABIA –*** anotação obtida da integração com SABIA – *Annotation*
Genome size (base pairs) - CG[1]	Tamanho do genoma completo em pares de bases
Sequence file - CG[1] (FASTA format)	Arquivo com a seqüência de nucleotídeos do genoma completo em formato FASTA
Annotation file - CG1 (.PTT format)	Arquivo no formato ptt (Genbank) com as informações de anotação do genoma completo
Select assembly of PG[2]	Item que informa de onde deve ser extraída a informação de montagem (seqüência) do genoma parcial: **Arquivo multifasta** (*Multifasta Sequence*), ou **SABIA –** seqüências obtidas da integração com SABIA - *Assembly*
File of Multifasta Sequence – PG [2] (Multifasta format)	Opção para inserção das informações do arquivo contendo as seqüências em formato FASTA do genoma parcial. Esse arquivo só deve ser inserido quando a opção for *Multifasta Sequence* no item *Show assembly* of *PG*[2] e pode conter qualquer tipo de seqüência, conforme já apresentado

[1] CG-*Complete Genome*; [2] PG-*Partial Genome*

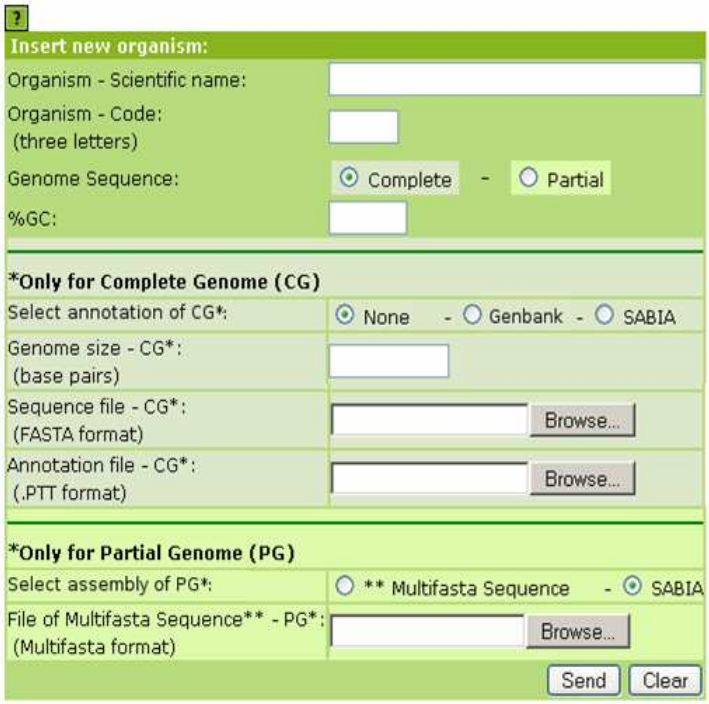


Figura 8: Tela que apresenta as opções (descritas na Tabela 5) disponíveis para o cadastro de informações sobre dos organismos que serão analisados (*Insert Organism*).

- ***Ginga Extraction***

A opção *Ginga Extraction* está ligada à função de extrair e armazenar o resultado do alinhamento entre as seqüências. Nesta opção, tem-se os subitens: (a) *Select Organisms* (Figura 9), que permite definir os genomas que serão analisados; e (b) *Data Extraction* (Figura 10 e Tabela 6) que, baseado nos genomas, na montagem parcial (Figura 10, opção *Assembly* da Tabela 6) e no tipo de segmento (*scaffold* ou *contig* isolado), realiza o alinhamento, extração e armazenamento das informações com o uso do programa *cross_match*. Outra opção desse item é o subitem *Genomic Library* (Figura 11), que contém a lista de bibliotecas genômicas utilizadas durante o processo de seqüenciamento e que permite definir a sua presença na visualização gráfica (*Comparative Tool*) do menu *Ginga View*.

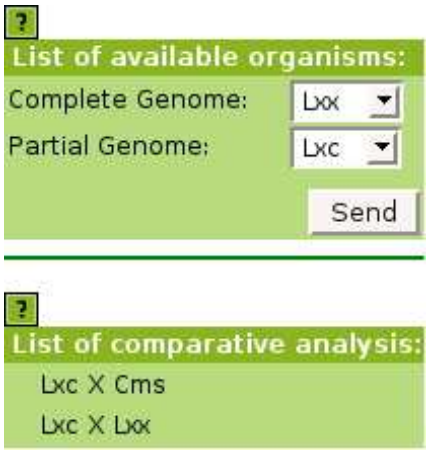


Figura 9: Tela que apresenta as opções de escolha dos genomas para a análise comparativa (*Select Organisms*). Neste exemplo, foram listadas duas análises comparativas disponíveis: Lxc X Cms, que contém as informações da comparação entre os genomas de *Leifsonia xyli* subsp. *cynodontis* (Lxc) e *Clavibacter michiganensis* subsp. *sepedonicus* e novamente o genoma de Lxc e *Leifsonia xyli* subsp. *xyli*.

A partir da inserção das informações de um genoma parcial, pode-se fazer a comparação com quantos genomas completamente seqüenciados forem de interesse.

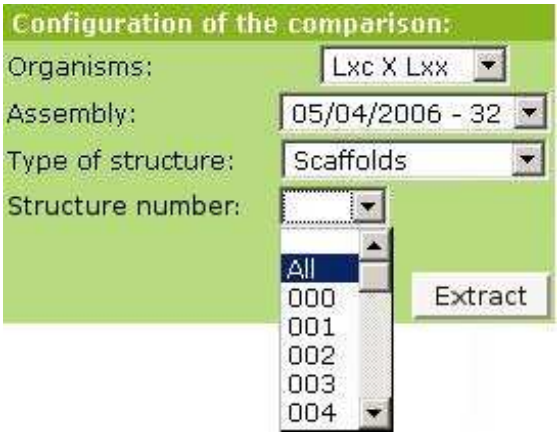


Figura 10: Tela que apresenta as opções de configuração para a extração e o armazenamento dos resultados do alinhamento realizado pelo *cross_match* (*Data Extraction*) e são descritas na Tabela 6, abaixo.

Tabela 6: Descrição das opções de configuração (*Data Extraction*) da comparação entre os genomas (parcial e completo).

Opção	***Descrição***
Organisms	Escolher entre os genomas que serão analisados
Assembly	Escolher dentre todas as montagens disponíveis do genoma parcial. Nessa opção é listado o nome e o número da montagem
Type of structure	Tipo de segmento que será analisada: *Scaffolds* ou *contigs* isolados
Structure number	Possibilita a escolha de um *scaffold* ou *contig* isolado em particular ou todos que serão alinhados

Figura 11: Tela que apresenta as opções de bibliotecas genômicas a serem visualizadas na representação gráfica.

- ***Ginga View***

Ginga View está ligado a todas as funções da representação gráfica do sistema. A visualização da análise é feita no subitem *Comparative Tool*, acessível dentro do menu *Ginga View*. Ao acessar o subitem *Comparative Tool* da ferramenta de visualização, o usuário é direcionado a uma tela de configuração (Figura 12). Cada quadro destaca um grupo de informação a ser visualizado posteriormente (Figura 12). O primeiro grupo, de **cor verde,** disponibiliza as opções de organismos, montagem, *scaffold* ou *contigs* isolados. O segundo grupo, de **cor magenta,** disponibiliza as informações de montagem do genoma parcial. O terceiro grupo, de **cor amarela,** é referente às informações de alinhamento. O grupo de opções, ainda em **cor amarela,** apresenta uma tabela de cores, e permite definir as cores para a visualização das regiões comuns, específicas e repetições. As cores pré-definidas para essas regiões são: verde para regiões comuns, azul para regiões específicas e vermelha para regiões de repetições. No quarto grupo de opções, de **cor azul**, estão as opções referentes à anotação, seqüências de inserção (ISs), ilhas genômicas (GI) e conteúdo de bases GC. Definida a configuração, o sistema buscará no banco as informações e construirá a representação gráfica da análise comparativa. Exemplo da visualização

de todas as opções disponíveis na análise comparativa entre as seqüências do *Scaffold* 000 de Lxc e genoma completo de Lxx (Figura 13 e Tabela 7).

Figura 12: Tela de configuração das informações disponíveis da análise comparativa entre os genomas para serem visualizadas na representa gráfica. Cada tabela é um grupo de informações, sendo: (a) em verde são opções sobre a comparação; (b) em azul as informações de anotação de ambos genomas; (c) em magenta são informações de montagem; e (d) em amarelo sobre as regiões alinhadas e não alinhadas.

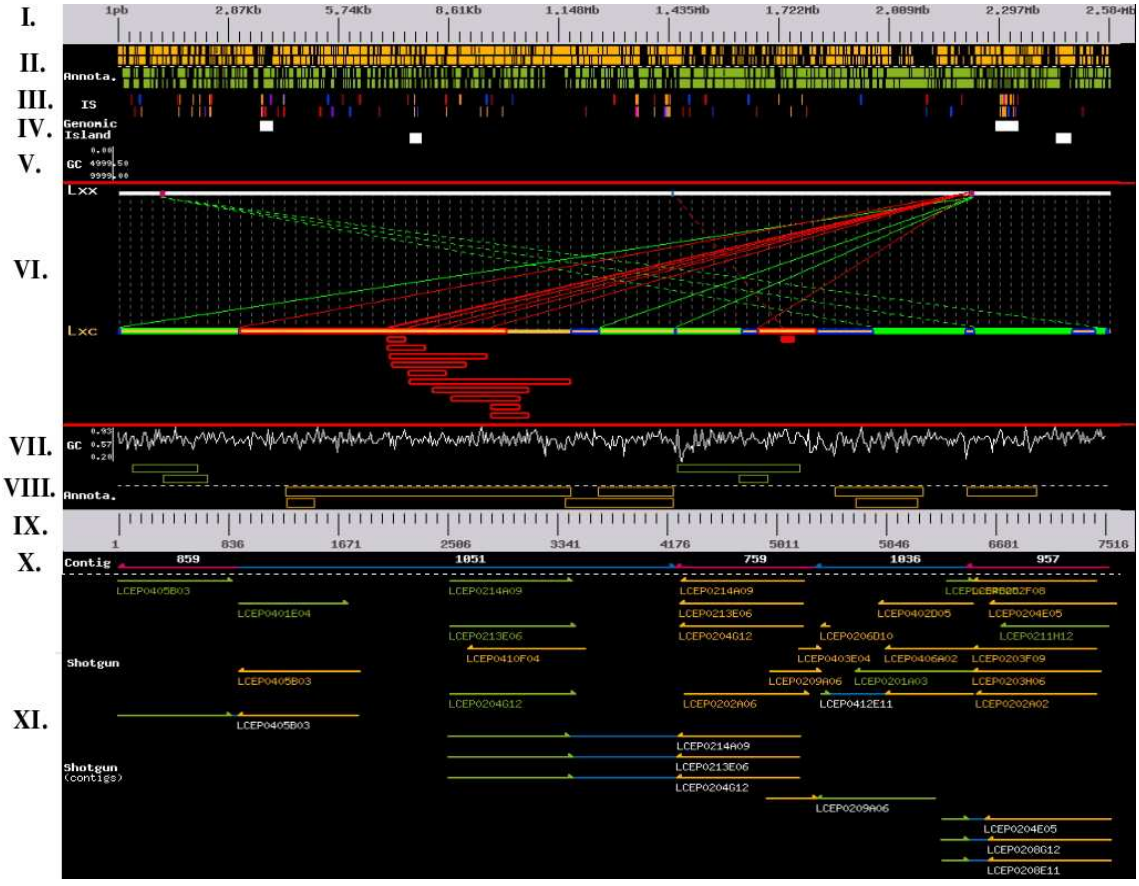


Figura 13: Tela que apresenta as opções de visualização da análise comparativa entre as seqüências do *Scaffold* 000 de Lxc e o genoma completo de Lxx. I e IX mostram as réguas em pares de bases para o genoma completo e parcial, respectivamente; II, III, IV e V mostram as informações de anotação (ORFs, ISs, ilhas genômicas e conteúdo GC) do genoma completo, VII e VIII são informações de anotação do genoma parcial (conteúdo GC e ORFs); VI é a visualização da comparação entre genomas; X e XI é a composição de *contigs* do *scaffold* sob análise e composição de *reads* em cada *contig*. Descrição detalhada na Tabela 7.

Tabela 7: Descrição de cada item da representação gráfica da Figura 13.

Item	***Descrição***
I.	Régua em pares de bases do genoma completo
II.	Informação de anotação do genoma completo mostrando as ORFs anotadas
III.	Posicionamento dos ISs do genoma completo, quando disponível
IV.	Posicionamento das GI do genoma completo, quando disponível
V.	Variação do conteúdo GC do genoma completo
VI.	1º) linha horizontal representando a seqüência do genoma completo (superior em branco – Lxx) e uma linha horizontal representando a seqüência do genoma parcial (inferior em laranja – Lxc). As linhas representam os genomas (completo e parcial) em escalas diferentes. 2º) Informações do resultado do alinhamento: (a) em verde são blocos comuns; (b) bem azul região específica; (c) em vermelho repetições. Linhas transversais ligam regiões alinhadas. Inversões são representadas por linhas tracejadas sendo que, nesse caso, essa região fica preenchida (pintado)
VII.	Mostra a variação do conteúdo GC do genoma parcial
VIII.	Informação de anotação do genoma parcial mostrando as ORFs anotadas
IX.	Régua em pares de bases do genoma parcialmente seqüênciado
X.	*Contigs* que compõem o *scaffold ou contig* isolado do genoma parcial
XI.	*Reads* de cada tipo de biblioteca genômica: *(a) reads* que compõem cada *contig*; *(b) reads* responsáveis pela ligação virtual entre os *contigs* daquele *scaffold*. As setas indicam a orientação do *read* em relação ao *contig*

- ***Reports***

O item *Reports* contém os diversos relatórios que complementam as informações extraídas na análise comparativa, como apresentado a seguir.

- ***Overview***

Esse item refere-se à apresentação de um relatório com informações gerais da montagem do genoma parcial e da análise comparativa (Figura 14 e Tabela 8).

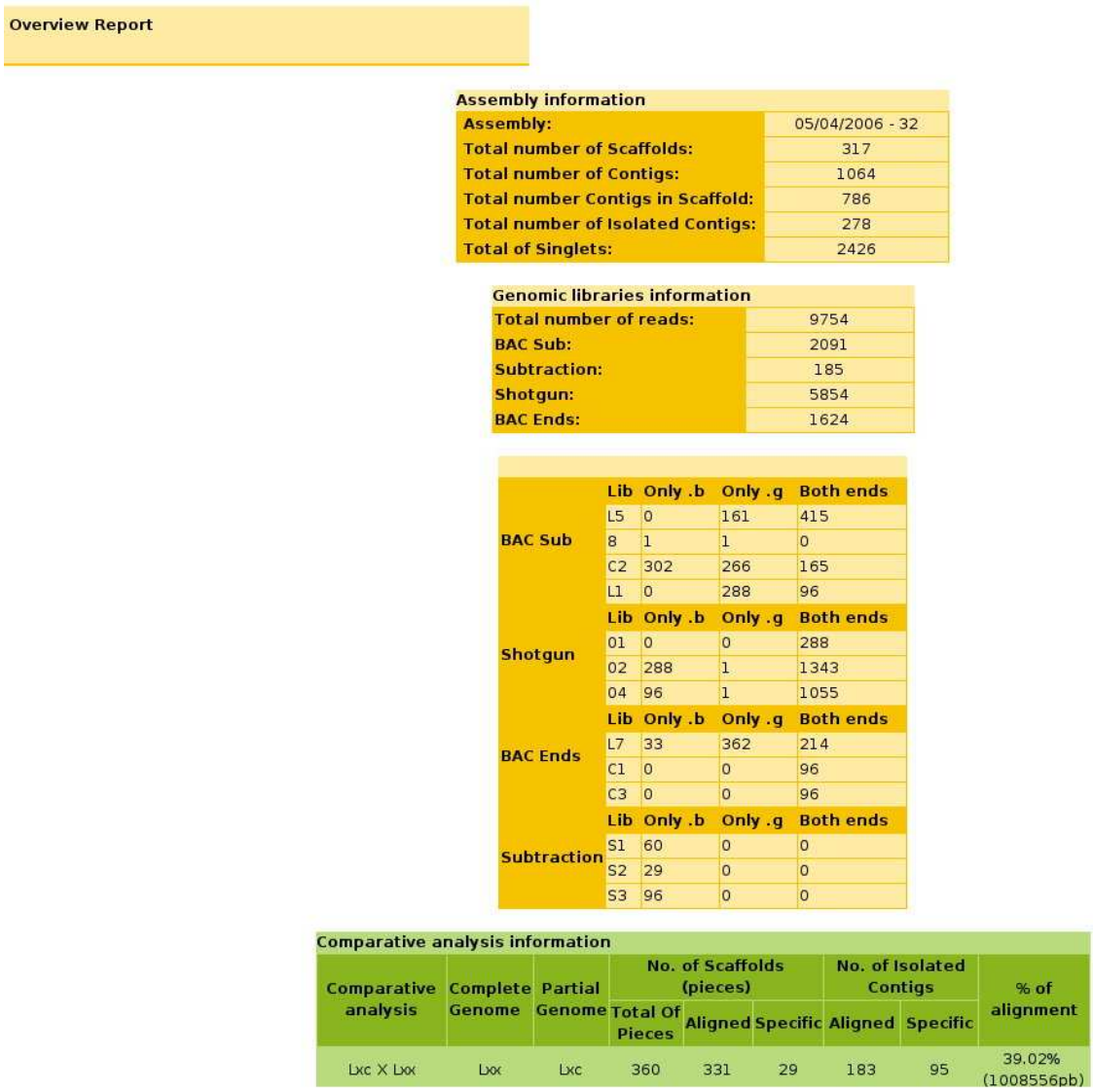
Tabela 8: Descrição da lista de informações apresentadas no relatório geral (*Overview*) apresentado na Figura 14.

Informações sobre a montagem	
Informação	***Descrição***
Assembly	Mostra a data e a numeração da montagem do genoma parcial
Total number of Scaffolds	Número total de *scaffolds* formados na montagem
Total number of Contigs	Número total de *contigs* formados na montagem
Total number Contigs in Scaffold	Número total de *contigs* que formaram *scaffolds*
Total number of Isolated Contigs	Número total de *contigs isolados*
Total of Singlets	Número total de *singlets*, ou seja, *reads* que não formaram *contigs*

Informações sobre as bibliotecas genômicas	
Informação	***Descrição***
Total number of reads	Número total de seqüências (*reads*) utilizadas na montagem
Libraries	Número total *reads* por cada tipo de biblioteca genômica

Informações sobre o seqüenciamento dos insertos clonados	
Informação	***Descrição***
Libraries	Tipo de biblioteca genômica
Lib	Nomenclatura da biblioteca genômica
Only .b	Número total de insertos que tiveram uma única extremidade seqüenciada (*reads* b)
Only .g	Número total de insertos que tiveram uma única extremidade seqüenciada (*reads* g)
Both ends	Número total de insertos que tiveram as duas extremidades seqüenciadas (*reads* casados - *reads* .b e .g)

Informações sobre a análise comparativa	
Informação	***Descrição***
Comparative analysis	Genoma parcial e genoma completo analisados
Complete Genome	Código utilizado para o genoma completo
Partial Genome	Código utilizado para o genoma parcial
N. of Scaffolds (pieces) Aligned / Specific	**Aligned** – N° de pedaços de *scaffolds* que *alinharam* **Specific** – N° de pedaços de *scaffolds* que *não alinharam*
N. of Isolated Contigs Aligned / Specific	**Aligned** – N° de *contigs isolados* que *alinharam* **Specific** – N° de *contigs isolados* que *não alinharam*
% of alignment	Porcentagem total de regiões alinhadas no genoma completo, e o número total de bases



Overview Report

Assembly information

Assembly:	05/04/2006 - 32
Total number of Scaffolds:	317
Total number of Contigs:	1064
Total number Contigs in Scaffold:	786
Total number of Isolated Contigs:	278
Total of Singlets:	2426

Genomic libraries information

Total number of reads:	9754
BAC Sub:	2091
Subtraction:	185
Shotgun:	5854
BAC Ends:	1624

	Lib	Only .b	Only .g	Both ends
BAC Sub	L5	0	161	415
	8	1	1	0
	C2	302	266	165
	L1	0	288	96
	Lib	**Only .b**	**Only .g**	**Both ends**
Shotgun	01	0	0	288
	02	288	1	1343
	04	96	1	1055
	Lib	**Only .b**	**Only .g**	**Both ends**
BAC Ends	L7	33	362	214
	C1	0	0	96
	C3	0	0	96
	Lib	**Only .b**	**Only .g**	**Both ends**
Subtraction	S1	60	0	0
	S2	29	0	0
	S3	96	0	0

Comparative analysis information

Comparative analysis	Complete Genome	Partial Genome	No. of Scaffolds (pieces)			No. of Isolated Contigs		% of alignment
			Total Of Pieces	Aligned	Specific	Aligned	Specific	
Lxc X Lxx	Lxx	Lxc	360	331	29	183	95	39.02% (1008556pb)

Figura 14: Tela que apresenta o relatório geral com informações sobre a montagem do genoma parcial e resultados da análise comparativa como descrito na Tabela 8.

- ***Macro Vision***

O Relatório *Macro Vision* (Figura 15 e Tabela 9) lista o alinhamento de todos os blocos comuns, com opções de listar os alinhamentos com base na posição do genoma completo ou pelo número do *scaffold* formado.

Macro Vision Report
Assembly: 05/04/2006 - 32
Comparative: Lxc X Lxx

Scaffold: 000

Contig	Contig size	Alignment (bp) - PG	Start - CG	End - CG	Alignment (bp) - CG	Gap	Orientation	Repeats
957	1078	78	111876	111954	78	24	+	-
957	1078	733	111978	112736	758	3666	+	-
1036	1150	695	116402	117095	693	464701	+	-
107	841	645	581796	582442	646	602	-	-
591	1138	1125	583044	584174	1130	559	-	-
760	996	890	584733	585630	897	1635105	-	-
859	912	907	2220735	2221645	910	3702	-	-
1051	3314	576	2225347	2225923	576	696	-	-
759	1061	498	2226619	2227116	497		-	-

Scaffold: 001

Contig	Contig size	Alignment (bp) - PG	Start - CG	End - CG	Alignment (bp) - CG	Gap	Orientation	Repeats
976	2777	111	1253400	1253511	111	1283920	+	-
307	909	893	2537431	2538326	895	680	-	-
504	1169	849	2539006	2539840	834	5136	-	-
614	873	872	2544976	2545837	861	589	-	-
631	489	137	2546426	2546563	137	1193	-	-
406	1117	1116	2547756	2548872	1116	986	-	-
817	1627	1578	2549858	2551437	1579	381	-	-
1054	5437	4948	2551818	2556748	4930	255	-	-
624	938	874	2557003	2557875	872		-	-

Scaffold: 002

Contig	Contig size	Alignment (bp) - PG	Start - CG	End - CG	Alignment (bp) - CG	Gap	Orientation	Repeats
983	1577	2992	1477247	1480243	2996	-	+	-
970	1512	2992	1477247	1480243	2996		+	-

Figura 15: Tela que apresenta o relatório visão macro (*Macro Vision*) contendo as informações sobre o alinhamento entre os genomas parcial e completo. As células em verde indicam diferenças no tamanho da região alinhada entre os genomas. As células em azul indicam a mudança de orientação do alinhamento, e em branco e laranja indicam cada vez que o alinhamento entre as seqüências tem uma discrepância maior do que 10.000 pb (*gap*). A sigla PG refere-se a *Partial Genome* e GC a *Complete Genome*, sendo que cada coluna é descrita na Tabela 9.

Tabela 9: Descrição das opções do relatório visão macro (*Macro Vision*).

Opção	***Descrição***
Contig	Número do *contig* alinhado
Contig size	Tamanho do *contig* em pares de bases
Size aligned – PG[1]	Tamanho em pares de bases da região alinhada no genoma parcial (PG[1])
Start – CG	Posição de início do alinhamento ao genoma completo (CG[2])
End – CG	Posição de fim do alinhamento ao genoma completo (CG[2])
Size aligned – CG[2]	Tamanho em pares de bases da região alinhada no genoma completo (CG[2])
Gap	Intervalo (distância em pares de bases) entre duas regiões alinhadas
Orientation	Orientação do alinhamento sendo: + não invertido e – invertido
Repeats	Número de repetições que aquele *contig* apresenta

[1] PG-Partial Genome; [2] CG-Complete Genome

- ***Scaffolds***

O relatório de *scaffolds* (Figura 16) apresenta todos os *scaffolds* alinhados (tabela representada em cor azul) e todos os não alinhados ao genoma completo (tabela representada em cor verde). Cada célula contendo o número do *scaffold* formado apresenta ligações para a seqüência de bases, qualidade *phred* de cada base, composição de *contigs*, seqüência e qualidade de cada *contig* e os resultados dos alinhamentos realizados pelo programa *cross_match* (entre *scaffold*/genoma completo e *contigs* que compõem o *scaffold*/genoma completo).

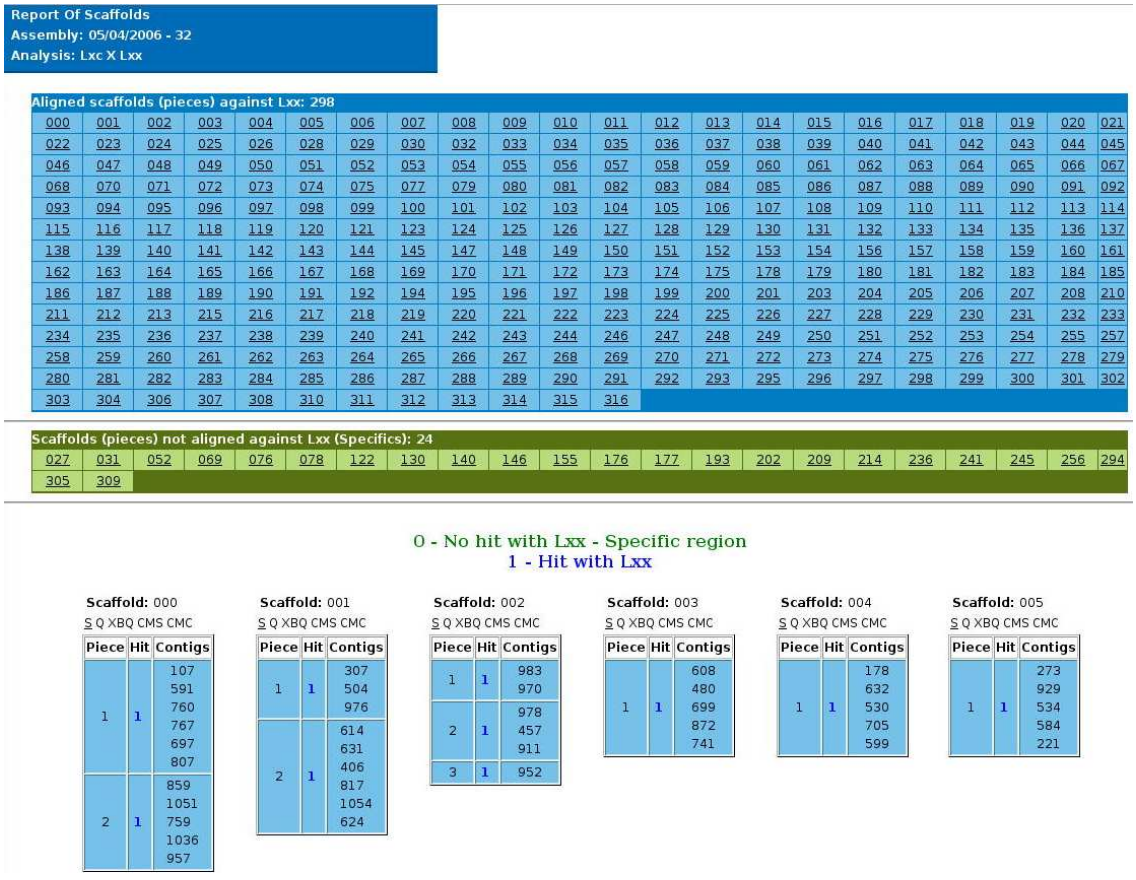
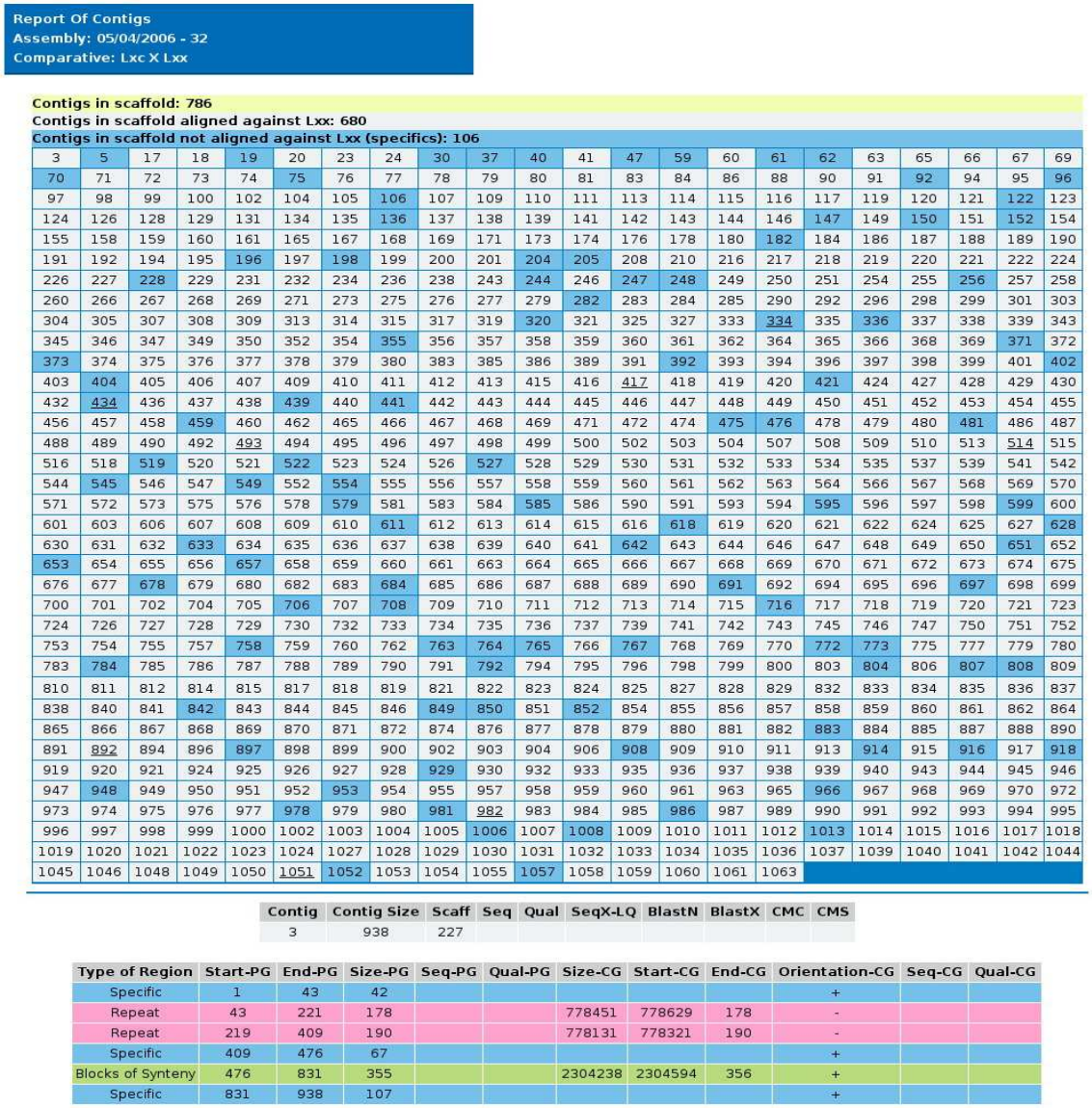


Figura 16: Tela que apresenta as informações de todos os *scaffolds* alinhados ao genoma completo (azul) e todos os *scaffolds* não alinhados (verde). O detalhe sobre o alinhamento dos *scaffolds* está apresentado nas tabelas inferiores, com a formação de *contigs* e a subdivisão em partes de cada *scaffold* (o item 5.2 explica a divisão do *scaffold* em partes).

- ***Contigs in Scaffolds***

Este relatório mostra as informações dos *contigs* que formaram *scaffolds* alinhados ao genoma de Lxx (células em cinza) e os *contigs* não alinhados (células em azul) (Figura 17). Informações de seqüência, qualidade, seqüência filtrada, os resultados dos alinhamentos e de BLAST (blastn e blastx) também são apresentados, quando disponíveis.



Report Of Contigs
Assembly: 05/04/2006 - 32
Comparative: Lxc X Lxx

Contigs in scaffold: 786
Contigs in scaffold aligned against Lxx: 680
Contigs in scaffold not aligned against Lxx (specifics): 106

3	5	17	18	19	20	23	24	30	37	40	41	47	59	60	61	62	63	65	66	67	69
70	71	72	73	74	75	76	77	78	79	80	81	83	84	86	88	90	91	92	94	95	96
97	98	99	100	102	104	105	106	107	109	110	111	113	114	115	116	117	119	120	121	122	123
124	126	128	129	131	134	135	136	137	138	139	141	142	143	144	146	147	149	150	151	152	154
155	158	159	160	161	165	167	168	169	171	173	174	176	178	180	182	184	186	187	188	189	190
191	192	194	195	196	197	198	199	200	201	204	205	208	210	216	217	218	219	220	221	222	224
226	227	228	229	231	232	234	236	238	243	244	246	247	248	249	250	251	254	255	256	257	258
260	266	267	268	269	271	273	275	276	277	279	282	283	284	285	290	292	296	298	299	301	303
304	305	307	308	309	313	314	315	317	319	320	321	325	327	333	334	335	336	337	338	339	343
345	346	347	349	350	352	354	355	356	357	358	359	360	361	362	364	365	366	368	369	371	372
373	374	375	376	377	378	379	380	383	385	386	389	391	392	393	394	396	397	398	399	401	402
403	404	405	406	407	409	410	411	412	413	415	416	417	418	419	420	421	424	427	428	429	430
432	434	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455
456	457	458	459	460	462	465	466	467	468	469	471	472	474	475	476	478	479	480	481	486	487
488	489	490	492	493	494	495	496	497	498	499	500	502	503	504	507	508	509	510	513	514	515
516	518	519	520	521	522	523	524	526	527	528	529	530	531	532	533	534	535	537	539	541	542
544	545	546	547	549	552	554	555	556	557	558	559	560	561	562	563	564	566	567	568	569	570
571	572	573	575	576	578	579	581	583	584	585	586	590	591	593	594	595	596	597	598	599	600
601	603	606	607	608	609	610	611	612	613	614	615	616	618	619	620	621	622	624	625	627	628
630	631	632	633	634	635	636	637	638	639	640	641	642	643	644	646	647	648	649	650	651	652
653	654	655	656	657	658	659	660	661	663	664	665	666	667	668	669	670	671	672	673	674	675
676	677	678	679	680	682	683	684	685	686	687	688	689	690	691	692	694	695	696	697	698	699
700	701	702	704	705	706	707	708	709	710	711	712	713	714	715	716	717	718	719	720	721	723
724	726	727	728	729	730	732	733	734	735	736	737	739	741	742	743	745	746	747	750	751	752
753	754	755	757	758	759	760	762	763	764	765	766	767	768	769	770	772	773	775	777	779	780
783	784	785	786	787	788	789	790	791	792	794	795	796	798	799	800	803	804	806	807	808	809
810	811	812	814	815	817	818	819	821	822	823	824	825	827	828	829	832	833	834	835	836	837
838	840	841	842	843	844	845	846	849	850	851	852	854	855	856	857	858	859	860	861	862	864
865	866	867	868	869	870	871	872	874	876	877	878	879	880	881	882	883	884	885	887	888	890
891	892	894	896	897	898	899	900	902	903	904	906	908	909	910	911	913	914	915	916	917	918
919	920	921	924	925	926	927	928	929	930	932	933	935	936	937	938	939	940	943	944	945	946
947	948	949	950	951	952	953	954	955	957	958	959	960	961	963	965	966	967	968	969	970	972
973	974	975	976	977	978	979	980	981	982	983	984	985	986	987	989	990	991	992	993	994	995
996	997	998	999	1000	1002	1003	1004	1005	1006	1007	1008	1009	1010	1011	1012	1013	1014	1015	1016	1017	1018
1019	1020	1021	1022	1023	1024	1027	1028	1029	1030	1031	1032	1033	1034	1035	1036	1037	1039	1040	1041	1042	1044
1045	1046	1048	1049	1050	1051	1052	1053	1054	1055	1057	1058	1059	1060	1061	1063						

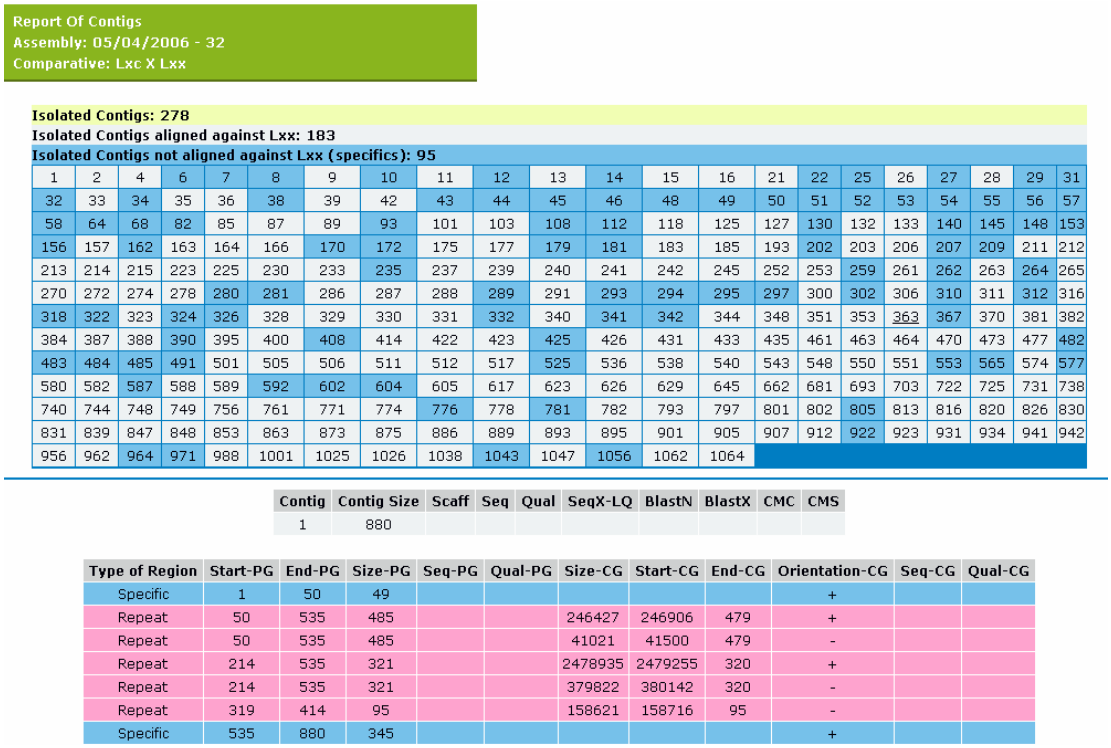
Contig	Contig Size	Scaff	Seq	Qual	SeqX-LQ	BlastN	BlastX	CMC	CMS
3	938	227							

Type of Region	Start-PG	End-PG	Size-PG	Seq-PG	Qual-PG	Size-CG	Start-CG	End-CG	Orientation-CG	Seq-CG	Qual-CG
Specific	1	43	42						+		
Repeat	43	221	178			778451	778629	178	-		
Repeat	219	409	190			778131	778321	190	-		
Specific	409	476	67						+		
Blocks of Synteny	476	831	355			2304238	2304594	356	+		
Specific	831	938	107						+		

Figura 17: Tela que apresenta o relatório de todos os *contigs* que formaram *scaffolds* e que: alinharam (em fundo cinza na tabela superior) e não alinharam (em fundo azul na tabela superior) ao genoma completo. As tabelas inferiores apresentam os detalhes das informações desses *contigs*, e as regiões de alinhamento e específica, quando essa informação for disponível. Cada cor representa o tipo de alinhamento (bloco comum, região específica ou rearranjo), conforme já descrito.

- ***Contigs Isolated***

O relatório que contém as informações dos *contigs* isolados (Figura 18) apresenta o número do *contig*, tamanho em pares de bases da seqüência, seqüência, qualidade da seqüência, seqüência filtrada, resultado de BLAST (blastn e blastx) e resultado do alinhamento pelo *cross_match* entre as seqüências do *contig* e do genoma completo.



Report Of Contigs
Assembly: 05/04/2006 - 32
Comparative: Lxc X Lxx

Isolated Contigs: 278
Isolated Contigs aligned against Lxx: 183
Isolated Contigs not aligned against Lxx (specifics): 95

1	2	4	6	7	8	9	10	11	12	13	14	15	16	21	22	25	26	27	28	29	31
32	33	34	35	36	38	39	42	43	44	45	46	48	49	50	51	52	53	54	55	56	57
58	64	68	82	85	87	89	93	101	103	108	112	118	125	127	130	132	133	140	145	148	153
156	157	162	163	164	166	170	172	175	177	179	181	183	185	193	202	203	206	207	209	211	212
213	214	215	223	225	230	233	235	237	239	240	241	242	245	252	253	259	261	262	263	264	265
270	272	274	278	280	281	286	287	288	289	291	293	294	295	297	300	302	306	310	311	312	316
318	322	323	324	326	328	329	330	331	332	340	341	342	344	348	351	353	363	367	370	381	382
384	387	388	390	395	400	408	414	422	423	425	426	431	433	435	461	463	464	470	473	477	482
483	484	485	491	501	505	506	511	512	517	525	536	538	540	543	548	550	551	553	565	574	577
580	582	587	588	589	592	602	604	605	617	623	626	629	645	662	681	693	703	722	725	731	738
740	744	748	749	756	761	771	774	776	778	781	782	793	797	801	802	805	813	816	820	826	830
831	839	847	848	853	863	873	875	886	889	893	895	901	905	907	912	922	923	931	934	941	942
956	962	964	971	988	1001	1025	1026	1038	1043	1047	1056	1062	1064								

Contig	Contig Size	Scaff	Seq	Qual	SeqX-LQ	BlastN	BlastX	CMC	CMS
1	880								

Type of Region	Start-PG	End-PG	Size-PG	Seq-PG	Qual-PG	Size-CG	Start-CG	End-CG	Orientation-CG	Seq-CG	Qual-CG
Specific	1	50	49						+		
Repeat	50	535	485			246427	246906	479	+		
Repeat	50	535	485			41021	41500	479	-		
Repeat	214	535	321			2478935	2479255	320	+		
Repeat	214	535	321			379822	380142	320	-		
Repeat	319	414	95			158621	158716	95	-		
Specific	535	880	345						+		

Figura 18: Tela que apresenta o relatório de todos os *contigs* isolados que: alinharam (em fundo cinza na tabela superior) e não alinharam (em fundo azul na tabela superior) ao genoma completo. As tabelas inferiores apresentam os detalhes das informações desses *contigs*, e as regiões de alinhamento e específica, quando essa informação for disponível. Cada cor representa o tipo de alinhamento (bloco comum, região específica ou rearranjo), conforme já descrito.

- ***Documentation***

Caso o usuário do sistema tenha eventuais dúvidas, uma documentação está disponível para auxiliá-lo. A documentação também possui um FAQ – Perguntas mais

freqüentes – para facilitar a uma rápida resposta às perguntas mais corriqueiras do usuário.

5.2 Resultados obtidos da aplicação do GINGA com o modelo biológico *Leifsonia xyli*

O modelo biológico que envolve o estudo de representantes da espécie de bactéria *Leifsonia xyli* foi utilizado como forma de validar o sistema GINGA. Entretanto, como já apresentado, o sistema GINGA pode ser utilizado em outros estudos comparativos entre genomas de organismos procariotos. *Leifsonia xyli* subsp. *xyli* (Lxx) é um patógeno de cana-de-açúcar e teve a seqüência do seu genoma completamente determinada [16]. A *Leifsonia xyli* subsp. *cynodontis* (Lxc) não é patogênica à cana-de-açúcar e apresenta seqüências do seu genoma resultantes de um projeto ainda em andamento. A montagem das seqüências utilizadas como teste foi realizada em 05/04/2006 e apresentaram os seguintes resultados: os 9.754 *reads* seqüenciados até essa data formaram 1.064 *contigs*, 317 *scaffolds* e restando 2.426 seqüências isoladas (*reads singlets*) (Figura 19) num total de 1.470.731 bases não redundantes do genoma de Lxc. As seqüências obtidas são provenientes de 4 tipos de bibliotecas genômicas diferentes (*Shotgun*, *BAC Ends*, *Sub-BAC* e Subtração). Dentre as informações disponibilizadas, o sistema apresenta o número de *reads* .b e .g e o total de clones (*reads* casados) que contém as duas extremidades seqüenciadas (Figura 19).

Assembly information	
Assembly:	05/04/2006 - 32
Total number of Scaffolds:	317
Total number of Contigs:	1064
Total number Contigs in Scaffold:	786
Total number of Isolated Contigs:	278
Total of Singlets:	2426

Genomic libraries information	
Total number of reads:	9754
BAC Sub:	2091
Subtraction:	185
Shotgun:	5854
BAC Ends:	1624

	Lib	Only .b	Only .g	Both ends
BAC Sub	L5	0	161	415
	8	1	1	0
	C2	302	266	165
	L1	0	288	96
	Lib	**Only .b**	**Only .g**	**Both ends**
Shotgun	01	0	0	288
	02	288	1	1343
	04	96	1	1055
	Lib	**Only .b**	**Only .g**	**Both ends**
BAC Ends	L7	33	362	214
	C1	0	0	96
	C3	0	0	96
	Lib	**Only .b**	**Only .g**	**Both ends**
Subtraction	S1	60	0	0
	S2	29	0	0
	S3	96	0	0

Figura 19: Tela que apresenta as informações de montagem como parte das informações disponibilizadas no relatório geral (*Overview*).

Dos 1.064 *contigs* formados, 786 (73,8%) foram agrupados em *scaffolds* e 278 (26%) ficaram isolados. Dentre os 786 *contigs* que estão em *scaffolds*, 680 *contigs* (63,9%) alinharam ao genoma de Lxx com os parâmetros de alinhamento utilizados e 106 *contigs* (9,9%) foram específicos ao genoma de Lxc.

Dentre os 278 *contigs* isolados, 183 (17%) alinharam ao genoma de Lxx, enquanto que 95 (8,9%) foram específicos (Figura 20). Portanto, do total de *contigs* da montagem do dia de 05/04/06 (1.064 *contigs*), 81% (863 *contigs*) puderam ser alinhados ao genoma de Lxx e ~19% (201 *contigs*) foram específicos ao genoma de Lxc. Os 201 *contigs* específicos representaram um total de 206.320 bases. Resultados esses que estão de acordo com o esperado em diferenças genéticas entre genomas de bactérias próximas [1,3,4,5,6]. As análises detalhadas de anotação dessas regiões podem vir a ajudar a compreender as diferenças no compatamento diferencial dessas

duas bactérias com relação ao hospedeiro da cana-de-açúcar. O GINGA apresenta, ainda, as informações sobre onde estas regiões estão ancoradas no genoma de Lxc em relação ao genoma de Lxx, uma vez que o sistema armazena as seqüências flanqueadoras de regiões específicas quando elas estão disponíveis. Essas informações estão armazenadas na tabela *CA_Specific* e *CA_NoSpecificNeighborParts* do banco de dados, e apresentam relacionamentos com as tabelas *CA_Organisms* e *CA_CrossmatchFirstAlign*.

Dentre os 317 *scaffolds* resultantes da montagem, 19 (5,9%) foram totalmente específicos correspondendo a um total de 56.884 bases. Durante o processo de montagem do genoma da Lxc e a construção dos *scaffolds*, optou-se por separar os *scaffolds* formados em partes cada vez que a ligação entre dois *contigs* fosse determinada somente por *reads* casados de bibliotecas de BAC. Essa estratégia foi necessária devido a uma limitação do programa para a construção de *scaffolds* utilizado, que aceita uma única ligação entre dois *contigs* para construção dos *scaffolds.* Cada vez que uma ligação era determinada por mais de um clone de BAC que apresenta insertos grandes, *contigs* ligados por *reads* casados de *shotgun* eram ignorados. Desta forma, *contigs* pequenos, que poderiam ser inseridos entre dois *contigs* ligados por *reads* casados de BACs, acabavam por ser considerados isolados. Assim, dos 317 *scaffolds*, 360 partes foram obtidas (Figura 20), sendo que 277 *scaffolds* não foram divididos, 37 *scaffolds* foram divididos em duas partes e 3 *scaffolds* em três partes. Entre as 360 partes, 331 (91%) puderam ser alinhadas ao genoma de Lxx, enquanto que 29 (8%) foram específicas. Essas regiões específicas representam inserções no genoma de Lxc ausentes no genoma de Lxx.

As seqüências alinhadas (*scaffolds* e *contigs* isolados) correspondem a 1.008.556 bases, sendo, aproximadamente, 40% do genoma de Lxx (Figura 21).

Information about the comparative analysis								
Comparative Organisms	Complete Genome	Partial Genome	No. of Scaffolds (pieces)			No. of Contigs Isolated		% of alignment
			Total Of Pieces	Align	Specific	Align	Specific	
Lxc X Lxx	Lxx	Lxc	360	331	29	183	95	39.02% (1008556pb)

Figura 20: Tela que apresenta as opções da tabela que faz parte do relatório geral (*Overview*) que apresenta dados gerais da comparação entre os genomas. No exemplo, apresentam-se dados da análise comparativa entre os genomas de Lxc (parcial) e Lxx (completo).

Figura 21: Tela que apresenta a uma representação gráfica da cobertura de ~40% (1.008.556 bases) referente a todas as regiões alinhadas do genoma parcial de Lxc com o genoma completo de Lxx. A barra horizontal branca representa o genoma de Lxx e as linhas verticais azuis representam regiões alinhadas do genoma de Lxc.

As inserções genômicas podem ser facilmente acompanhadas durante o processo de montagem quando utilizado o sistema GINGA. A representação gráfica do *scaffold* 005 de 5.707 bases é um exemplo desses casos (Figura 22). A seqüência do *contig* 929 (em destaque – em azul) é totalmente específica ao genoma de Lxc, pois não apresentou alinhamento com o genoma de Lxx. A composição de *reads* do *contig* 929 demonstra a confiabilidade na formação deste *contig*, pois é composto por *reads* de diferentes bibliotecas. Essa inserção no genoma de Lxc está ausente no genoma de Lxx. Esse exemplo demonstra também a qualidade do sistema quanto ao número ilimitado de tipos de bibliotecas genômicas que podem ser utilizadas durante o processo de seqüenciamento e visualizadas na análise comparativa. Essas informações também podem ser obtidas analisando-se os relatórios que ficam disponíveis pelo sistema.

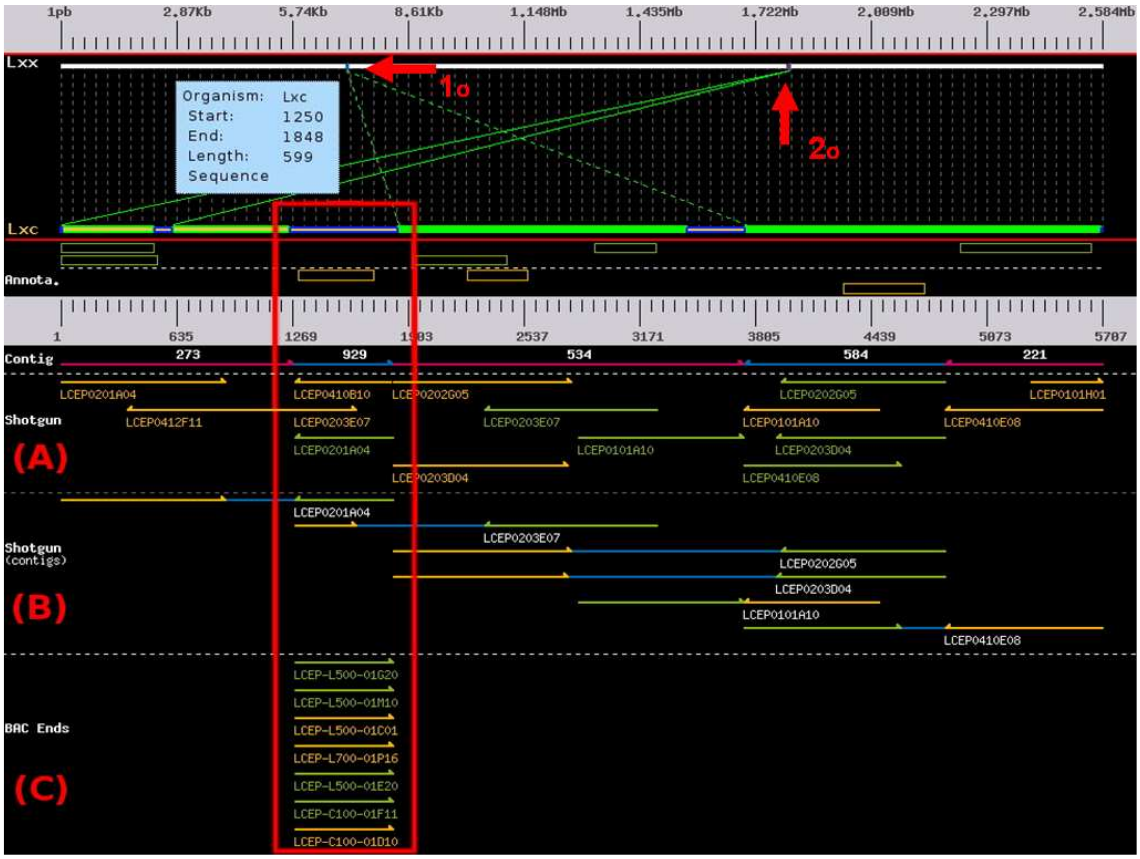


Figura 22: Tela que apresenta a representação gráfica do alinhamento entre o *scaffold* 005 da montagem do genoma de Lxc e o genoma completo de Lxx. Em destaque o *contig* 929 totalmente específico ao genoma de Lxc. (A) e (B) são os *reads* das bliblioteca de *Shotgun* e BAC que formaram cada *contig* e (C) os *reads* casados que fizeram a ligação entre os *contigs*. A caixa em azul mostra informações dessa região específica em destaque. Os itens 1° e 2° mostram dois grandes eventos de reorganização do genoma.

Esse mesmo exemplo (Figura 22) apresenta ainda um grande evento de reorganização. Parte do *scaffold* formado pelas seqüências do genoma de Lxc está mapeado na posição de 1805396 a 1806598 (número 2º - Figura 22) do genoma de Lxx, sendo que outra parte está mapeada na posição de 709620 a 713115 (número 1º - Figura 22). Isso indica que esta região do genoma de Lxc, apesar de comum ao genoma de Lxx, está organizada de maneira diferente. Esse tipo de resultado pode rapidamente ser avaliado e decisões quanto à continuidade do seqüenciamento nesta região podem ser tomadas.

Da mesma forma, podem-se acompanhar as ligações entre os *contigs* dentro de *scaffolds* analisando a composição dos *reads* que ligaram dois *contigs* e também a

composição de *reads* em cada *contig* (Figura 23).

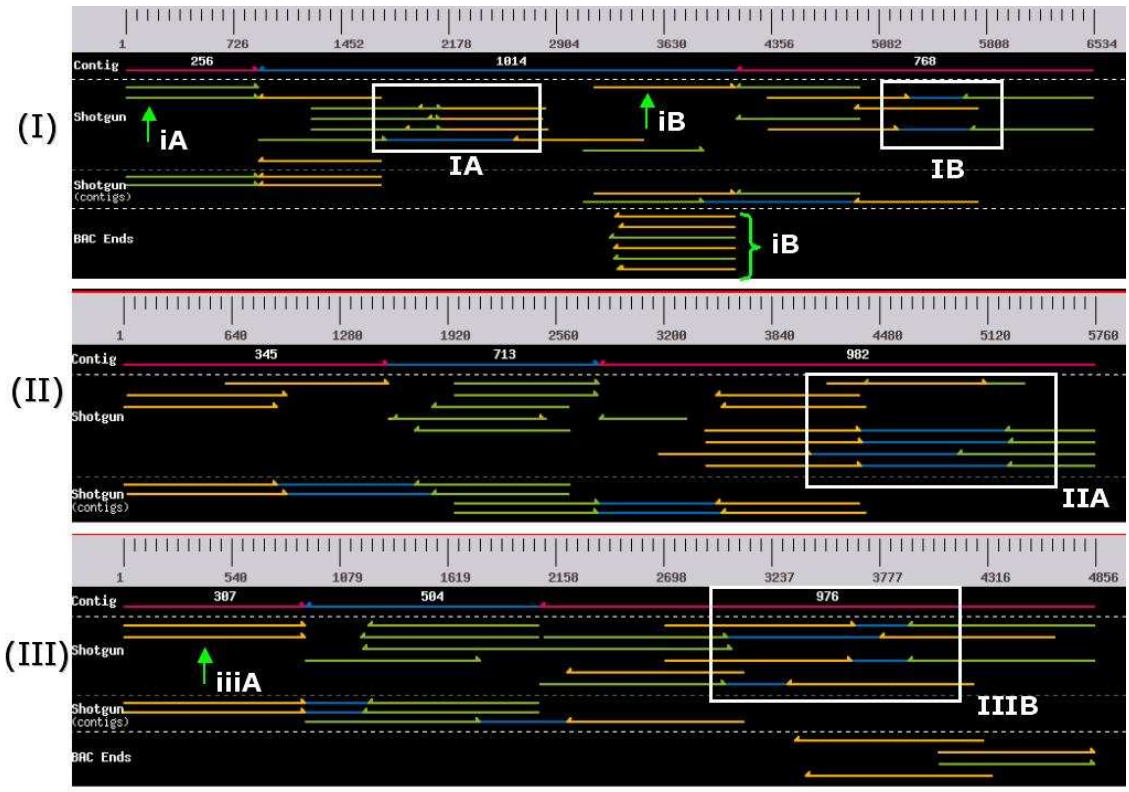


Figura 23: Exemplos de três casos (I, II e III) de como GINGA guia o processo de seqüênciamento e montagem do genoma parcial. Os exemplos, iA, iB, iiiA, IA, IB, IIA e IIIB, mostram casos de como *reads* .b e .g que podem formar *reads* casados e qual o *gap* (região em azul que liga os *reads*). Permite também visualizar o quanto uma região tem de cobertura.

A utilização do sistema permite a identificação de regiões de repetição (representado em vermelho/amarelo) (Figura 24). Repetições (idênticas ou não) podem ser facilmente interpretadas por intermédio da visualização e com a apresentação do posicionamento destas regiões repetidas no genoma completo. O posicionamento é apresentado em uma caixa contendo essas informações ao passar do *mouse*. Das 9 regiões que alinharam entre o *scaffold* 002 do genoma de Lxc e o genoma completo de Lxx, 8 regiões foram identificadas contendo ORFs que codificam transposases (Figura 24). Três das oito transposases (em destaque com setas em amarelo - Figura 24), foram identificadas por meio do alinhamento entre os genomas como presentes em uma mesma ilha genômica. Essa ilha ainda contém um total de 11 transposases e 6 tipos de IS diferentes (Figura 25).

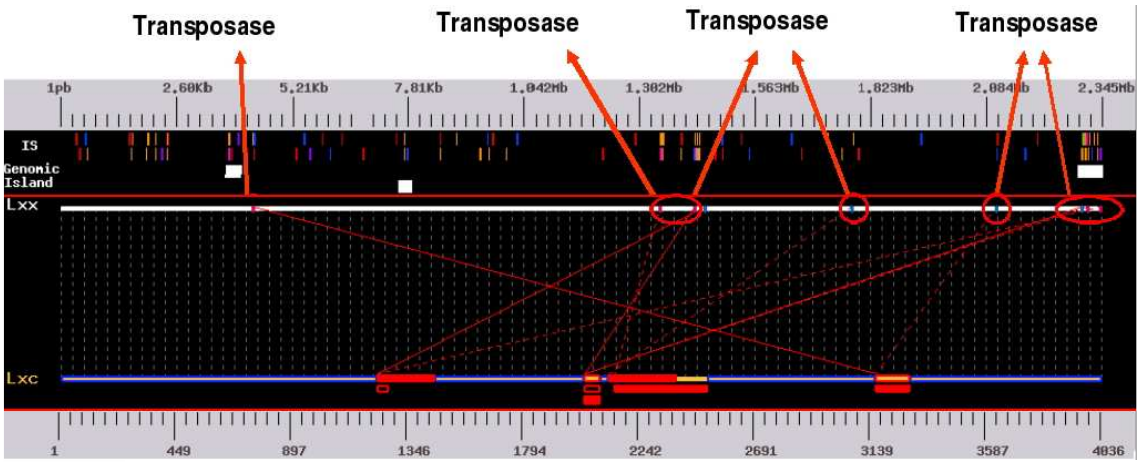


Figura 24: Exemplo do resultado de alinhamento entre o *scaffold* 2 de Lxc contra Lxx. Neste resultado do alinhamento tem-se 8 transposases identificadas referente as 9 regiões que alinharam em Lxx. As linhas tracejadas representam inversão das regiões alinhadas entre os genomas.

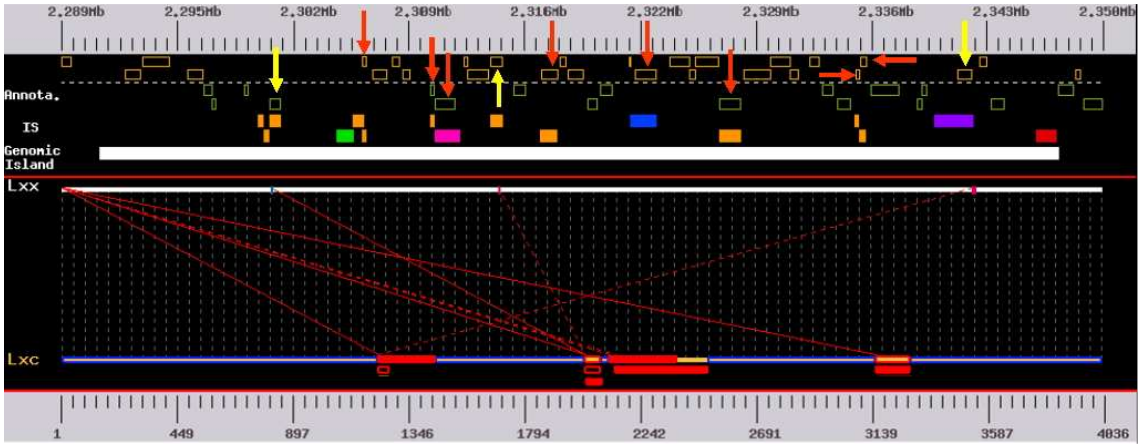


Figura 25: Exemplo do zoom de três regiões do *scaffold* 2 de Lxc que alinharam em contra Lxx. Nestra região identificou-se 11 transposases inseridas em 6 diferentes tipos de IS de uma única ilha genômica. Dessas 11 transposases, 3 estão em regiões sobreposta ao alinhamento entre os genomas (setas em amarelo) e outras 8 transposases visinhas e localizadas na mesma região (setas em vermelho). As linhas tracejadas entre as regiões alinhadas representam evento de inversão genômica.

O alinhamento pode sempre ser detalhado ulitizando as opções de *zoom* (Figura 26). Nesta figura foram apresentados também os três tipos de regiões (bloco comum, específica e rearranjo). Foram analisadas 3 sobreposições na região de 7.288 a 10.188 no genoma de Lxc (A, B e C), cada qual alinhada em um lugar diferente do genoma de Lxx (Figura 26). As regiões A, B e C (representadas em cor amarela) têm em média 1.793 pb, apesar de existir uma variação de 1.990 pb entre a maior e a menor repetição. A análise dos resultados de anotação permitiu identificar o tipo de repetição envolvida neste caso. A anotação do genoma de Lxx mostrou que a região

repetitiva contém um gene que codifica uma integrase associada a fago; as demais ORFs na vizinhança indicam a presença de uma região de fago (Figura 26).

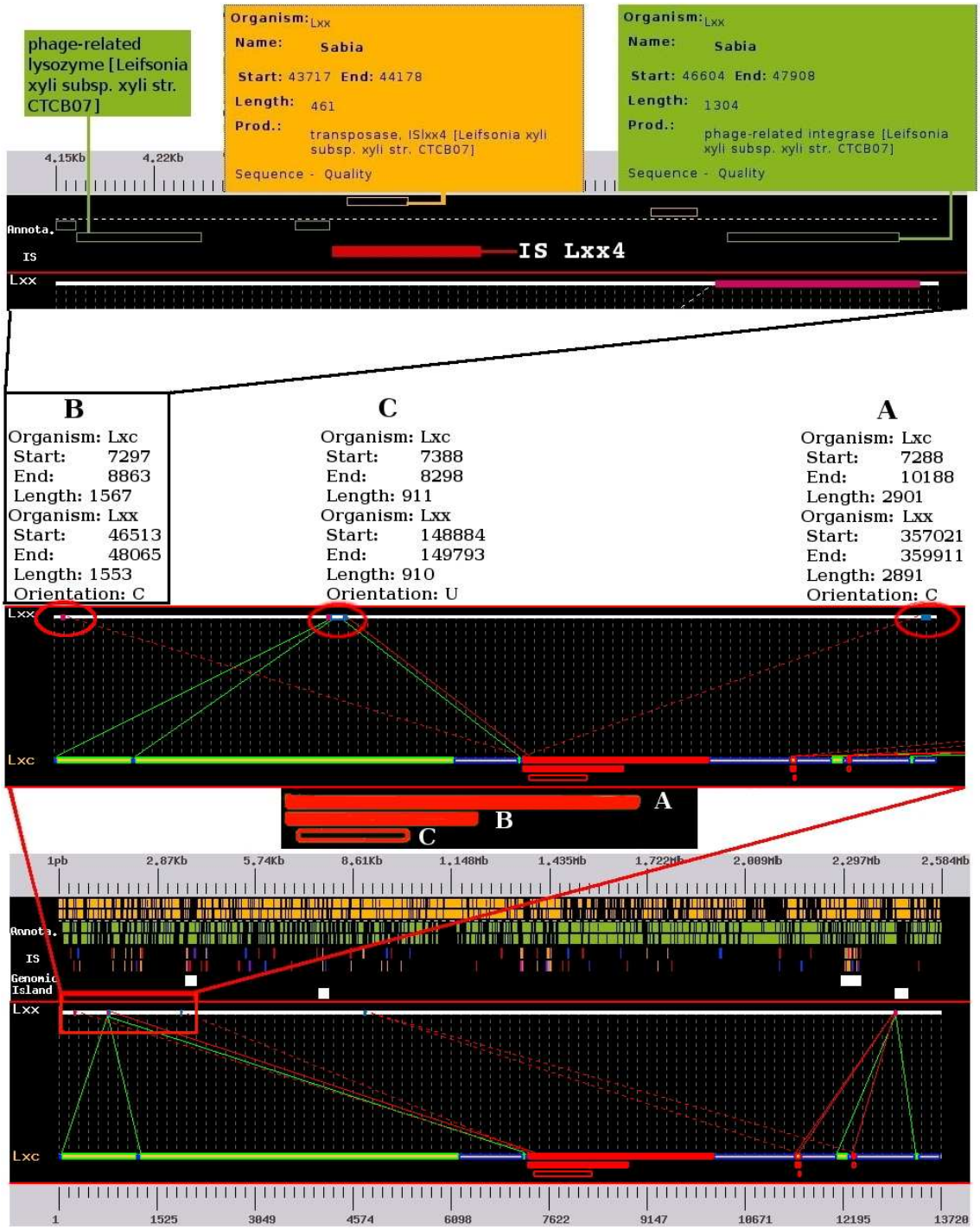


Figura 26: Exemplo de 3 rearranjos (em amarelo – A, B e C) do alinhamento entre o *scaffold* 148 de Lxc contra Lxx. A partir da opção de *zoom* da região B, pode-se obsevar uma possível região de fago.

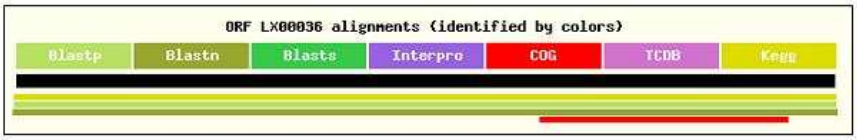
A. Início: 35.8429; Fim: 35.9733; Tamanho: 1.304; e Produto: *phage-related integrase* (*Leifsonia xyli* subsp. *xyli* str. CTCB07);

B. Início: 46.604; Fim: 47.908; Tamanho: 1.304; e Produto: *phage-related integrase* (*Leifsonia xyli* subsp. *xyli* str. CTCB07);

C. Início:148.962; Fim: 149.885; Tamanho: 923; e Produto: *phage-related integrase* (*Leifsonia xyli* subsp. *xyli* str. CTCB07).

Em particular, é apresentado um detalhe da região do alinhamento da repetição B e a presença da ORF que contém como produto gênico de uma integrase (Figura 26 parte superior). Ao lado tem-se duas outras ORFs, sendo: (a) em laranja, uma ORF que codifica a transposase do ISLxx4 (sequência de inserção); e (b) em verde, outra ORF que está associada a fago (Figura 26). Cada uma das ORFs representadas na seqüência do genoma completo e do genoma parcial, no caso específico com a utilização do sistema SABIA, apresentam uma ligação para a página de anotação desse sistema (Figura 27 e 28).

Functional annotation | InterPro | GO (Gene Onthology) | Blast | Psort | Annotation | Search



ORF information

ORF ID	LX00036 (frame -3)	Origin	Glimmer (Contig 1) (Old | New)
Position and sequences	complement(46604...47908) (1305 bp) (435 aa)	Upstream extragenic region	-
Molecular weight	0.00	Theoretical pI	0.00
Optional start codon	-	Nucleotides percentage	A (20.23%) | C (32.64%) | G (31.49%) | T (15.63%)
Percent CG	64.13%	Percent AT	35.86%
Overlaps	-		

Transcriptional regulation

RBS

New start position	Stop position	RBS pattern	RBS position	New start codon	Shift	Old start codon	Old start position
47908	46604	GGGAG	47915	ATG	0	ATG	47908

Promoter

Box -35	distance to	Box -10	Distance from ORF
GTGATG	19	TGCAAG	50
TTGAGG	17	TAGGGT	42
GTTTCA	18	GATGGT	8

Functional annotation

KEGG database

Organism	*L.xyli*	Gene entry	Lxx03570
Gene name	-	Synonyms	-
Definition	phage-related integrase		
EC number	-		
Class	-		
Classification	Unassigned		
DBLinks	GI		

COG Clusters of Orthologous Groups of proteins

	Gene	Classification	Coords (size)	Product
COG0582	*XerC*	L	278 .. 409 (132)	Integrase

Blast results (KEGG)

	BlastN (output)		BlastP (output)	
	L.xyli	*E.coli*	*L.xyli*	
Score	2587	-	885	-
Expect	0.0	-	0.0	-
Query coverage	100.00%	-	100.00%	-
Subject coverage	100.00%	-	100.00%	-

InterPro

No results found !

Figura 27: Tela que apresenta uma primeira parte das informações de anotação manual do sistema SABIA apresentando o exemplo da ORF de 46.604pb a 47.908pb do *Scaffold* 148.

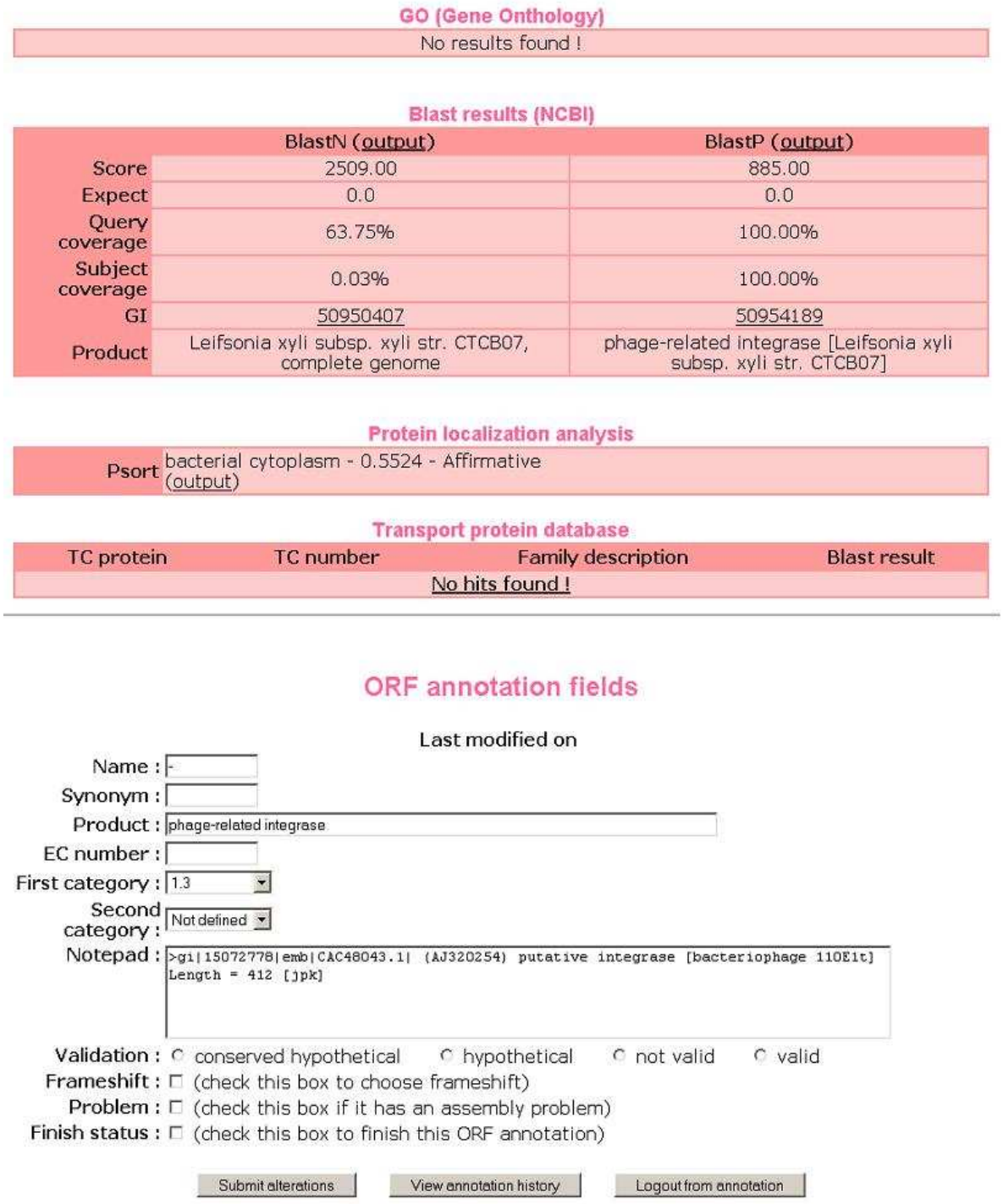


Figura 28: Tela que apresenta uma segunda parte das informações de anotação manual do sistema SABIA, apresentando o exemplo da ORF de 46.604pb a 47.908pb do *Scaffold* 148.

Inversões puderam também ser visualizadas (Figura 29). Neste exemplo, duas regiões de blocos comuns (1° e 2° Figura 29) estão alinhadas ao genoma de Lxx juntas, uma ao lado da outra (em cor magenta e azul), localizados em 711.534 a 713.115 e em 709.620 a 711.560.

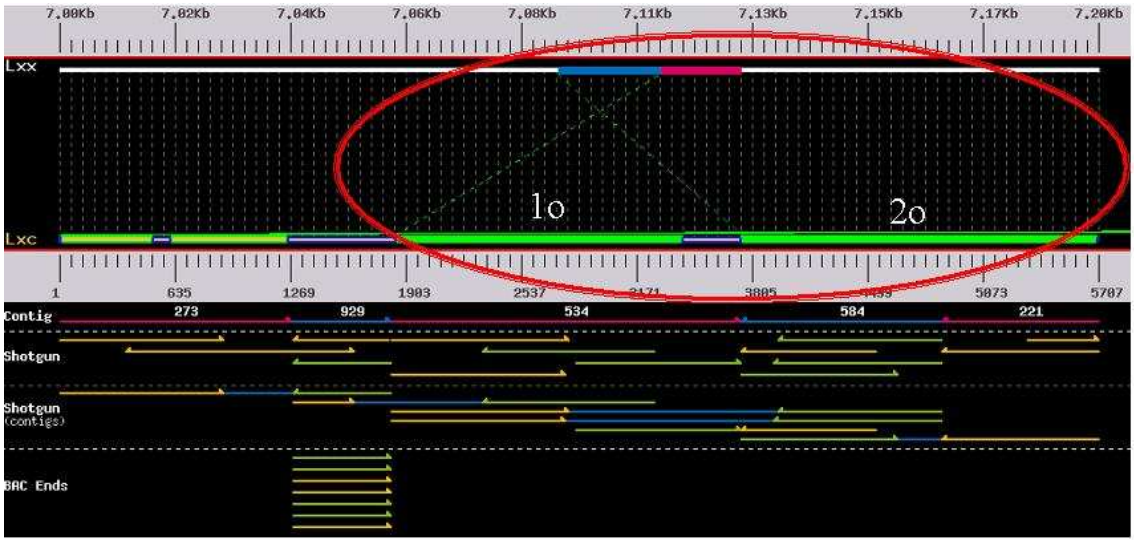


Figura 29: Tela que apresenta o resultado do alinhamento do *Scaffold* 005 apresentando as regiões de blocos comuns 1° e 2° com o alinhamento em orientação invertida ao genoma de Lxx. Além disso, a região específica entre 1° e 2° indica uma possível inserção no genoma parcial.

O relatório apresentado permitiu a identificação de pelo menos 6 regiões envolvidas em grandes eventos de reorganização nos genomas de Lxc e Lxx (Figura 30). As coordenadas de todos os alinhamentos são apresentadas no relatório, e a coloração diferencial das células (linhas) cada vez que um bloco comum apresentava uma quebra de colinearidade ajudou na identificação desses eventos. Foram consideradas somente as quebras na colinearidade que envolviam *contigs* com pelo menos 800 pb e alinhamentos envolvendo pelo menos 800 pb (Figura 30).

Scaffold: 000

Contig	Contig size	Alignment (bp) - PG	Start - CG	End - CG	Alignment (bp) - CG	Gap	Orientation	Repeats
957	1078	78	111876	111954	78	24	+	-
957	1078	733	111978	112736	758	3666	+	-
1036	1150	695	116402	117095	693	464701	+	-
107	841	645	581796	582442	646	602	-	-
591	1138	1125	583044	584174	1130	559	-	-
760	996	890	584733	585630	897	1635105	-	-
859	912	907	2220735	2221645	910	3702	-	-
1051	3314	576	2225347	2225923	576	696	-	-
759	1061	498	2226619	2227116	497		-	-

Scaffold: 005

Contig	Contig size	Alignment (bp) - PG	Start - CG	End - CG	Alignment (bp) - CG	Gap	Orientation	Repeats
584	1105	1949	709620	711560	1940	-	+	-
221	859	1949	709620	711560	1940	-	+	-
534	1920	1574	711534	713115	1581	1092281	+	-
273	1281	507	1805396	1805903	507	52	-	-
273	1281	639	1805955	1806598	643		-	-

Scaffold: 036

Contig	Contig size	Alignment (bp) - PG	Start - CG	End - CG	Alignment (bp) - CG	Gap	Orientation	Repeats
497	1352	1331	368578	369908	1330	34378	+	-
362	1340	2602	404286	406905	2619	-	+	-
958	1274	2602	404286	406905	2619		+	-

Scaffold: 071

Contig	Contig size	Alignment (bp) - PG	Start - CG	End - CG	Alignment (bp) - CG	Gap	Orientation	Repeats
560	957	529	1464119	1464643	524	44	+	-
560	957	129	1464687	1464816	129	942397	+	-
1049	4373	403	2407213	2407618	405	11820	-	-
1049	4373	2725	2419438	2422194	2756		+	-

Scaffold: 075

Contig	Contig size	Alignment (bp) - PG	Start - CG	End - CG	Alignment (bp) - CG	Gap	Orientation	Repeats
862	987	923	651105	652031	926	504268	-	-
1018	1152	1105	1156299	1157407	1108	763	+	-
184	940	867	1158170	1159036	866		+	-

Scaffold: 087

Contig	Contig size	Alignment (bp) - PG	Start - CG	End - CG	Alignment (bp) - CG	Gap	Orientation	Repeats
520	1953	1766	878180	879946	1766	445923	-	-
984	1895	1845	1325869	1327715	1846	134	+	-
755	992	960	1327849	1328809	960	192	+	-
520	1953	109	1329001	1329113	112		+	-

Figura 30: Tela que apresenta a informação de 6 grandes rearranjos na organização entre os genomas identificados a partir do relatório visão macro (*Macro Vision*). A sigla PG refere-se a *Partial Genome* e GC a *Complete Genome*.

Todos os exemplos e a maneira como são apresentados refletem o resultado de como as informações foram organizadas no banco de dados e como elas foram tratadas pelo sistema. O sistema GINGA busca integrar toda a base de informação para facilitar a análise desses dados, disponibilizando-os de uma forma padronizada e organizada.

5.3 Resultado da performance do sistema utilizando o modelo biológico

Na comparação do modelo biológico de Lxc e Lxx foram utilizadas as seqüências dos 317 *scaffolds* e 278 *contigs* isolados*,* referentes a uma montagem escolhida, no caso de número 32. Esses *scaffolds* e *contigs* isolados foram comparados contra o genoma completo da Lxx que contém uma seqüência de 2.584.158 pb. A Tabela 10 apresenta o resultado do tempo de execução da análise comparativa separado por *scaffolds*, *contigs* isolados, e, depois, ambos.

Tabela 10: Tempo de execução da análise comparativa dos *scaffolds* e *contigs* isolados do genoma parcial de Lxc contra o genoma completo de Lxx.

Seqüência	***Tempo***
Scaffolds	6 horas
Contigs Isolados	1 hora
Scaffolds + *Contigs* Isolados	7 horas

É importante observar que o tempo de execução dependerá da quantidade e do tamanho médio das seqüências comparadas e do número de rearranjos que disparam novas análises.

5.4 Análise comparativa do GINGA com outros sistemas

Conforme apresentado, existe uma coleção de sistemas que podem ser utilizados para a análise comparativa de genomas; cada um desses sistemas apresenta suas características, usam métodos de análises diferentes e integram-se a ferramentas e banco de dados particulares.

Dentre os cinco sistemas analisados, quatro utilizam BLAST [63] (ou variações) para fazer o alinhamento de seqüências. Dentre esses quatro, dois utilizam outras ferramentas associadas ao BLAST sendo: PatternHunter [122,123] (COMBO) e MSPcrunch [124] (ACGT). O GenAlyzer utiliza somente o programa Vmatch para fazer o alinhamento. O ACT aceita, além de alinhamento local resultante de BLAST, resultados de alinhamento global do programa MUMmer [95,96,97]. O GINGA utiliza o programa *cross_match*, de alinhamento local, para alinhar as seqüências.

As informações de entrada para cada um dos sistemas, em sua maioria, têm como base arquivos contendo seqüências em formato FASTA, Genbank ou EMBL. As informações de anotação são extraídas do Genbank (arquivo no formato .ptt) ou de arquivos no formato .GFF. Apenas os sistemas ACGT, ACT, COMBO, GenAlyzer e GINGA disponibilizam informação de anotação. ACGT e GenAlyzer utilizam a informação de anotação obtida de um arquivo de entrada fornecido pelo usuário. Desses sistemas, COMBO e GINGA estão integrados a algum sistema de anotação possibilitando a anotação manual. O ACT possui integração com o visualizador de anotação Artemis [91], que disponibiliza diversas funções. ACT e GINGA disponibilizam informações complementares como conteúdo GC dos genomas. Ainda, GINGA acrescenta informações de ilha genômica e IS (quando disponível) e o ACT[10] permite marcar informações nas seqüências que estão sendo analisadas. Esse sistema apresenta também informações de viés na utilização dos códons e nas assinaturas de dinucleotídeos.

O programa BACCardI, além do SABIA, é o único que aceita informações de montagem e arquivos em formato .ace, resultantes do programa Phrap [32]. Esse sistema permite a comparação de informações de clones e de insertos seqüenciados de uma montagem e utiliza o genoma completo para o mapeamento

dessas seqüências. Entretanto, BACCardl aceita somente informações de bibliotecas genômicas de insertos grandes como as construídas em vetores como BAC e cosmídeo/fosmídeo. O sistema GINGA pode utilizar qualquer tipo de biblioteca genômica para organismos procariotos, sem limitação quanto ao tamanho. Além disso, GINGA disponibiliza informação de *reads*/clones e bibliotecas genômicas utilizadas na montagem via integração com o sistema SABIA.

As representações gráficas mais utilizadas são circular, paralela ou *dot plot* (como disponibilizado pelo MUMmer [95,96,97]). A visão paralela predomina em quatro (COMBO, ACGT, ACT e GenAlyzer) dos cinco sistemas analisados**.** BACCardl é o único sistema que possui uma visualização circular. A visualização linear do BACCardl possibilita uma análise mais detalhada de regiões do genoma. Apesar das pequenas variações GINGA extende a apresentação dos resultados além da representação gráfica através de diversos relatórios.

O principal diferencial presente no sistema GINGA em relação aos outros sistemas é a centralização das diversas informações de genômica comparativa em um único sistema. Informações como a integração de resultados de alinhamento entre genomas com informações de anotação, além das informações de seqüenciamento e montagem do genoma em andamento de maneira comparativa à seqüência do genoma completo.

6 CONCLUSÕES E PERSPECTIVAS

A análise comparativa entre genomas é uma opção de estudo para que se possa entender a relação biológica existente entre os organismos. Sistemas computacionais com aplicação dos mais diversos métodos e algoritmos existem para auxiliar nessa tarefa. Hoje, o grande desafio é integrar toda a base de conhecimento de diversas áreas e meios de modo adequado, rápido e fácil para que sejam analisadas. Esse é também um dos maiores desafios da genômica, que é responsável por uma grande quantidade de dados que são gerados de maneira independente nos mais diversos centros de pesquisa do mundo.

Considerando apenas organismos procariotos (em bactérias e arqueobactérias), mais de mil genomas estão sendo seqüenciados atualmente. Existem diversas ferramentas computacionais disponíveis para serem utilizadas nas análises comparativas entre genomas. Cada ferramenta possui seus objetivos, peculiaridades e aplicação. A necessidade do usuário é um fator determinante na escolha da ferramenta que será utiliza. No presente trabalho foi desenvolvido o sistema GINGA, que busca ser um sistema voltado para a análise comparativa entre genomas procariotos, disponibilizando a informação por intermédio de uma *interface web* amigável ao usuário. O diferencial do GINGA está em dois principais fatores: (a) a apresentação das informações de um genoma parcialmente seqüênciado de maneira comparativa a um outro completamente sequênciado, sendo ambos genomas próximos do ponto de vista filogenético (~80% de identidade); e (b) a integração com diversas fontes de informação, no caso, informações de seqüenciamento e montagem

de genoma, anotação e análise comparativa de genomas centralizada num único sistema.

A utilidade do sistema foi demonstrada na análise comparativa do genoma de duas bactérias (Lxc e Lxx) importantes para o setor agrícola. Esse modelo foi um passo inicial e importante para mostrar a aplicação e as funções desse sistema, além de ajudar a testá-lo. Porém, outros modelos biológicos deverão ser utilizados até mesmo para ajudar a visualizar e definir novas funções ao sistema. Por fim, as conclusões da proposta e objetivos traçados para o sistema GINGA são:

1. A organização do banco de dados e as tabelas com seus relacionamentos definidos permitem que o sistema apresente as informações de interesse definidas na proposta.

2. A Integração ao sistema SABIA permite a visualização das informações de montagem e anotação nas telas de comparação.

3. A utilização do sistema GINGA na busca e na visualização das diferenças genéticas entre Lxc e Lxx leva a definição de que os dois genomas apresentam pelo menos 80% de regiões comuns, sendo que Lxc contém 20% de regiões específicas, e pelo menos grandes eventos de reorganização.

Referente às perspectivas quanto às funções a serem incluídas no sistema, destacam-se:

- Na representação gráfica, a possibilidade de buscar onde estão as duas pontas .b e .g de um *read* não casado (que não se ligou virtualmente, formando clone).

- A utilização de outros modelos biológicos de organismos procariotos.

- Disponibilizar os resultados de blastx contra o banco de dados NR do GenBank [51,52] das regiões alinhadas e não alinhadas do genoma parcial para uma rápida inspeção.

- Anexar ao GINGA uma visualização gráfica para a análise comparativa entre proteínas.

- Disponibilizar o GINGA como um serviço web em que qualquer pesquisador possa usar o sistema a partir de nossos servidores sem a necessidade de instalá-lo, por meio da submissão das informações (seqüências, anotação etc) dos genomas de interesse.

- Por fim, validar o sistema GINGA a partir de seqüência parcial de um genoma que já foi completamente seqüenciado e validado.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] BINNEWIES, T. T., MOTRO, Y., HALLIN, P. F., *et al.* "Ten years of bacterial genome sequencing: comparative-genomics-based discoveries". *Funct Integr Genomics*, v.6, n.3, pp.165-85, Jul. 2006.

[2] LAN, R. e REEVES, P. R. "Intraspecies variation in bacterial genomes: the need for a species genome concept". *Trends Microbiol*, v.8, n.9, pp.396-401, Sep. 2000.

[3] ALM, R. A., LING, L. S., MOIR, D. T., *et al.* "Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori". *Nature*, v.397, n.6715, pp.176-80, Jan 14. 1999.

[4] PARKHILL, J., ACHTMAN, M., JAMES, K. D., *et al.* "Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491". *Nature*, v.404, n.6777, pp.502-6, Mar 30. 2000.

[5] TAKAMI, H., NAKASONE, K., TAKAKI, Y., *et al.* "Complete genome sequence of the alkaliphilic bacterium Bacillus halodurans and genomic sequence comparison with Bacillus subtilis". *Nucleic Acids Res*, v.28, n.21, pp.4317-31, Nov 1. 2000.

[6] FLEISCHMANN, R. D., ALLAND, D., EISEN, J. A., *et al.* "Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains". *J Bacteriol*, v.184, n.19, pp.5479-90, Oct. 2002.

[7] BARTELS, D., KESPOHL, S., ALBAUM, S., *et al.* "BACCardI--a tool for the validation of genomic assemblies, assisting genome finishing and intergenome comparison". *Bioinformatics*, v.21, n.7, pp.853-9, Apr 1. 2005.

[8] ENGELS, R., YU, T., BURGE, C., *et al.* "Combo: a whole genome comparative browser". *Bioinformatics*, v.22, n.14, pp.1782-3, Jul 15. 2006.

[9] XIE, T. e HOOD, L. "ACGT-a comparative genomics tool". *Bioinformatics*, v.19, n.8, pp.1039-40, May 22. 2003.

[10] CARVER, T. J., RUTHERFORD, K. M., BERRIMAN, M., *et al.* "ACT: the Artemis Comparison Tool". *Bioinformatics*, v.21, n.16, pp.3422-3, Aug 15. 2005.

[11] ABBOTT, J. C., AANENSEN, D. M., RUTHERFORD, K., *et al.* "WebACT--an online companion for the Artemis Comparison Tool". *Bioinformatics*, v.21, n.18, pp.3665-6, Sep 15. 2005.

[12] CHOUDHURI, J. V., SCHLEIERMACHER, C., KURTZ, S., *et al.* "GenAlyzer: interactive visualization of sequence similarities between entire genomes". *Bioinformatics*, v.20, n.12, pp.1964-5, Aug 12. 2004.

[13] ALMEIDA, L. G., PAIXAO, R., SOUZA, R. C., *et al.* "A System for Automated Bacterial (genome) Integrated Annotation--SABIA". *Bioinformatics*, v.20, n.16, pp.2832-3, Nov 1. 2004.

[14] *Agrianual 2002: anuário da agricultura brasileira*. FNP – Consultoria & Comércio. São Paulo, p.537, 2002.

[15] ALBINO, J. D. C., CRESTE, S. e FIGUEIRA, A. "Mapeamento genético da Cana-de-açúcar". *Biotecnologia Ciência & Desenvolvimento*, n.36, Jan/Jun. 2006.

[16] MONTEIRO-VITORELLO, C. B., CAMARGO, L. E., VAN SLUYS, M. A., *et al.* "The genome sequence of the gram-positive sugarcane pathogen Leifsonia xyli subsp. xyli". *Mol Plant Microbe Interact*, v.17, n.8, pp.827-36, Aug. 2004.

[17] WATSON, J. D. e CRICK, F. H. "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid". *Nature*, v.171, n.4356, pp.737-8, Apr 25. 1953.

[18] SANGER, F., NICKLEN, S. e COULSON, A. R. "DNA sequencing with chain-terminating inhibitors". *Proc Natl Acad Sci U S A*, v.74, n.12, pp.5463-7, Dec. 1977.

[19] FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O.*, et al.* "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd". *Science*, v.269, n.5223, pp.496-512, Jul 28. 1995.

[20] MARGULIES, M., EGHOLM, M., ALTMAN, W. E.*, et al.* "Genome sequencing in microfabricated high-density picolitre reactors". *Nature*, v.437, n.7057, pp.376-80, Sep 15. 2005.

[21] SHENDURE, J., PORRECA, G. J., REPPAS, N. B.*, et al.* "Accurate multiplex polony sequencing of an evolved bacterial genome". *Science*, v.309, n.5741, pp.1728-32, Sep 9. 2005.

[22] SIMPSON, A. J., REINACH, F. C., ARRUDA, P.*, et al.* "The genome sequence of the plant pathogen Xylella fastidiosa. The Xylella fastidiosa Consortium of the Organization for Nucleotide Sequencing and Analysis". *Nature*, v.406, n.6792, pp.151-9, Jul 13. 2000.

[23] DA SILVA, A. C., FERRO, J. A., REINACH, F. C.*, et al.* "Comparison of the genomes of two Xanthomonas pathogens with differing host specificities". *Nature*, v.417, n.6887, pp.459-63, May 23. 2002.

[24] VETTORE, A. L., DA SILVA, F. R., KEMPER, E. L.*, et al.* "Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane". *Genome Res*, v.13, n.12, pp.2725-35, Dec. 2003.

[25] NASCIMENTO, A. L., KO, A. I., MARTINS, E. A.*, et al.* "Comparative genomics of two Leptospira interrogans serovars reveals novel insights into physiology and pathogenesis". *J Bacteriol*, v.186, n.7, pp.2164-72, Apr. 2004.

[26] NASCIMENTO, A. L., VERJOVSKI-ALMEIDA, S., VAN SLUYS, M. A.*, et al.* "Genome features of Leptospira interrogans serovar Copenhageni". *Braz J Med Biol Res*, v.37, n.4, pp.459-77, Apr. 2004.

[27] VIEIRA, L. G. E., ANDRADE, C. A., COLOMBO, C. A.*, et al.* "Brazilian coffee genome project: an EST-based genomic resource". *Braz. J. Plant Physiol.*, v.18, n.1. 2006.

[28] VERJOVSKI-ALMEIDA, S., DEMARCO, R., MARTINS, E. A.*, et al.* "Transcriptome analysis of the acoelomate human parasite Schistosoma mansoni". *Nat Genet*, v.35, n.2, pp.148-57, Oct. 2003.

[29] "The complete genome sequence of Chromobacterium violaceum reveals remarkable and exploitable bacterial adaptability". *Proc Natl Acad Sci U S A*, v.100, n.20, pp.11660-5, Sep 30. 2003.

[30] VASCONCELOS, A. T., FERREIRA, H. B., BIZARRO, C. V.*, et al.* "Swine and poultry pathogens: the complete genome sequences of two strains of Mycoplasma hyopneumoniae and a strain of Mycoplasma synoviae". *J Bacteriol*, v.187, n.16, pp.5568-77, Aug. 2005.

[31] XU, J. "Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances". *Mol Ecol*, v.15, n.7, pp.1713-31, Jun. 2006.

[32] GREEN, P. Documentation for Phrap. Disponível em: http://bozeman.mbt.washington.edu/phredphrap/general.html. Acesos em: 11 de fevereiro de 2007

[33] HUANG, X. e MADAN, A. "CAP3: A DNA sequence assembly program". *Genome*

Res, v.9, n.9, pp.868-77, Sep. 1999.

[34] SUTTON, G., WHITE, O., ADAMS, M.*, et al.* "TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects". *Genome Science & Technology*, v.1, n.1, pp.9-19. 1995.

[35] POP, M. e KOSACK, D. "Using the TIGR assembler in shotgun sequencing projects". *Methods Mol Biol*, v.255, pp.279-94. 2004.

[36] MYERS, E. W. *A suite of unix filters for gragment assembly*. Depto. of CS. U. of Arizona, Tucson, AZ, TR96-07, 1996.

[37] MYERS, E. W., JAIN, M., ANSON, E.*, et al. An Interface for a Fragment Assembly Kernel*. Depto. of CS. U. of Arizona, Tucson, AZ, TR96-04, 1996.

[38] DEAR, S. e STADEN, R. "A sequence assembly and editing program for efficient management of large projects". *Nucleic Acids Res*, v.19, n.14, pp.3907-11, Jul 25. 1991.

[39] CHEN, T. e SKIENA, S. STROLL: A new fragment assembly program. The Eighth Symposium on Combinatorial Pattern Matching: Aarhus, Denmark 1997.

[40] CHEN, T. e SKIENA, S. Trie-based data structures for fragment assembly. The Eighth Symposium on Combinatorial Pattern Matching: Aarhus, Denmark 1997.

[41] SMITH, T. F. e WATERMAN, M. S. "Identification of common molecular subsequences". *J Mol Biol*, v.147, n.1, pp.195-7, Mar 25. 1981.

[42] EWING, B. e GREEN, P. "Base-calling of automated sequencer traces using phred. II. Error probabilities". *Genome Res*, v.8, n.3, pp.186-94, Mar. 1998.

[43] EWING, B., HILLIER, L., WENDL, M. C.*, et al.* "Base-calling of automated sequencer traces using phred. I. Accuracy assessment". *Genome Res*, v.8, n.3, pp.175-85, Mar. 1998.

[44] GORDON, D., ABAJIAN, C. e GREEN, P. "Consed: a graphical tool for sequence finishing". *Genome Res*, v.8, n.3, pp.195-202, Mar. 1998.

[45] HUANG, X. "A contig assembly program based on sensitive detection of fragment overlaps". *Genomics*, v.14, n.1, pp.18-25, Sep. 1992.

[46] MUKHERJEE, S. e MITRA, S. "Hidden Markov Models, grammars, and biology: a tutorial". *J Bioinform Comput Biol*, v.3, n.2, pp.491-526, Apr. 2005.

[47] DELCHER, A. L., HARMON, D., KASIF, S.*, et al.* "Improved microbial gene identification with GLIMMER". *Nucleic Acids Res*, v.27, n.23, pp.4636-41, Dec 1. 1999.

[48] BORODOVSKY, M. e MCININCH, J. "GeneMark: parallel gene recognition for both DNA strands". *Comput. Chem*, v.19, pp.123-133. 1993.

[49] BESEMER, J. e BORODOVSKY, M. "GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses". *Nucleic Acids Res*, v.33, n.Web Server issue, pp.W451-4, Jul 1. 2005.

[50] LARSEN, T. S. e KROGH, A. "EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance". *BMC Bioinformatics*, v.4, pp.21, Jun 3. 2003.

[51] WHEELER, D. L., BARRETT, T., BENSON, D. A.*, et al.* "Database resources of the National Center for Biotechnology Information". *Nucleic Acids Res*, v.34, n.Database issue, pp.D173-80, Jan 1. 2006.

[52] BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J.*, et al.* "GenBank". *Nucleic Acids Res*, v.34, n.Database issue, pp.D16-20, Jan 1. 2006.

[53] TATENO, Y., IMANISHI, T., MIYAZAKI, S.*, et al.* "DNA Data Bank of Japan (DDBJ) for genome scale research in life science". *Nucleic Acids Res*, v.30, n.1, pp.27-

30, Jan 1. 2002.

[54] OKUBO, K., SUGAWARA, H., GOJOBORI, T., *et al.* "DDBJ in preparation for overview of research activities behind data submissions". *Nucleic Acids Res*, v.34, n.Database issue, pp.D6-9, Jan 1. 2006.

[55] BIRNEY, E., ANDREWS, D., CACCAMO, M., *et al.* "Ensembl 2006". *Nucleic Acids Res*, v.34, n.Database issue, pp.D556-61, Jan 1. 2006.

[56] KANEHISA, M. "A database for post-genome analysis". *Trends Genet*, v.13, n.9, pp.375-6, Sep. 1997.

[57] KANEHISA, M. e GOTO, S. "KEGG: kyoto encyclopedia of genes and genomes". *Nucleic Acids Res*, v.28, n.1, pp.27-30, Jan 1. 2000.

[58] KANEHISA, M., GOTO, S., HATTORI, M., *et al.* "From genomics to chemical genomics: new developments in KEGG". *Nucleic Acids Res*, v.34, n.Database issue, pp.D354-7, Jan 1. 2006.

[59] BRU, C., COURCELLE, E., CARRERE, S., *et al.* "The ProDom database of protein domain families: more emphasis on 3D". *Nucleic Acids Res*, v.33, n.Database issue, pp.D212-5, Jan 1. 2005.

[60] BATEMAN, A., COIN, L., DURBIN, R., *et al.* "The Pfam protein families database". *Nucleic Acids Res*, v.32, n.Database issue, pp.D138-41, Jan 1. 2004.

[61] LETUNIC, I., COPLEY, R. R., SCHMIDT, S., *et al.* "SMART 4.0: towards genomic data integration". *Nucleic Acids Res*, v.32, n.Database issue, pp.D142-4, Jan 1. 2004.

[62] BOECKMANN, B., BAIROCH, A., APWEILER, R., *et al.* "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003". *Nucleic Acids Res*, v.31, n.1, pp.365-70, Jan 1. 2003.

[63] ALTSCHUL, S. F., GISH, W., MILLER, W., *et al.* "Basic local alignment search tool". *J Mol Biol*, v.215, n.3, pp.403-10, Oct 5. 1990.

[64] PEARSON, W. R. e LIPMAN, D. J. "Improved tools for biological sequence comparison". *Proc Natl Acad Sci U S A*, v.85, n.8, pp.2444-8, Apr. 1988.

[65] STOTHARD, P. e WISHART, D. S. "Automated bacterial genome analysis and annotation". *Curr Opin Microbiol*, v.9, n.5, pp.505-10, Oct. 2006.

[66] TAKEUCHI, F., WATANABE, S., BABA, T., *et al.* "Whole-genome sequencing of staphylococcus haemolyticus uncovers the extreme plasticity of its genome and the evolution of human-colonizing staphylococcal species". *J Bacteriol*, v.187, n.21, pp.7292-308, Nov. 2005.

[67] BUELL, C. R., JOARDAR, V., LINDEBERG, M., *et al.* "The complete genome sequence of the Arabidopsis and tomato pathogen Pseudomonas syringae pv. tomato DC3000". *Proc Natl Acad Sci U S A*, v.100, n.18, pp.10181-6, Sep 2. 2003.

[68] SETUBAL, J. C., MOREIRA, L. M. e DA SILVA, A. C. "Bacterial phytopathogens and genome science". *Curr Opin Microbiol*, v.8, n.5, pp.595-600, Oct. 2005.

[69] PERNA, N. T., PLUNKETT, G., 3RD, BURLAND, V., *et al.* "Genome sequence of enterohaemorrhagic Escherichia coli O157:H7". *Nature*, v.409, n.6819, pp.529-33, Jan 25. 2001.

[70] CALCUTT, M. J., LEWIS, M. S. e WISE, K. S. "Molecular genetic analysis of ICEF, an integrative conjugal element that is present as a repetitive sequence in the chromosome of Mycoplasma fermentans PG18". *J Bacteriol*, v.184, n.24, pp.6929-41, Dec. 2002.

[71] VAN SLUYS, M. A., DE OLIVEIRA, M. C., MONTEIRO-VITORELLO, C. B., *et al.* "Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of Xylella fastidiosa". *J Bacteriol*, v.185, n.3, pp.1018-26, Feb. 2003.

[72] MOREIRA, L. M., DE SOUZA, R. F., ALMEIDA, N. F., JR., *et al.* "Comparative genomics analyses of citrus-associated bacteria". *Annu Rev Phytopathol*, v.42, pp.163-84. 2004.

[73] BHATTACHARYYA, A., STILWAGEN, S., IVANOVA, N., *et al.* "Whole-genome comparative analysis of three phytopathogenic Xylella fastidiosa strains". *Proc Natl Acad Sci U S A*, v.99, n.19, pp.12403-8, Sep 17. 2002.

[74] BHATTACHARYYA, A., STILWAGEN, S., REZNIK, G., *et al.* "Draft sequencing and comparative genomics of Xylella fastidiosa strains reveal novel biological insights". *Genome Res*, v.12, n.10, pp.1556-63, Oct. 2002.

[75] QIAN, W., JIA, Y., REN, S. X., *et al.* "Comparative and functional genomic analyses of the pathogenicity of phytopathogen Xanthomonas campestris pv. campestris". *Genome Res*, v.15, n.6, pp.757-67, Jun. 2005.

[76] PARKHILL, J., SEBAIHIA, M., PRESTON, A., *et al.* "Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica". *Nat Genet*, v.35, n.1, pp.32-40, Sep. 2003.

[77] KLEE, S. R., NASSIF, X., KUSECEK, B., *et al.* "Molecular and biological analysis of eight genetic islands that distinguish Neisseria meningitidis from the closely related pathogen Neisseria gonorrhoeae". *Infect Immun*, v.68, n.4, pp.2082-95, Apr. 2000.

[78] MUKHOPADHYAY, A. K., CHAKRABORTY, S., TAKEDA, Y., *et al.* "Characterization of VPI pathogenicity island and CTXphi prophage in environmental strains of Vibrio cholerae". *J Bacteriol*, v.183, n.16, pp.4737-46, Aug. 2001.

[79] NAKAYAMA, K., KANAYA, S., OHNISHI, M., *et al.* "The complete nucleotide sequence of phi CTX, a cytotoxin-converting phage of Pseudomonas aeruginosa: implications for phage evolution and horizontal gene transfer via bacteriophages". *Mol Microbiol*, v.31, n.2, pp.399-419, Jan. 1999.

[80] OELSCHLAEGER, T. A. e HACKER, J. "Impact of pathogenicity islands in bacterial diagnostics". *Apmis*, v.112, n.11-12, pp.930-6, Nov-Dec. 2004.

[81] HACKER, J., BENDER, L., OTT, M., *et al.* "Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal Escherichia coli isolates". *Microb Pathog*, v.8, n.3, pp.213-25, Mar. 1990.

[82] HACKER, J., SCHROTER, G., SCHRETTENBRUNNER, A., *et al.* "Hemolytic Escherichia coli strains in the human fecal flora as potential urinary pathogens". *Zentralbl Bakteriol Mikrobiol Hyg*, v.254, n.3, pp.370-8, May. 1983.

[83] GROISMAN, E. A. e OCHMAN, H. "Pathogenicity islands: bacterial evolution in quantum leaps". *Cell*, v.87, n.5, pp.791-4, Nov 29. 1996.

[84] HACKER, J. e KAPER, J. B. "Pathogenicity islands and the evolution of microbes". *Annu Rev Microbiol*, v.54, pp.641-79. 2000.

[85] NOEL, L., THIEME, F., NENNSTIEL, D., *et al.* "Two novel type III-secreted proteins of Xanthomonas campestris pv. vesicatoria are encoded within the hrp pathogenicity island". *J Bacteriol*, v.184, n.5, pp.1340-8, Mar. 2002.

[86] JACKSON, R. W., ATHANASSOPOULOS, E., TSIAMIS, G., *et al.* "Identification of a

pathogenicity island, which contains genes for virulence and avirulence, on a large native plasmid in the bean pathogen Pseudomonas syringae pathovar phaseolicola". *Proc Natl Acad Sci U S A*, v.96, n.19, pp.10875-80, Sep 14. 1999.

[87] TURNER, S. A., LUCK, S. N., SAKELLARIS, H.*, et al.* "Nested deletions of the SRL pathogenicity island of Shigella flexneri 2a". *J Bacteriol*, v.183, n.19, pp.5535-43, Oct. 2001.

[88] SULLIVAN, J. T., TRZEBIATOWSKI, J. R., CRUICKSHANK, R. W.*, et al.* "Comparative sequence analysis of the symbiosis island of Mesorhizobium loti strain R7A". *J Bacteriol*, v.184, n.11, pp.3086-95, Jun. 2002.

[89] MAHILLON, J. e CHANDLER, M. "Insertion sequences". *Microbiol Mol Biol Rev*, v.62, n.3, pp.725-74, Sep. 1998.

[90] SIGUIER, P., FILEE, J. e CHANDLER, M. "Insertion sequences in prokaryotic genomes". *Curr Opin Microbiol*, v.9, n.5, pp.526-31, Oct. 2006.

[91] RUTHERFORD, K., PARKHILL, J., CROOK, J.*, et al.* "Artemis: sequence visualization and annotation". *Bioinformatics*, v.16, n.10, pp.944-5, Oct. 2000.

[92] VAN DOMSELAAR, G. H., STOTHARD, P., SHRIVASTAVA, S.*, et al.* "BASys: a web server for automated bacterial genome annotation". *Nucleic Acids Res*, v.33, n.Web Server issue, pp.W455-9, Jul 1. 2005.

[93] MEYER, F., GOESMANN, A., MCHARDY, A. C.*, et al.* "GenDB--an open source genome annotation system for prokaryote genomes". *Nucleic Acids Res*, v.31, n.8, pp.2187-95, Apr 15. 2003.

[94] GAASTERLAND, T. e SENSEN, C. W. "MAGPIE: automated genome interpretation". *Trends Genet*, v.12, n.2, pp.76-8, Feb. 1996.

[95] KURTZ, S., PHILLIPPY, A., DELCHER, A. L.*, et al.* "Versatile and open software for comparing large genomes". *Genome Biol*, v.5, n.2, pp.R12. 2004.

[96] DELCHER, A. L., PHILLIPPY, A., CARLTON, J.*, et al.* "Fast algorithms for large-scale genome alignment and comparison". *Nucleic Acids Res*, v.30, n.11, pp.2478-83, Jun 1. 2002.

[97] DELCHER, A. L., KASIF, S., FLEISCHMANN, R. D.*, et al.* "Alignment of whole genomes". *Nucleic Acids Res*, v.27, n.11, pp.2369-76, Jun 1. 1999.

[98] BRAY, N., DUBCHAK, I. e PACHTER, L. "AVID: A global alignment program". *Genome Res*, v.13, n.1, pp.97-102, Jan. 2003.

[99] HUSON, D. H., REINERT, K. e MYERS, E. W. "The greedy path-merging algorithm for contig scaffolding". *J. ACM*, v.49, pp.603–615. 2002.

[100] KURTZ, S., CHOUDHURI, J. V., OHLEBUSCH, E.*, et al.* "REPuter: the manifold applications of repeat analysis on a genomic scale". *Nucleic Acids Res*, v.29, n.22, pp.4633-42, Nov 15. 2001.

[101] KURTZ, S. e SCHLEIERMACHER, C. "REPuter: fast computation of maximal repeats in complete genomes". *Bioinformatics*, v.15, n.5, pp.426-7, May. 1999.

[102] DEB, S. e NARAYANAN, P. J. *RepVis: A Remote Visualization System for Large Environments*. Workshop on Computer Vision, Graphics and Image Processing (WCVGIP). Gwalior, p.54-57, 2004.

[103] BURGE, C. e KARLIN, S. "Prediction of complete gene structures in human genomic DNA". *J Mol Biol*, v.268, n.1, pp.78-94, Apr 25. 1997.

[104] DAVIS, M. J., GILLASPSIE JR, A. G., VIDAVER, A. K.*, et al.* "Clavibacter: a new genus containing some phytopathogenic coryneform bacteria, including

Clavibacter xyli subsp. xyli sp. nov., subsp. nov. and Clavibacter xyli subsp. cynodontis subsp. nov., pathogens that cause ratoon stunting disease of sugarcane and Bermudagrass stunting disease". *Int. J. Syst. Bacteriol.*, v.34, pp.107-117. 1984.

[105] LIAO, C. H. e CHEN, T. A. "Isolation, culture and pathogenicity to Sudan Grass of a corynebacterium associated with ratoon stunting of sugarcane and with Bermuda grass". *Phytopathology*, v.71, pp.1303–1306. 1981.

[106] MILLS, L., LEAMAN, T. M., TAGHAVI, S. M.*, et al.* "Leifsonia xyli-like bacteria are endophytes of grasses in eastern Australia". *Aust. Plant Pathol.*, v.30, pp.145-151. 2001.

[107] SUZUKI, K. I., SUZUKI, M., SASAKI, J.*, et al.* "Leifsonia gen. nov., a genus for 2,4-diaminobutyric acid-containing actinomycetes to accommodate "Corynebacterium aquaticum" Leifson 1962 and Clavibacter xyli subsp. cynodontis Davis et al. 1984". *J Gen Appl Microbiol*, v.45, n.5, pp.253-262, Oct. 1999.

[108] EVTUSHENKO, L. I., DOROFEEVA, L. V., SUBBOTIN, S. A.*, et al.* "Leifsonia poae gen. nov., sp. nov., isolated from nematode galls on Poa annua, and reclassification of 'Corynebacterium aquaticum' Leifson 1962 as Leifsonia aquatica (ex Leifson 1962) gen. nov., nom. rev., comb. nov. and Clavibacter xyli Davis et al. 1984 with two subspecies as Leifsonia xyli (Davis et al. 1984) gen. nov., comb. nov". *Int J Syst Evol Microbiol*, v.50 Pt 1, pp.371-80, Jan. 2000.

[109] BOUCHER, Y., NESBO, C. L. e DOOLITTLE, W. F. "Microbial genomes: dealing with diversity". *Curr Opin Microbiol*, v.4, n.3, pp.285-9, Jun. 2001.

[110] *Terceiro Levantamento - Cana-de-açúcar, Safra 2006/2007*. CONAB – Companhia Nacional de Abastecimento, 2006.

[111] DEAN, J. L. e DAVIS, M. J. "Yield losses caused by ratoon stunting. disease of sugarcane in Florida". *Journal of the American Society of Sugarcane Technologists*, v.10, pp.66-72. 1989.

[112] FEGAN, M., CROFT, B. J., TEAKLE, D. S.*, et al.* "Sensitive and specific detection of Clavibacter xyli subsp. xyli, causal agent of ratoon stunting disease of sugarcane, with a polymerase chain reaction-based assay". *Plant pathology*, v.47, n.47, pp.495-504. 1998.

[113] GIGLIOTI, E. "RSD impact on sugar industries – Brazil". *International Congress of Plant Pathology*, v.7, pp.9-16. 1998.

[114] LI, T. Y., YIN, P., ZHOU, Y.*, et al.* "Characterization of the replicon of a 51-kb native plasmid from the gram-positive bacterium Leifsonia xyli subsp. cynodontis". *FEMS Microbiol Lett*, v.236, n.1, pp.33-9, Jul 1. 2004.

[115] DIATCHENKO, L., LAU, Y. F., CAMPBELL, A. P.*, et al.* "Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries". *Proc Natl Acad Sci U S A*, v.93, n.12, pp.6025-30, Jun 11. 1996.

[116] STAJICH, J. E., BLOCK, D., BOULEZ, K.*, et al.* "The Bioperl toolkit: Perl modules for the life sciences". *Genome Res*, v.12, n.10, pp.1611-8, Oct. 2002.

[117] SETUBAL, J. C. e WERNECK, R. *A program for building contig scaffolds in double-barrelled shotgun genome sequencing*. Institute of Computing. Unicamp, IC-01-05, 2001.

[118] GOTOH, O. "An improved algorithm for matching biological sequences". *J Mol Biol*, v.162, n.3, pp.705-8, Dec 15. 1982.

[119] TATUSOV, R. L., FEDOROVA, N. D., JACKSON, J. D.*, et al.* "The COG database: an updated version includes eukaryotes". *BMC Bioinformatics*, v.4, pp.41, Sep 11. 2003.

[120] TATUSOV, R. L., KOONIN, E. V. e LIPMAN, D. J. "A genomic perspective on protein families". *Science*, v.278, n.5338, pp.631-7, Oct 24. 1997.

[121] MULDER, N. J., APWEILER, R., ATTWOOD, T. K.*, et al.* "InterPro, progress and status in 2005". *Nucleic Acids Res*, v.33, n.Database issue, pp.D201-5, Jan 1. 2005.

[122] LI, M., MA, B., KISMAN, D.*, et al.* "PatternHunter II: highly sensitive and fast homology search". *Genome Inform*, v.14, pp.164-75. 2003.

[123] LI, M., MA, B., KISMAN, D.*, et al.* "Patternhunter II: highly sensitive and fast homology search". *J Bioinform Comput Biol*, v.2, n.3, pp.417-39, Sep. 2004.

[124] SONNHAMMER, E. L. e DURBIN, R. "A workbench for large-scale sequence homology analysis". *Comput Appl Biosci*, v.10, n.3, pp.301-7, Jun. 1994.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo