UNIVERSIDADE FEDERAL DO PARANÁ

Jesse Teixeira da Silva

GENETIC TRANSCRIPT ANALYZER - FERRAMENTA COMPUTACIONAL PARA ANÁLISE DE TRANSCRIÇÃO GÊNICA POR RNA-SEQ

CURITIBA
2012

JESSE TEIXEIRA DA SILVA

GENETIC TRANSCRIPT ANALYZER - FERRAMENTA COMPUTACIONAL PARA ANÁLISE DE TRANSCRIÇÃO GÊNICA POR RNA-SEQ

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, área de concentração em Bioinformática.

Orientador: Prof. Dr. Luiz Antônio Pereirá Neves

Co-orientadora: Prof. Drª Rose Adele Monteiro

CURITIBA
2012

Silva, Jesse Teixeira da
S586 Genetic transcript analyzer – ferramenta computacional para análise de transcrição gênica por RNA-SEQ / Jesse Teixeira da Silva. - Curitiba, 2012.
64 f.: il., tabs, grafs.

Orientador: Prof . Dr. Luiz Antônio Pereira Neves
Co-orientador: Profa. Dra. Rose Adele Monteiro
Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Curso de Pós-Graduação em Bioinformática.
Inclui Bibliografia.

1. Java (Linguagem de programação de computador). 2. Ferramenta computacional. 3. RNA-SEQ (Sequenciamento). 4. Transcrição gênica. I. Neves, Luiz Antônio Pereira. II. Monteiro, Rose Adele. III.Título. IV. Universidade Federal do Paraná.

CDD 574.192

TERMO DE APROVAÇÃO

JESSÉ TEIXEIRA DA SILVA

Genetic Transcript Analyzer – Ferramenta Computacional para Análise de Transcrição Gênica por RNA-seq

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador: Prof. Dr. Luiz Antonio Pereira Neves

Coorientador: Profª. Drª. Rose Adele Monteiro

Prof. Dr. Gustavo Benvenutti Borba
Universidade Tecnológica Federal do Paraná - UTFPR

Prof. Dr. Leonardo Magalhães Cruz
Universidade Federal do Paraná - UFPR

Curitiba, 27 de fevereiro de 2012

AGRADECIMENTOS

- À Universidade Federal do Paraná.
- Ao Prof. Dr. Luiz Antônio Pereira Neves, pela sua orientação e por sua e fé de que a ciência é o futuro do Brasil e que todos têm algo a contribuir.
- À Prof.ª. Drª Rose Adele Monteiro, pelo auxilio e por me fazer acreditar que era capaz de realizar as tarefas.
- À mestra Michelle Zibetti Tadra Sfeir, que sempre teve a paciência e boa vontade para me auxiliar nas questões biológicas e me acompanhar durante todo o projeto.
- Ao Mestre Vinícius A. Weiss, por todo auxilio recebido durante todas as etapas do processo e pela troca de conhecimento.
- À Prof.ª Dra. Jeroniza Marchaukoski, por acreditar que eu venceria este desafio.
- Ao mestre Dieval Guizelini, que me deu dicas importantíssimas nos momentos necessários e também forneceu uma ferramenta excepcional para me ajudar a ter um sistema mais confiável.
- Aos companheiros de curso, que sempre me deram apoio moral e intelectual nos momentos mais difíceis e também nas conquistas.
- Ao pessoal da Iddéia, empresa que, acreditando em meu potencial, abriu mão de horas de trabalhos valiosas a fim de me ver correr atrás de meu sonho, sem ter exigido nada em troca.
- Aos colegas do setor de biologia em geral, que sempre me receberam de braços abertos.
- Aos meus grandes amigos, Greg Santos, Eduardo Santos, Janealysson Araújo, Jhonisson de Paula e Ronaldo dos Santos, com os quais sempre pude contar nos momentos de dificuldades.
- Ao sempre entusiasta Rafael Stanger (O Boss), por quem tenho não apenas respeito, mas admiração.
- Ao meu pai, que me ensinou que caráter não se compra ou vende, é construído.
- À minha mãe, que mesmo perante todas as dificuldades, sempre lutou ao meu lado e acreditou em minhas vitórias.
- Às minhas filhas, que apesar de estarem longe, são o motivo pelo qual me esforço todos os dias.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

LISTA DE SIGLAS

BAM	-	SAM Binário
GBK	-	Formato padrão de arquivos biológicos do Genbank
cDNA	-	Ácido desoxirribonucleico complementar
DNA	-	Ácido desoxirribonucleico
GPL	-	Licença pública Geral
GPU	-	Unidade de processamento gráfico
GTA	-	Genetic Transcript Analyzer
HD	-	Disco rígido
HQL	-	Hibernate query language
mRNA	-	Ácido ribonucleico mensageiro
MS	-	Microsoft
PB	-	Pares de bases
PCR	-	Reação de Polimerização em Cadeia
RAM	-	Memória de acesso aleatório
RNA	-	Ácido ribonucleico
RPKM	-	Reads por Kilobase de regiões expressas a cada milhão de Reads mapeados
SAGE	-	Análise Seriada da Expressão Gênica
SAM	-	Sequence Alignment/Map
SQL	-	Linguagem de consulta estruturada
TI	-	Tecnologia da Informação
XML	-	Linguagem de marcação estendida

LISTA DE SÍMBOLOS

GB	-	GigaByte
GHz	-	GigaHertz
MB	-	MegaByte
MHz	-	MegaHertz
ms	-	Millisegundo
TB	-	TeraByte
®	-	Marca Registrada

RESUMO

Atualmente, com as inúmeras tecnológicas disponíveis para sequenciamento de transcritos, o problema consiste no número de ferramentas computacionais para a análise de dados. Neste trabalho apresentamos o desenvolvimento do *Genetic Transcript Analyzer* (GTA), uma nova ferramenta para comparação de resultados da análise de expressões gênicas por RNA-Seq para diferentes amostras, gerando informações mais ricas a partir de dados brutos no formato BAM ou SAM. O sistema proposto é *freeware* e foi desenvolvido utilizando a linguagem de programação Java® com auxílio da ferramenta SamTools®, uma biblioteca Java para leitura de arquivos no formato SAM e BAM (Heng li et al, 2009). As validações do sistema deram-se através de entrevistas de satisfação dos usuários e comparações diretas com programas de planilhas eletrônicas e também com comparações descritivas com outros softwares disponíveis que possuam módulos semelhantes ao aqui proposto. O resultado desta pesquisa foi o desenvolvimento de uma solução simples e prática para analisar os resultados de sequenciamentos realizados por máquinas de sequenciamento da nova geração através do método Rna-Seq.

Palavras-chave: Análise, ferramenta computacional, java, RNA-Seq, samtools, tecnologia da informação, transcriptoma.

ABSTRACT

Nowadays, with a large number of technological tools available, the bottleneck has moved from collection to data analysis. This paper presents the development of the GTA (Genetic Transcript Analyzer), a new tool for comparing biological samples and analysis of statistical data, comparing their gene expression and generating richer information from the raw data type coming from sorted BAM files. The proposed system is freeware and was designed and built using the Java® language with the tool Samtools®, a Java library for reading files in SAM and BAM files(Heng li et al, 2009) . The system validations are made through user satisfactions interviews and direct comparisons with the program actually used by the professional, the Microsoft Excel and also descriptive comparison with others softwares that can make genetic data comparisons . As result we had developed a new computer software able to analyze the results of the new high output machines that works with Rna-Seq.

Keywords: Analyze, information technology, Java, RNA-Seq, samtools, software, Transcriptome.

1. INTRODUÇÃO

É cada vez mais necessária a interação entre profissionais das áreas de pesquisas biológicas e tecnológicas, uma vez que a demanda por soluções tecnológicas para análise de dados biológicos tem crescido fortemente nos últimos anos. Dentre estas soluções estão o aperfeiçoamento de novas ferramentas que nos auxiliam a conhecer organismos que até então eram pouco explorados e, em alguns casos até desconhecidos, mas que são de extrema importância para o entendimento do ecossistema e consequentemente na melhoria de processos que envolvem a manipulação de dados biológicos.

A técnica de RNA-Seq é o sequenciamento em larga escala de cDNA (DNA complementar) utilizando sequenciadores de nova geração, nos ajudando a representar os resultados dos transcriptomas analisados, gerando informações que são analisadas por softwares específicos, tornando estas informações mais claras aos pesquisadores que as utilizam em novas pesquisas e comparações de organismos.

Os profissionais da área da bioquímica executam o processo de comparação de *Reads* por amostras utilizando softwares complexos de terceiros, como o *CLC DNA WorkBench®,* uma ferramenta comercial, o *BioConductor®,* que é um sistema *OpenSource* e também ferramentas genéricas de planilhas eletrônicas, as quais exigem um período muito grande entre a preparação dos dados a serem analisados (cadastro, separação e cálculos iniciais) e o resultado final desejado, que é a comparação precisa de cada amostra e seus valores de Reads.

Este trabalho surgiu da necessidade de se desenvolver uma ferramenta simples que permita aos profissionais da biologia analisar os dados de forma rápida e eficaz diretamente de sua estação de trabalho, não necessitando para isso recorrer a ferramentas complexas, de alto custo ou que demandem muito esforço para preparar os dados a serem interpretados. Todo o projeto foi desenvolvido utilizando ferramentas de domínio público e fácil acesso e aprendizado, visando oferecer um ambiente simplificado e funcional tanto para usuários como para desenvolvedores. O programa foi desenvolvido na plataforma *Java® standard edition*. O software trabalha diretamente

com arquivos disponibilizados no formato de alinhamentos de sequências (SAM) e sua versão binária (BAM).

2. OBJETIVOS

2.1. Objetivo Geral

O objetivo geral do projeto é desenvolver uma ferramenta computacional para a área biológica, aqui chamada de *Genetic Transcript Analyser (GTA)*, tendo como meta facilitar a aquisição e a manipulação dos dados biológicos obtidos através do método Rna-Seq, dispensando o profissional de tarefas repetitivas e aumentando sua produtividade.

2.2. Objetivos específicos

- Desenvolver um sistema capaz de manipular e armazenar dados biológicos de análise de expressões gênicas.
- Agregar a esta ferramenta novas modalidades de compartilhamento e armazenamento dos dados analisados.
- Criar relatórios e gráficos a partir da análise dos dados inseridos.
- Avaliar o sistema proposto a fim de garantir sua viabilidade de acordo com o objetivo geral desta pesquisa.

2.3. Justificativa

A principal motivação para este projeto foi a visível necessidade que os profissionais do setor de bioquímica possuem de ferramentas amigáveis e de baixo custo para auxiliar na análise de resultados da análise de expressão gênica efetuadas pelo método RNA-Seq e também em suas eventuais manipulações. O foco do projeto é valorizar ainda mais as horas de trabalho do profissional e evitar que ele seja

deslocado do processo de pesquisa e análise para o processo de preparação dos dados para novas análises.

3. REVISÃO BIBLIOGRÁFICA

3.1. Conceitos de biologia e genética

3.1.1. DNA

O ácido desoxirribonucleico é um composto orgânico que contém instruções genéticas que coordenam o desenvolvimento e funcionamento de todos os seres vivos e alguns vírus. O seu principal papel é armazenar as informações necessárias para a construção das proteínas e do ácido desoxirribonucleico (RNA). O DNA é um longo polímero de unidades simples (monômeros) de nucleotídeos, cuja cadeia principal é formada por moléculas de açúcares e fosfato intercalados unidos por ligações fosfodiéster. A sequência de bases ao longo da molécula de DNA constitui a nossa informação genética, sendo estas interpretadas através do código genético, que especifica a sequência linear dos aminoácidos das proteínas (E. Houdebine, 2000). A sequência de DNA de um conjunto de cromossomos de um organismo é conhecida como o *genoma* do organismo.

3.1.2. RNA

O ácido ribonucleico é o responsável pela síntese de proteínas de determinada célula, sendo conhecido como um composto químico de elevada massa molecular (polímero) de compostos com grande energia que auxiliam no processo metabólico do organismo (nucleotídeos), formado na sua grande maioria por cadeias simples. As dimensões das moléculas formadas por RNA possuem suas dimensões muito inferiores às moléculas formadas por DNA (J.M

Amabis et al, 2004). Dentre as características que diferenciam uma molécula de RNA de uma molécula de DNA, estão:

- O RNA é formado por uma cadeia simples de nucleotídeo, enquanto o DNA é formado por uma dupla hélice.
- O RNA possui açúcar ribose em seus nucleotídeos, sendo que o DNA possui a desoxirribose.
- No RNA, em comparação ao DNA, a única base a se alterar é a Timina (T), que é substituída pela pirimidina.

3.1.2.1. RNA ribossômico (rRNA)

É inicialmente armazenado nos nucléolos e passa para o citoplasma, associando-se a proteínas para formar os ribossomos, responsáveis pela formação de proteínas.

3.1.2.2. RNA mensageiro (mRNA)

É uma molécula de vida curta, de tamanho variável, que tem como função principal levar para o citoplasma as informações para a síntese de proteínas. O DNA sintetizado pelas moléculas de mRNA é conhecido como DNA Complementar (cDNA), onde os introns já foram removidos.

3.1.2.3. RNA transportador (tRNA)

É o RNA responsável por transportar os aminoácidos aos locais onde a síntese proteica está ocorrendo, sendo formado por apenas uma cadeia de nucleotídeos, que se dobra sobre si próprio em forma de trevo.

3.1.3. Gene

Uma definição moderna descreve que o gene é uma sequência de nucleotídeos de DNA que podem ser transcritas em RNA, sendo este um segmento de um cromossomo correspondente a um código genético distinto, uma informação para produção de proteínas ou controle de alguma característica, como exemplo a cor dos olhos de um individuo. O gene possui seções que codificam proteínas, conhecidas como exons e seções que não codificam nenhuma, conhecidas por introns. Os introns são encontrados principalmente (mas não exclusivamente) em células eucarióticas, sendo inicialmente transcritos na molécula de RNA Mensageiro, mas eliminado durante o processo que une os exons após a transcrição, conhecido como splicing. (S. Farrel, 2007).

3.1.4. Transcriptoma

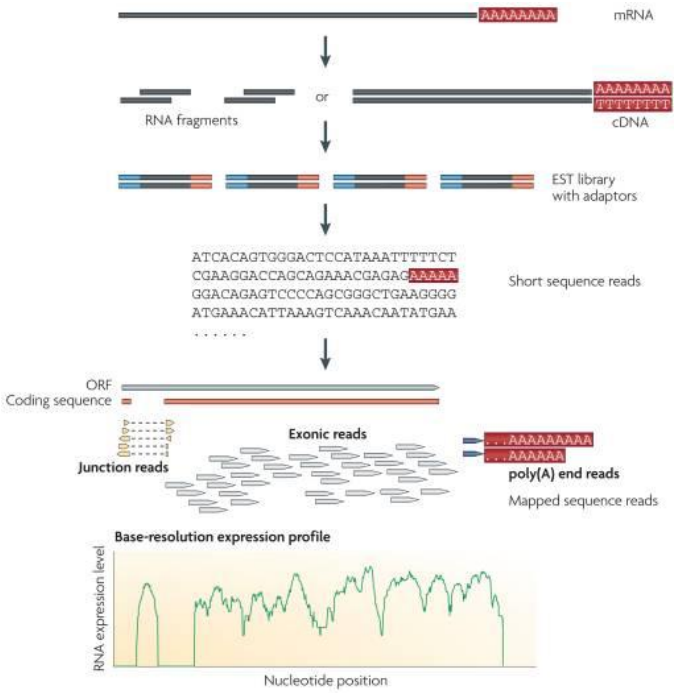
O transcriptoma é o conjunto completo de RNA de um organismo, um grupo de células ou mesmo de uma célula específica em uma determinada condição fisiológica. O RNA é sintetizado do DNA através de um processo conhecido como transcrição. O RNA presente em uma célula determina quais genes serão expressos e pode mudar durante a vida do organismo, diferentemente do DNA, que se mantem estático. Alterações ambientais são uma das principais causas de mudanças, tendo em vista que o organismo tenta se adaptar a estas mudanças para continuar vivo (J. Moore, 2011).

O transcriptoma pode demonstrar grandes variações entre as células dos organismos. Por exemplo, células de diferentes partes do corpo possuem a mesma cópia do DNA, mas a diferença entre cada uma delas é determinada apenas pelos seus respectivos transcriptomas (S.E. Smith, 2003).

3.1.5. RNA-Seq

Nas palavras de Pinto et al. (2011), o termo RNA-Seq representa o transcriptoma revelado pelo sequenciamento de DNA complementar (cDNA) e é considerado pelos pesquisadores um método revolucionário, pois possui alta sensibilidade e pode ser utilizado para caracterizar o transcriptoma de um organismo. É útil para descobrir novas transcrições, identificações de mutações, deleções e inserções, *splicings* alternativos e também oferece uma cobertura elevada. Uma das suas grandes vantagens é a ausência quase total de ruídos e a capacidade de detectar um número elevado de cópias de mRNA por célula. A Figura 1 ilustra de forma simplificada um processo típico de sequenciamento por RNA-Seq.

Figura 1 - Etapas de um experimento utilizando RNA-Seq. FONTE: pubMed (2011).



De uma forma simplificada, uma quantidade de RNA é convertida em uma biblioteca contendo fragmentos de cDNA. Em seguida estes fragmentos recebem

adaptadores (bases de DNA) e passam pelo sequenciamento, gerando uma sequência curta (na ordem de 30 a 400 pares de base). Em seguida estas leituras são alinhadas a um genoma de referência (ou outro transcriptoma) ou até mesmo remontadas sem um genoma de referência a fim de criar um mapa em escala genômica que é composto pela estrutura transcricional ou o nível de expressão de cada gene individualmente (Wang et al, 2008).

- ***Principais vantagens do RNA-Seq e seus desafios***

O RNA-Seq possui algumas vantagens sobre seus concorrentes que devem ser frisadas, sendo elas:

- Detecção de transcritos não restrita a aqueles pré-existentes em uma sequência genômica, como acontece com as abordagens baseadas em hibridizações. Isso torna o método mais atrativo para pesquisas com organismo cujos genomas ainda não foram determinados.
- Resoluções de até uma única base podem ser determinadas devido às precisas localizações dos limites de transcrição
- Detecção de variações da sequência genômica nas regiões transcritas
- Ruído de fundo muito menor se comparado aos microarranjos de DNA
- Requer uma quantidade muito menor de amostras de RNA.

Apesar de apresentar vantagens sobre as outras técnicas, o método RNA-Seq ainda possui muitos desafios em sua metodologia, como por exemplo, um problema comum na área da informática, que é o armazenamento, a leitura e o processamento das informações geradas, o que pode resultar em erros no momento da análise de imagens, além de que mesmo com leituras de boa qualidade, ainda é um desafio para a bioinformática a precisão durante as etapas de alinhamento ou montagem destas leituras. Outro desafio a ser considerado é o custo do sequenciamento, pois quanto maior a cobertura desejada, mais sequenciamento deve ser executado, o que ocorre comumente em grandes genomas a serem cobertos. Este método deveria ser capaz de

identificar qualquer tipo de RNA de qualquer tamanho, mas como o RNA-Seq possui varias etapas de manipulação durante a produção das bibliotecas de cDNA, ocorre que moléculas maiores de RNA por exemplo, acabam tendo que ser fragmentadas para se adequarem as tecnologias de nova geração para sequenciamento , sendo que cada um dos métodos usuais pode influenciar diferentemente o resultado produzido.

Mesmo com as limitações citadas, o RNA-Seq tem sido considerado um marco, tendo em vista que com sua alta resolução e sensibilidade, revelou varias regiões de transcrição e isoformas de *splicing* novas em genes já conhecidos, inclusive mapeando seus limites 5' (cinco linha) e 3' (três linha). Novas regiões de transcrição também já foram descobertas com RNA-Seq para cada genoma em que ele foi empregado. Essas novas regiões, combinadas com as muitas variantes de *splicing* ainda não descobertas, sugerem uma complexidade transcricional ainda maior do que a observada até aqui. Este método já provou ser capaz, também, de rastrear com precisão mudanças na expressão gênica durante diferentes estágios do desenvolvimento de alguns organismos, assim como em comparações de diferentes tecidos. A Tabela 1 mostra de forma simplificada as principais vantagens do RNA-Seq sobre as outras metodologias segundo Z Wang et al (2009).

Tabela 1- Principais vantagens do RNA-Seq sobre outras metodologias. FONTE: Z Wang et al (2009).

Tecnologia	**Tiling microarray**	**seqüenciamento de cDNA ou EST**	**RNA-seq**
Especificações da tecnologia			
Princípio	Hibridização	Sequenciamento Sanger	Alta capacidade de sequenciamento
Resolução	até 100 pb	Única base	Única base
Vazão	Alta	Baixa	Alta
Confiança em seqüência genômica	Sim	Não	Em alguns casos
Ruído de fundo	Alto	Baixo	Baixo
Aplicação			
Simultaneamente mapear regiões transcritas e expressão gênica	Sim	Limitado pela expressão do gene	Sim
Faixa dinâmica para	Até 100 vezes	-	> 8000 vezes

quantificar nível de expressão de genes	em média		
Capacidade de distinguir diferentes isoformas	Limitado	Sim	Sim
Capacidade de distinguir expressão alélica	Limitada	Sim	Sim
Questões práticas			
Quantidade necessária de RNA	Alto	Alto	Baixo
Custo para mapeamento de genomas grandes	Alto	Alto	Relativamente baixa

- ***Sistemas de sequenciamento***

Os três principais sistemas de sequenciamento na nova-geração são: GS FLX Genome Analyzer (Roche-454), Illumina Solexa 1G sequencer e SOLiD system (AppliedBiosystems) (van Vliet, 2009).

- ***Reads***

É uma sequência de dados não tratados provenientes das maquinas de sequenciamento. Um read pode consistir de múltiplos segmentos, conhecidos como sub Reads.

- ***RPKM***

A sigla RPKM (*Reads* Per Kilobase of exon model Per Million Mapped Reads), ou *Reads* por Kilobase de regiões expressas a cada milhão de *Reads* mapeados (Mortazavi et al, 2008) e é utilizado para normalizar os *Reads* de cada amostra, considerando o tamanho do gene, o tamanho da biblioteca (amostra) e a quantidade de

Reads expressos no gene em questão. O Cálculo do RPKM esta exemplificado na Equação 1, detalhando cada componente.

Equação 1

$$RPKM = \frac{10^9 * C}{N * L}$$

FONTE: Mortazavi et al (2008)

Na equação do RPKM, o valor 10^9 representa o resultado de 1 mil pares de bases multiplicado por 1 milhão de *Reads*, enquanto C representa o total de *Reads* do gene analisado, N o total de *Reads* da amostra analisada e L o tamanho da transcrição. O resultado desta equação mede a densidade do *Read* em uma região gênica de interesse, normalizando a contagem do *Read* em suas regiões exônicas correspondentes comparados ao tamanho original do gene ou exon.

3.1.6. Arquivos SAM e BAM

O formato SAM (Sequence Alignment/Map) é um formato genérico para armazenar grandes sequências de nucleotídeos alinhados. Suas informações são apresentadas em formato texto separado por tabulações, contendo um cabeçalho (opcional) e uma seção de alinhamento. O que diferencia um cabeçalho é que o inicio da linha dos cabeçalhos contem sempre um símbolo de arroba (@). Cada linha que descreve o alinhamento possui onze campos contendo informações essenciais como, por exemplo, posição do mapeamento.

A Figura 2 ilustra um alinhamento contendo um par de Reads (r001/1 e r001/2) um *Read* chimérico (r003) e um alinhamento em *Split* no formato SAM, contendo um cabeçalho e uma seção que descreve o alinhamento.

Figura 2 - Exemplo de um arquivo SAM contendo um cabeçalho e informações sobre determinado alinhamento. FONTE: O autor, com base em arquivos SAM (2012).

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

No exemplo da Figura 2 temos que:

- @HD simboliza a linha do cabeçalho
- VN simboliza a versão do arquivo
- SO simboliza a ordem do alinhamento, com os valores sendo desconhecido (*unknown*), não ordenado (*unsorted*), nome da query (*queryname*) e coordenada (*coordinate*). O padrão é o valor desconhecido.
- SN simboliza o nome da sequência de referência. Cada linha contendo @SQ deve conter uma única marcação SN.
- LS simboliza o tamanho da sequência de referência.

Outros campos são encontrados em cada sequência e suas determinações podem ser verificadas na documentação oficial do formato. Dentre as características do formato SAM destacam-se os seguintes pontos (Homer N. et al, 2009).

- É flexível para armazenar toda a informação do alinhamento.
- Pode ser facilmente gerado por programas de alinhamento ou convertido de formatos de alinhamentos já existentes.
- É compacto, ou seja, ocupa pouco espaço no disco.
- Permite que a maioria das operações no alinhamento seja executada sem a necessidade de carregar todo o alinhamento na memória.

- Permite que o arquivo seja indexado, fazendo assim que seja pratico recuperar todos os *Reads* alinhados.

O formato Bam nada mais é do que um arquivo SAM em formato binário, o que reduz drasticamente o seu tamanho. O formato BAM utiliza o esquema de compressão em blocos, garantindo um alinhamento mais compacto e podendo ser indexado e ordenado, fazendo com que o acesso aos dados seja extremamente rápido. Os índices de um arquivo BAM encontram-se em outro arquivo que possui sua extensão como .**bai** . Por exemplo, um arquivo com o nome de *amostra.bam* terá seus índices contidos em um outro arquivo com o nome de *amostra.bam.bai* (samtools, 2009).

3.1.7. Genbank Flat File Format (GBK)

De acordo com a *National Center for Biotechnology Information* (NCBI), o Genbank é um banco de dados contendo uma coleção com sequências de DNA. Até abril de 2011, já havia em seus registros aproximadamente 191.401.393.188 (mais de 191 bilhões) bases em 62.715.288 sequências (NCBI, 2012). Os arquivos provenientes do genbank podem ser disponibilizados em arquivos no formato GBK, contendo informações completas ou parciais sobre determinado organismo, dna, gene, etc. Um arquivo no formato GBK deve ser constituído de três partes, sendo elas: *Header*, *Features* e *Sequence*. A Tabela 2 detalha as informações apresentadas em cada uma destas seções.

Tabela 2 - Principais componentes contidos em um arquivo no formato GBK. FONTE: NCBI (2011).

	ITEM	DESCRIÇÃO
HEADER (CABEÇALHO)	LOCUS	Um nome mnemônico curto para a citação.
	DEFINITION	Uma descrição da sequência.
	ACCESSION	Usado para citações da sequência em jornais e revistas.

	VERSION	O número de acession primário seguido com um número de versão dos dados sequenciados.
	KEYWORDS	Frases curtas descrevendo informações sobre a citação.
	SOURCE	Nome comum do organismo ou o nome mais utilizado na literatura.
	ORGANISM	Nome científico formal do organismo e níveis taxonômicos.
	REFERENCE	Citações para todos os artigos contendo os dados reportados.
	AUTHORS	Lista com os autores da citação.
	TITLE	Titulo completo da citação.
	JOURNAL	Listas com os nomes dos jornais, ano, volume e número de paginas da citação.
	MEDLINE	Identificador único da citação perante na Medline.
	PUBMED	Identificador único da citação perante na Pubmed.
	REMARK	Relevância da citação.
	COMMENT	Referências cruzadas com outras citações, notas de mudanças, etc.
FEATURES (CARACTERISTICAS)	SOURCE	Contém informações sobre o organismo, mapeamento, cromossomos, alinhamento de tecidos, identificação de clones, etc.
	CDS	Instruções de como unir sequências para construir sequências de aminoácidos através das coordenadas passadas. Inclui referencias cruzadas com outras bases de dados.
	GENE FEATURE	Um segmento de DNA identificado por um nome.
	RNA FEATURE	Utilizado para anotar o RNA na sequência genômica, como exemplo mRNA,tRNA e rRNA.
SEQUENCE (SEQUÊNCIA)	SEQUENCE	A sequência completa

Mais informações sobre o formato GBK e suas variações estão disponíveis nas documentações oficiais da *National Center for Biotechnology Information* (NCBI),

podendo ser acessadas pelo endereço http://www.ncbi.nlm.nih.gov.

3.2. Ferramentas Computacionais

3.2.1. Apache Derby Data-Base

É uma base de dados relacional escrita na linguagem JAVA e disponível sob a versão 2.0 da licença Apache, sendo utilizado para processamento de transações Online (Apache Foundation, 2011).

3.2.2. Bioconductor

O Bioconductor fornece ferramentas para análise de dados genômicos utilizando uma linguagem de programação estatística, sendo uma ferramenta de código aberto e apta a novos desenvolvimentos pela comunidade, que disponibilizam dois releases a cada ano. O Bioconductor consegue exportar diversos tipos de arquivos de sequências, incluindo o formato fasta, fastq, BAM, gff, bed e wig, entre muitos outros e suporta operações de manipulações comuns e avançadas, tais como trimming, transformações e alinhamentos. Consegue trabalhar com Chip-seq, RNA-Seq e outros, tornando a ferramenta de grande abrangência entre as áreas da pesquisa e manipulação de dados genômicos (Bioconductor, 2012).

Devido ao fato de ser um projeto desenvolvido e mantido por uma comunidade, torna-se mais fácil procurar e fornecer ajuda a usuários e também desenvolvedores por todo o mundo, não restringido as possíveis soluções e melhorias a apenas poucos laboratórios, como é comum quando se utiliza uma ferramenta comercial desta natureza.

3.2.3. Bioperl

É uma coleção de módulos na linguagem Perl que facilita o desenvolvimento de scripts para aplicações na área da bioinformática (Bioperl, 2002). A primeira versão estável foi disponibilizada em 11 de junho de 2002 e sua versão mais recente, encontrada durante a criação deste documento é a 1.6.9, de abril de 2011. Sua licença é distribuída sob a "*Perl Artistic License*" e sua distribuição pode ser encontrada em bioperl.org.

3.2.4. Clcbio Dna Workbench

Segundo clcbio (2012), o CLCbio Dna WorkBench cria um ambiente onde os usuários possam executar diversas análises de sequência com dados de fácil manipulação e excelentes ferramentas gráficas com diversas opções de entrada e saída de dados. A ferramenta é disponível para os sistemas operacionais Windows®, Mac OS® e plataformas Linux®. As características que são reforçadas pelos desenvolvedores são:

- Fácil acesso a um grande número de ferramentas de pesquisa.
- Transformar seu próprio computador em um centro de alto desempenho.
- Facilitar o trabalho de pesquisa.
- É uma ferramenta que esta sempre evoluindo.
- Ambiente de trabalho personalizável.

O CLC DNA pode ser adquirido com uma licença estudante ou uma licença industrial, tendo cada uma delas preços e acessórios diferenciados segundo informações do fabricante.

3.2.5. Hibernate

O Hibernate é um framework para o mapeamento objeto-relacional escrito na linguagem Java que facilita o mapeamento dos atributos entre uma base tradicional de dados relacionais e o modelo objeto de uma aplicação, mediante o uso de arquivos (XML) para estabelecer esta relação. A HQL (Hibernate Query Language) é um dialeto SQL para o Hibernate. Ela é uma poderosa linguagem de consulta que se parece muito com a SQL, mas totalmente orientada a objetos, incluindo os paradigmas de herança, polimorfismo e encapsulamento (Gavin King, 2004).

3.2.6. Java

O Java é uma linguagem de programação e uma plataforma de computação lançada pela Sun Microsystems® em 1995 e hoje mantida pela Oracle®. É a tecnologia que capacita muitos programas da mais alta qualidade, como utilitários, jogos e aplicativos corporativos, entre muitos outros (Deitel et al, 2001). O Java é executado em mais de 850 milhões de computadores pessoais e em bilhões de dispositivos em todo o mundo, inclusive telefones celulares e dispositivos de televisão (Oracle, 2011).

3.2.7. JGBParser

O JGBParser é um analisador de dados no formato Genbank desenvolvido por Dieval Guizelini em 2010 e em constante processo de melhorias. Segundo Dieval Guizelini (2010), o produto final apresentou nos testes velocidades 15, 9 e 2 vezes mais rápidas que a disponível pelo BioPython, BioPerl e GBParsy, respectivamente. Estima-se que o analisador extraia todas as informações contidas em um arquivo GenBank com velocidades de aproximadamente 16.6Mb/s e de 30Mb/s para extrair apenas as informações da seção features. Com estas informações extraídas de forma padronizada, as aplicações de bioinformática processarão muito mais informações em um tempo muito menor.

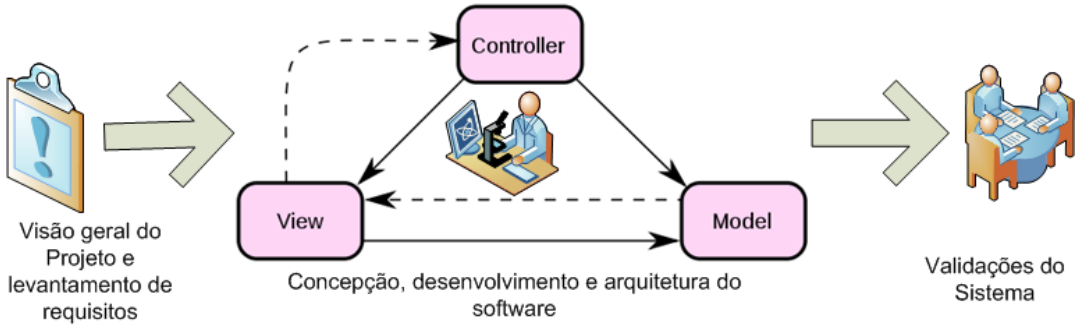
3.2.8. Samtools

O SamTools é uma biblioteca para manipular alinhamentos no formato BAM e SAM. É capaz de converter de outros formatos, ordenar e mesclar alinhamentos, remover duplicatas PCR, gerar informações por posições, etc. O Samtools possui duas implementações distintas, sendo uma delas em linguagem C e a outra em linguagem Java (Heng Li et al, 2009).

4. RESULTADOS E DISCUSSÃO

Este trabalho foi planejado e desenvolvido em três etapas, conforme ilustrado na Figura 3. Estas etapas estão descritas em ordem de ocorrência no decorrer deste capítulo, assim como as suas eventuais considerações e resultados de cada item.

Figura 3 - Exibição das 3 principais etapas do sistema GTA. Fonte: O Autor (2012).



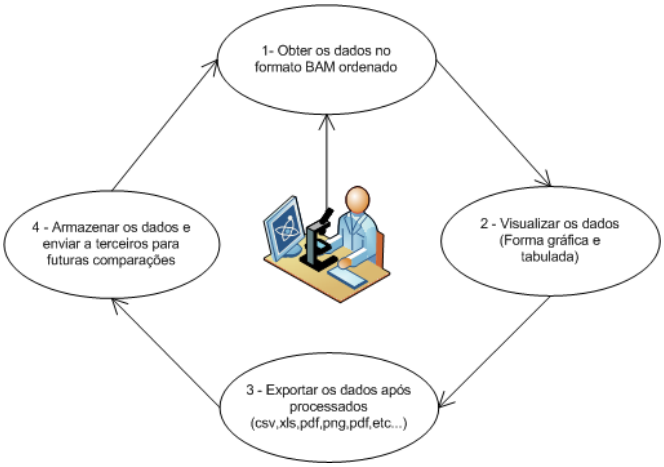
4.1. Visão Geral Do Projeto e levantamento de requisitos

A criação do projeto foi baseada na necessidade que os profissionais da área biológica possuem de uma ferramenta de fácil utilização capaz de apresentar, de forma eficaz, as principais diferenças de expressões gênicas entre duas ou mais amostras manipuladas geneticamente e exibindo estas informações tanto na forma gráfica como tabulada, filtrando por valores como totais de *Reads*, nomes dos genes, nomes das funções expressas, entre outros.

Com base nestes relatos foram levantados os requisitos necessários para que o software possa ser utilizado como ferramenta preferencial para coleta e manipulação das informações que são recebidas no formato SAM ou BAM, que possuem informações de expressões gênicas provenientes do método Rna-Seq. Com estas informações, foi levantada a real necessidade dos envolvidos e estabelecido o escopo do projeto, esclarecendo os pontos discutidos, as necessidades de melhorias e também as facilidades que o programa deve oferecer a cada um dos usuários do

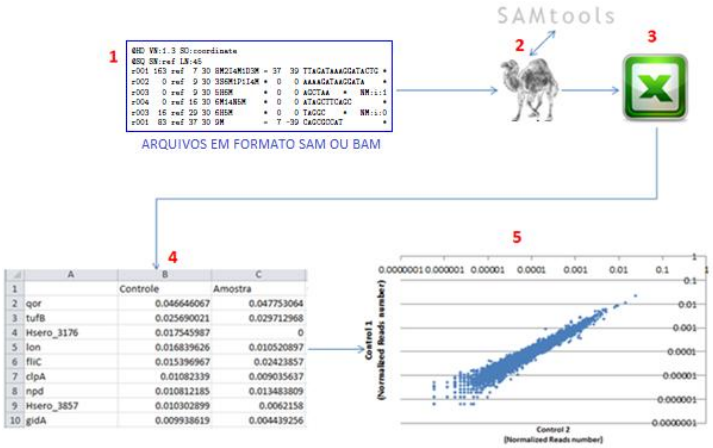
sistema. Após filtrar as requisições dos futuros usuários e analisar a necessidades de desenvolvimento, foram representadas graficamente as possibilidades de desenvolvimento e a abrangência do programa, conforme ilustrado na Figura 4.

Figura 4 - Imagem exibindo as necessidades levantadas e o escopo dos módulos que o projeto deve abranger para ser utilizado como ferramenta padrão de análise de amostras. FONTE: o autor (2011).



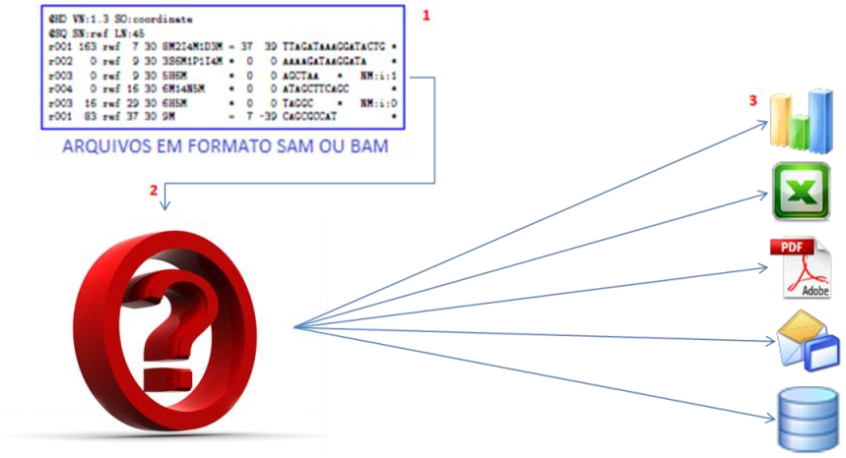
Foi levantada a informação de que atualmente existe uma rotina predefinida para análise de amostras provenientes dos sequenciadores da nova geração (Figura 5). Onde o usuário precisa converter estes arquivos para um formato texto e após a conversão dos dados deve importar estes dados para um software de manipulação de planilhas eletrônicas, preparar os cálculos necessários manualmente e após todo o processo visualizar os gráficos condizentes, que no caso descrito são apenas referentes à dispersão dos genes na comparação dos Reads de duas amostras distintas.

Figura 5 - Demonstração do fluxo atual para comparação de amostras utilizando as ferramentas já existentes. FONTE: o autor (2011).



Focando na necessidade de melhorias e nas possibilidades de agregar valores, um novo fluxo de trabalho foi proposto com o uso da nova ferramenta desenvolvida, seguindo os passos indicados na Figura 6, onde os arquivos no formato SAM ou BAM são importados diretamente para o software GTA e através dos tratamentos automáticos do software o usuário já pode executar as diversas tarefas disponíveis pelo programa.

Figura 6 - Imagem demonstrando o novo fluxo proposto para comparações, agregando novas funcionalidades e diminuindo a intervenção humana durante os processos. FONTE: o autor (2011).



4.2. Concepção, desenvolvimento e arquitetura do programa

O desenvolvimento inclui a concepção do *software* e o teste inicial do sistema. Para obtenção das necessidades e possibilidades de desenvolvimento do sistema foram entrevistados sete profissionais que desenvolvem o trabalho de pesquisa e análise no laboratório de Bioquímica da UFPR e as características iniciais do sistema estão ilustradas na Tabela 3.

Tabela 3 - Levantamento inicial dos requisitos para o desenvolvimento do sistema GTA. FONTE: o autor (2011).

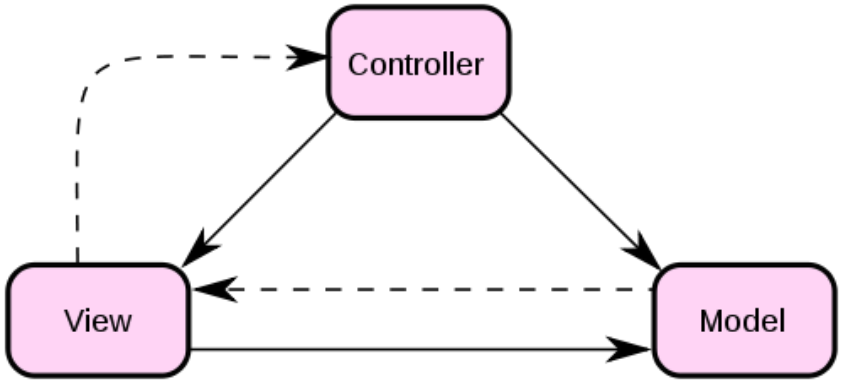
REQUISITOS FUNCIONAIS	REQUISITOS NÃO FUNCIONAIS
Importação dos dados brutos.	O sistema deve ser simples e prático.
Comparar e visualizar as diferenças entre as amostras cadastradas.	O sistema deve atender ao usuário com rapidez e eficiência
Armazenamento dos dados brutos e também os resultados analisados.	Desempenho do sistema: alto grau de assertividade nas comparações
Mobilidade dos dados, oferecendo opções de arquivamento local e envio para terceiros dos resultados analisados.	Alterações: O sistema deve permitir alterações futuras de forma simples e prática no nível de código.

Os principais requisitos levantados referem-se ao armazenamento dos dados e a facilidade de sua respectiva manipulação. O sistema tem como núcleo a base de dados embarcada que fornece não apenas as informações dos genes dos organismos cadastrados, mas também é referenciada constantemente no momento do cadastro dos dados das amostras. Outro aspecto importante do sistema é a utilização do cálculo de RPKM para análise dos genes das amostras comparadas, visando maior acurácia nas comparações correntes.

A arquitetura do GTA é divida em três camadas seguindo o conceito *Model-View-*Controller (MVC), um padrão de arquitetura de aplicações que visa separar a lógica da

aplicação da interface do usuário e do fluxo da aplicação (Trygve Reenskaug, 1979). A utilização deste padrão de projeto garante que futuras alterações no sistema sejam feitas sem causar grandes impactos no software como um todo, devido ao fraco acoplamento entre as classes do sistema (Figura 7).

Figura 7 - Conceito de MVC, exibindo seus principais componentes e fluxo de trabalho. FONTE: Trygve Reenskaug (1979).

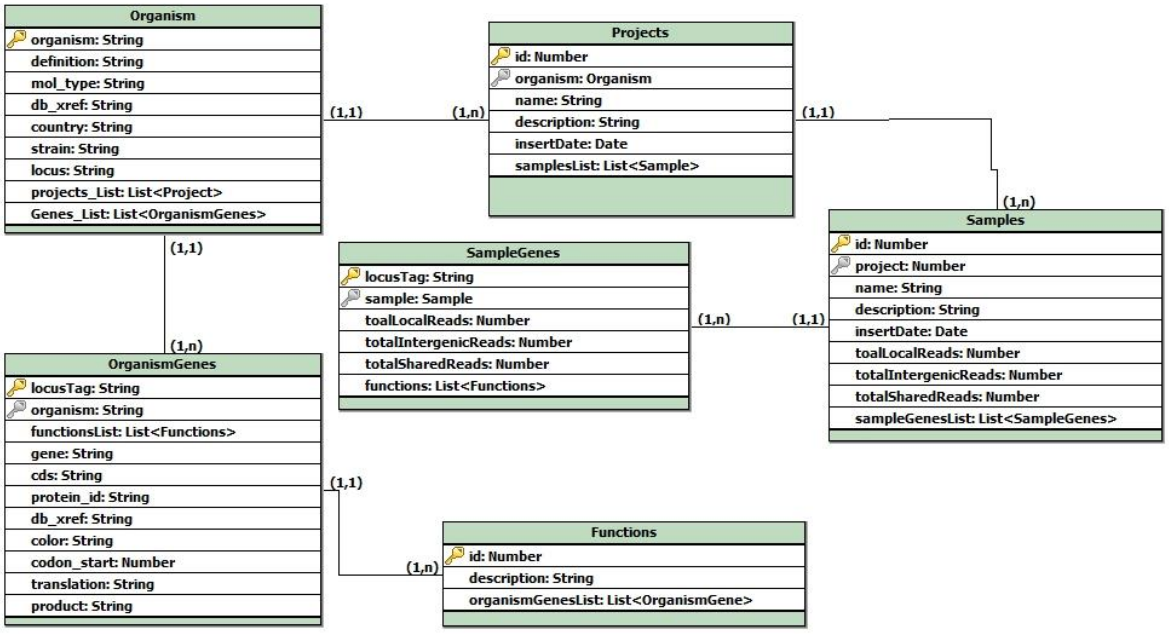


Com os requisitos levantados, os conhecimentos bibliográficos estudados, as tecnologias definidas e a base de dados arquitetada, foi iniciada a etapa de codificação do sistema GTA pelo desenvolvedor responsável, etapa esta que passou por constantes validações de usabilidade e funcionalidade pelos profissionais do setor químico da UFPR envolvidos no projeto.

A modelagem da base de dados (Figura 8) foi executada utilizando ferramentas nativas ao ambiente de desenvolvimento Netbeans®. As tabelas apresentadas representam os dados lidos durante o cadastro dos organismos através dos arquivos GBK, visando facilitar o entendimento e a manutenção do sistema. A tabela *Organism* é responsável por manter arquivadas as informações referentes ao organismo da mesma forma que pode ser encontrada em um arquivo no formato do Genbank, sendo que sua lista de genes esta disponível na tabela *OrganismGenes*, a qual possui todos os genes e suas informações essenciais para que o usuário possa referenciar durante o cadastro de novas amostras genéticas condizente com o organismo cadastrado. A tabela *projects* contém referências para todas as amostras (samples) cadastradas no software pelo usuário, que por sua vez possui conexão direta com a tabela *SampleGenes*, tabela aonde os dados de cada nova amostra cadastrada pelo usuário irá disponibilizar

suas informações de genes. A tabela *Functions* descreve a função definida para cada gene cadastrado, permitindo assim a construção de gráficos e a demonstração dos dados do gene selecionado de forma tabulada.

Figura 8 - Disposição das tabelas da base de dados do sistema, permitindo acesso e manipulação mais rápida dos dados inseridos. FONTE: o autor (2011).



Com a base de dados desenhada e as principais funcionalidades do programa já definida, foi iniciada a etapa de construção da interação entre o programa e seus usuários. Estudos foram feitos junto aos profissionais da área bioquímica visando proporcionar uma experiência agradável na utilização da ferramenta e seus opcionais. Toda a parte de visualização e manipulação de dados condizente ao usuário procurou ser o mais intuitiva possível, colocando as principais funcionalidades a poucos cliques de distância. A maioria destas funcionalidades não exige que o usuário saia da visualização principal do sistema, sempre abrindo as opções em novas abas ou em janelas suspensas, alterando a visualização somente se o usuário confirmar as alterações. A Figura 9 representa o diagrama contemplando as classes condizentes com as ações que o usuário pode tomar a partir da tela principal do programa. Nota-se

que a única classe que não possui ligação direta com a classe principal do sistema é a classe *MailConfiguration*, não impactando de forma alguma no fluxo normal do programa.

Figura 9 - Representação das classes e das opções de visualização, manipulação e exportação dos dados comparativos pelo usuário. FONTE: o autor (2011).

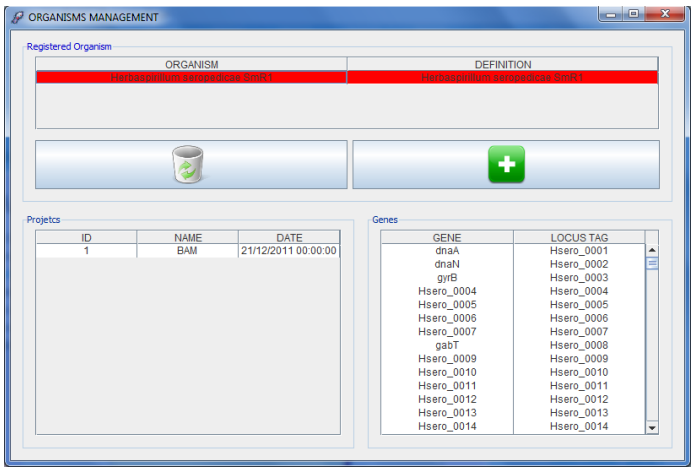
4.3. Utilização Do Sistema Gta

Para a utilização do sistema GTA, é necessária a obtenção do instalador e um período curto de treinamento, que pode ser concluído com a leitura do manual do usuário, o qual acompanha o pacote de instalação. Os testes foram executados durante todo o processo de desenvolvimento, tornando o sistema confiável e de acordo com s expectativas de todos os envolvidos. O software conta com inúmeras funcionalidades que atendem a todos os requisitos levantados na etapa anterior e também fornece aos usuários ferramentas que permitem uma maior riqueza na visualização, armazenamento e manipulação das amostras. Estas funcionalidades estão descritas do item 4.3.1 ao 4.3.4.

4.3.1. Cadastro De Novos Organismos

O cadastro de novos organismos pode ser executado pela tela de manipulação de organismos (Figura 10). O sistema aceita apenas organismos que estejam descritos em arquivos no formato GenBank, uma forma de compartilhar sequências de nucleotídeos, proteínas, genomas, etc. (Genbank, 1982) . A leitura dos arquivos no formato GBK são executadas através do uso da ferramenta *JGBParser*, uma ferramenta de uso livre desenvolvida por Dieval Guizelini em 2010 como requisito para obtenção de grau de mestre pela Universidade federal do Paraná. O programa só permite o cadastro de novas amostras se ao menos um organismo estiver cadastrado, o que garante a consistência dos dados cadastrados e que os cálculos executados no momento da inserção de novas amostras sejam executados corretamente.

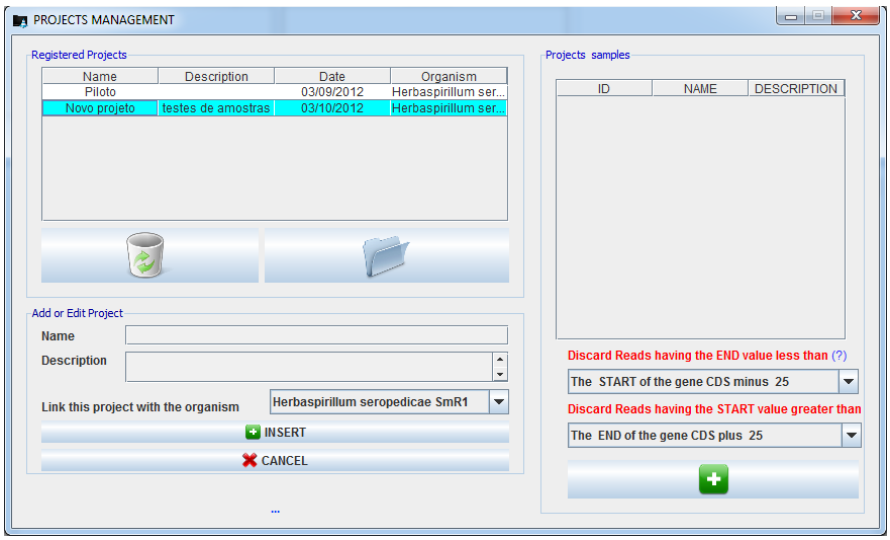
Figura 10 - Tela de cadastro de novos organismos do programa GTA. FONTE: o autor (2011).



4.3.2. Cadastro De Projetos E Amostras

O sistema oferece uma interface para cadastro e remoção de projetos e amostras (Figura 11). O cadastro das amostras é feito através da leitura de arquivos no formato BAM ou SAM, o usuário deverá sempre selecionar nomes distintos para cada amostra, não esquecendo que uma amostra pode fazer parte de apenas um projeto, conforme indica o diagrama da base de dados apresentado. O sistema não permite que o usuário cadastre um projeto que não contenha um organismo referenciado. No momento de cadastro da amostra o usuário pode escolher a tolerância para a área a esquerda e a direita do gene. Esta tolerância possibilita que mesmo *Reads* que não estejam localizados exatamente dentro das extremidades dos genes possam ser considerados e sua utilização esta descrita no manual do usuário. Caso o usuário tenha dúvidas de como o sistema trata estes limites, pode acessar a ajuda rápida clicando no ícone de interrogação (?) logo abaixo da tabela de amostras.

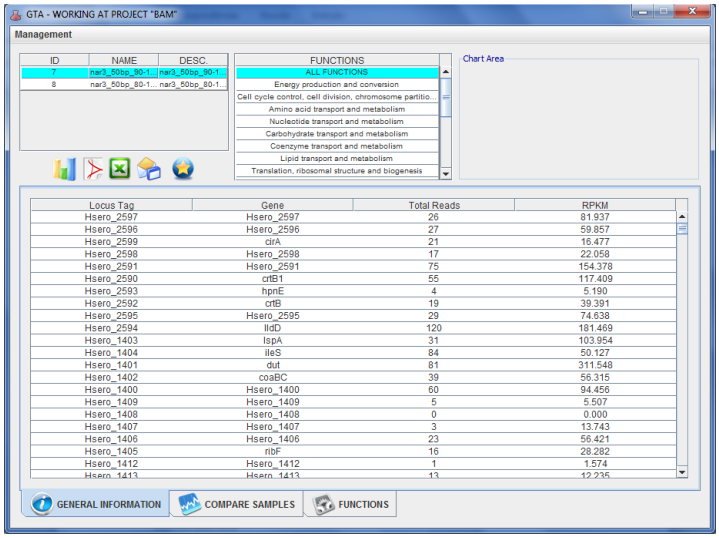
Figura 11 - Interface para cadastro de projetos e suas respectivas amostras no sistema GTA. FONTE: o autor (2011).



4.3.3. Visualização E Comparação De Amostras

Após o cadastro dos projetos e das amostras que se deseja comparar, o usuário poderá visualizar as informações referentes a cada amostra, detalhadas pelos nomes dos Genes expressos e seus respectivos valores de *Reads* e cálculos de RPKM relativos (Figura 12). O software disponibiliza também uma lista das funções dos genes de acordo com o organismo utilizado para cadastro da amostra. Estes dados não são estáticos, se alterando de acordo com os organismos selecionados, ou seja, uma função só será apresentada na lista de funções se fizer parte do organismo no qual o gene da amostra selecionada pelo usuário esteja relacionada, oferecendo ao usuário não apenas a opção de visualizar estes dados de forma gráfica mas também tabuladas.

Figura 12 - Interface de visualização dos genes de uma amostra. FONTE: o autor (2011).



Conforme proposto, o programa executa uma comparação entre os *Reads* encontrados em cada amostra cadastrada (limitado a selecionar duas amostras a cada comparação). Os valores são exibidos de forma tabulada ou em gráficos, conforme pode ser visualizado na Figura 13, na Figura 14 e na Figura 15.

Figura 13 - Exemplo de comparação de duas amostras cadastradas, exibidas de forma tabulada. FONTE: o autor (2011).

Gene	Size (Org)	nar3_50bp_90-16S	nar3_50bp_80-16S	nar3_50bp_90-16S/nar...	nar3_50bp_80-16S/nar...
Hsero_1181	1196	1237.475	1196.201	1.035	0.967
Hsero_1180	1076	227.910	177.281	1.286	0.778
Hsero_3923	1463	151.096	124.887	1.210	0.827
Hsero_3924	4676	93.071	91.910	1.013	0.988
Hsero_0092	602	137.699	152.707	0.902	1.109
Hsero_4449	2126	27.619	28.647	0.964	1.037
Hsero_4448	803	163.450	138.810	1.178	0.849
Hsero_2970	1370	886.179	884.069	1.002	0.998
Hsero_4447	1034	46.765	48.899	0.956	1.046
Hsero_4446	1640	11.583	11.211	1.033	0.968
Hsero_4445	1058	22.852	16.292	1.403	0.713
Hsero_4321	2660	12.985	11.232	1.156	0.865
Hsero_4444	1028	25.199	23.474	1.073	0.932
Hsero_4443	314	5.500	14.638	0.376	2.662
Hsero_2784	1955	34.451	25.863	1.332	0.751
Hsero_1873	1349	15.362	14.481	1.061	0.943
Hsero_3356	665	18.179	15.552	1.169	0.856
Hsero_3355	371	69.824	55.753	1.252	0.798
Hsero_1259	707	41.526	56.887	0.730	1.370
Hsero_3354	668	62.047	48.167	1.288	0.776

Show the results at column 4 only if it lies between 0 and 20 Show Infinity values at comparison

Figura 14 - Exemplo de gráfico de dispersão de duas amostras comparadas, onde cada ponto representa um gene expresso e a reta central representa o valor de regressão calculado para as amostras. FONTE: o autor (2011).

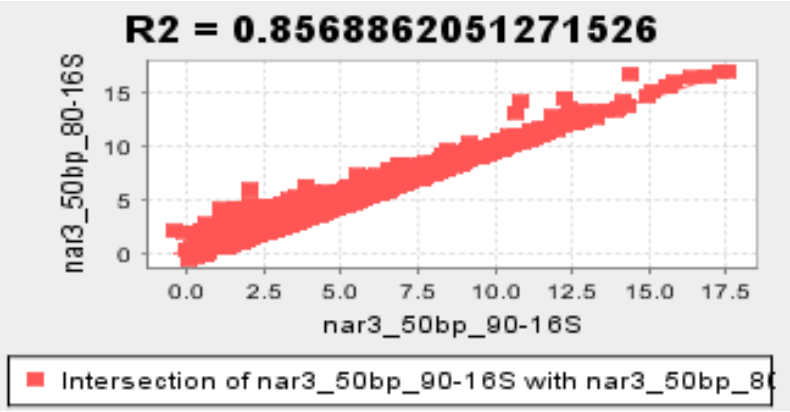
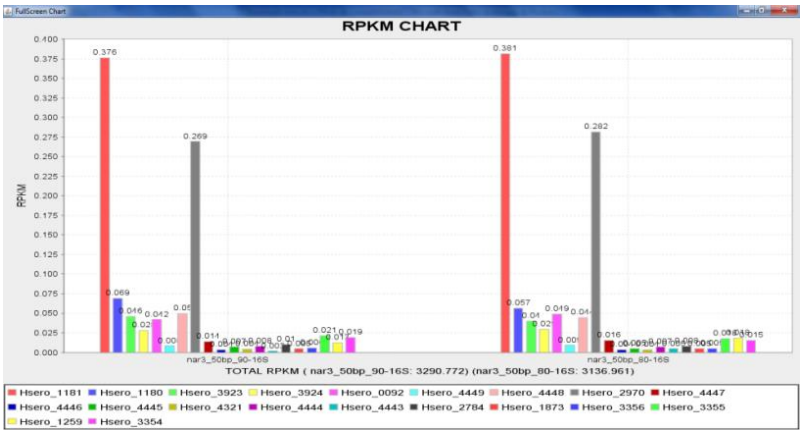


Figura 15 - Exemplo de gráfico comparando os valores de RPKM das amostras previamente selecionadas, exibindo de forma normalizada os valores relativos para cada uma individualmente. FONTE: o autor (2011).



É possível gerar também dados tabulados e gráficos exibindo o valor de *Reads* de cada função expressa pelos Genes nas amostras. Estes valores podem ser relativos a apenas uma amostra ou a um conjunto delas, tanto no formato de Pizza (Figura 16) ou de barras (Figura 17).

Figura 16 - Exemplo de gráfico de funções no formato de Pizza. FONTE: o autor (2011).

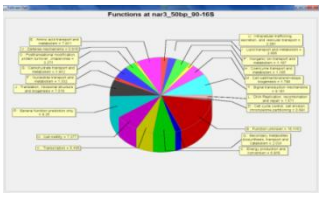
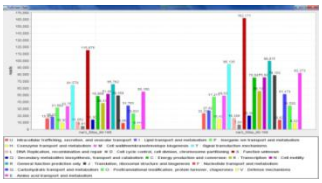


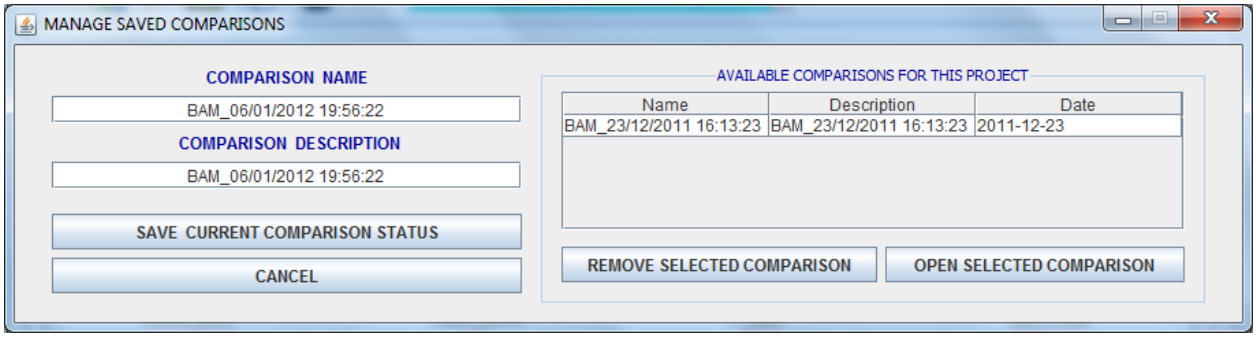
Figura 17 - Exemplo de gráficos de funções no formato de Barras. FONTE: o autor (2011).



4.3.4. Exportação Dos Dados

Tendo em vista a proposta principal do projeto de substituir o sistema Microsoft Excel®, foi inserida no programa uma opção para *salvar o estado atual de sua comparação*, ou seja, o usuário pode salvar sua pesquisa a qualquer momento e recuperar o estado desejado quando julgar necessário. A única restrição para que esta funcionalidade seja executada, assim como no programa Excel, é que os dados brutos estejam contidos na base de dados (Figura 18).

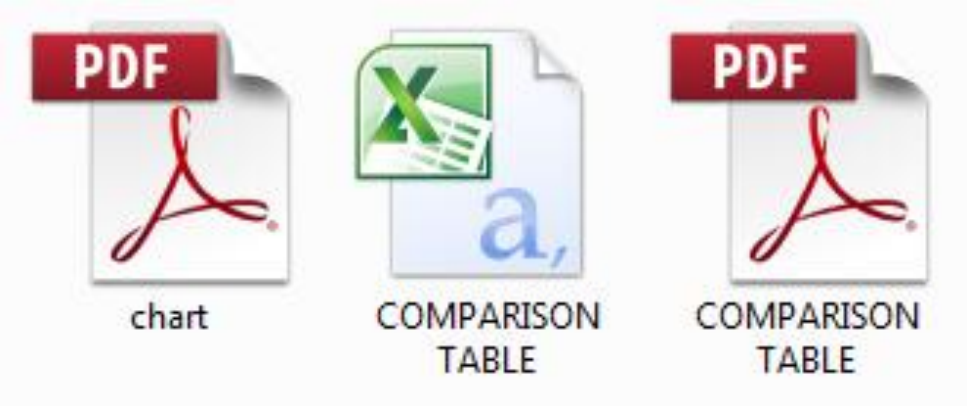
Figura 18 - Interface para que o usuário possa salvar ou recuperar os estados das suas comparações. FONTE: o autor (2011).



A ferramenta também oferece aos seus usuários a opção de enviar os dados comparados para um destinatário especifico de e-mail. Quando esta opção é utilizada, automaticamente é enviada como anexo uma cópia exata do estado atual da comparação, incluindo os dados tabulados e também os gráficos correspondentes. O destinatário recebe um e-mail contendo três arquivos distintos (Figura 19), sendo dois deles referentes aos dados tabulados, um em formato PDF® e o outro em um formato

separado por pontos e vírgulas (;), além de um arquivo separado contendo apenas o gráfico referente aos dados enviados.

Figura 19 - Exemplo de arquivos recebidos com comparações. FONTE: o autor (2011).



Estes dados também podem ser adquiridos diretamente pela interface da ferramenta, com as opções padrões de exportação de dados ou recuperados diretamente através do arquivo do banco de dados, contido na pasta de instalação do software.

4.4. Validações Do Projeto

4.4.1. Ambiente De Desenvolvimento

O hardware utilizado para o desenvolvimento do software pode ser facilmente encontrado no mercado local e contem as seguintes especificações:

- Modelo: Notebook Acer Aspire 5741G;
- CPU: Intel Core I5-430M com 2.26GHz e 3.0 MB de Cache L3;
- RAM: 4.0 GB;
- GPU: ATI Radeon HD5470 com 512MB Dedicados;

- HD: 500 GB.
- SO: Windows 7 Home Edition Premium 64 Bits

Os requerimentos mínimos para o funcionamento do sistema são: Processador de 1.2 GHz, 1024 MB de RAM e espaço mínimo em disco de 300 MB para armazenamento e manipulação, podendo varia de acordo com o tamanho das amostras selecionadas. As ferramentas de desenvolvimento e testes do software foram escolhidas com o intuito de jamais onerar valor financeiro ao programa e não intervir em nenhuma lei de proteção intelectual ou direitos autorais, dando preferência sempre às ferramentas *Open Source* (Código aberto) ou de domínio público. As ferramentas utilizadas são:

- IDE: Netbeans 7.0
- Linguagem de programação: JAVA 1.5
- Biblioteca para gráficos: JfreeChart
- Biblioteca para leitura de arquivos SAM e BAM: SamTools
- Biblioteca para envio de mensagens: Java Mail API
- Biblioteca para manipulação de arquivos PDF: IText
- Biblioteca de persistência: Hibernate
- Biblioteca para testes de integridade dos códigos: Junit

As ferramentas acima não necessitam estar presente nos computadores pessoais dos usuários, sendo necessárias apenas para o desenvolvimento do sistema. Outras metodologias e ferramentas foram utilizadas, mas por serem ferramentas incorporadas ou relacionadas às acima citadas, não necessitam ser descritas ou mencionadas neste documento.

4.4.2. Ambiente De Testes

4.4.2.1. Estudo de caso

Iremos apresentar neste item um breve estudo de caso da ação do GTA durante o processo de cadastro e exibição de uma amostra proveniente de um arquivo BAM. Não iremos apresentar os resultados de forma gráfica, mas sim o comportamento interno do sistema para ilustrar de forma simplificada o funcionamento da ferramenta proposta. Os passos seguem conforme item abaixo:

a) Leitura do arquivo BAM ou SAM

Existem algumas ferramentas específicas para a leitura de arquivos nos formatos SAM e BAM. Para o nosso projeto utilizamos a biblioteca *samtools* para ambiente Java. O método de leitura pode ser verificado no trecho de código ilustrado na Figura 20.

Figura 20 - Trecho de código responsável pela leitura do arquivo BAM Ordenado e da inserção destes valores em uma coleção contendo arrays de inteiros. FONTE: o autor (2011).

```
final SAMFileReader inputSam = new SAMFileReader(arquivo);
List<int[]> values = new ArrayList();
for (final SAMRecord samRecord : inputSam) {
    values.add(new int[]{samRecord.getAlignmentStart(), samRecord.getAlignmentEnd()});
}
```

b) Filtrando os Reads

Após preencher o vetor com todos os *Reads* lidos, o programa precisa percorrer o vetor, calcular o tamanho dos *Reads* e atribuir aos seus respectivos genes. O grande problema nesta iteração é que temos 13.566.990 milhões de *Reads* que devem ser iterados com aproximadamente 4800 genes do organismo cadastrado para encontrar seus valores, gerando um total de 65.121.552.000 (mais de 65 bilhões) de iterações.

Para amenizar esta curva de processamento, foram adotadas as seguintes metodologias: Ordenar e descartar *Reads* não candidatos. O funcionamento destes métodos é simples, sendo que o primeiro já é inerente ao arquivo BAM ordenado, conforme ilustra a Figura 21.

Figura 21 - Exemplo de leitura de um arquivo BAM onde os Reads estão ordenados pelo inicio de sua leitura. FONTE: o autor (2011).

```
Nome: 2239_999_589_F3 Start: 556 End:604
Nome: 2261_938_805_F3 Start: 556 End:604
Nome: 2341_1011_1592_F3 Start: 556 End:604
Nome: 2344_1969_1425_F3 Start: 556 End:604
Nome: 40_1033_1001_F3 Start: 557 End:603
Nome: 80_1461_335_F3 Start: 557 End:604
Nome: 103_1122_422_F3 Start: 557 End:604
Nome: 104_1341_251_F3 Start: 557 End:603
Nome: 142_1660_1948_F3 Start: 557 End:605
Nome: 167_526_733_F3 Start: 557 End:604
Nome: 208_541_31_F3 Start: 557 End:606
Nome: 230_1458_616_F3 Start: 557 End:604
Nome: 235_1215_1935_F3 Start: 557 End:604
Nome: 258_1783_467_F3 Start: 557 End:604
Nome: 275_796_371_F3 Start: 557 End:603
```

A segunda etapa consiste em analisar os *Reads* para descobrir os pertencentes ao gene sendo verificado no momento. Para que um *Read* seja considerado valido para um gene, o mesmo tem que atender ao menos uma das duas regras de validação. Estas regras são consideradas peças chaves para que o sistema execute a verificação de um grande número de *Reads* sem exigir um alto processamento. A primeira condição é ignorar os reads que não estão ao alcance do gene analisado. Este procedimento é feito antes das equações de verificação de reads contidos, pois evita que uma parcela de reads não candidatos seja analisada. Esta regra é demonstrada Equação 2, onde IR é o inicio do *Read*, FG o fim da área ocupada pelo gene e C uma constantes definida pelo usuário no momento do cadastro da amostra.

Equação 2

$$candidato\ se\ IR \leq FG + C$$

FONTE: o autor (2011).

Analisando a Equação 2, entendemos que ela desagrega do software a necessidade de percorrer todos os *Reads* de determinada amostra a procura de candidatos, pois quando encontra o primeiro read que não atende a esta equação, entende que todos os *Reads* após o verificado não serão candidatos também, devido ao fato de estarem ordenados pelo seu valor inicial. Passando por este primeiro filtro, o *Read* deverá satisfazer apenas uma das duas condições expostas, a primeira recebe o nome de ***contido pela esquerda***, conforme ilustra Equação 3, onde *FR* é o fim da leitura do *Read*, *IG* o inicio da área ocupada pelo gene, *FG* o fim da área ocupada pelo gene e *C* uma constante inserida pelo usuário no momento o cadastro da amostra.

Equação 3

$$contido\ pela\ esquerda\ se\ FR \geq IG - C\ \&\ FR \leq FG + C$$

FONTE: o autor (2011).

Para que um *Read* esteja contido pela esquerda do gene analisado, ele precisa estar com o seu inicio contido na região ocupada pelo gene, com uma aceitação de deslocamento máximo à esquerda e á direita conforme inserido pelo usuário (C). Esta distância limite é inserida no momento do cadastro da amostra, podendo variar de uma amostra para outra, mas não se alterando para os *Reads* da mesma amostra. Caso esta condição não seja satisfeita, a próxima condição que pode validar o read é a condição do ***contido pela direita*** (Equação 4), sendo *IR* o inicio do *Read, FR* o fim do *Read, IG* o inicio da área ocupada pelo gene e *FG* o fim da área ocupada pelo gene.

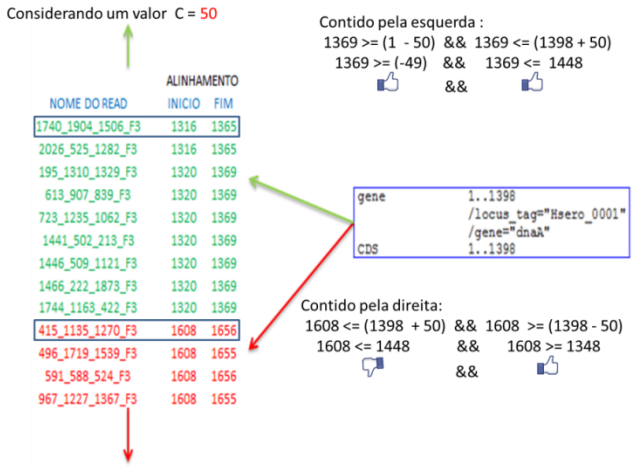
Equação 4

$$contido\ pela\ direita\ se\ IR \leq FG + C\ \&\ IR \geq IG - C$$

FONTE: o autor (2011).

A diferença entre a regra do *contido pela direita* e a regra do *contido pela esquerda* é a extremidade do read analisado, sendo uma verificando o inicio do read e a outra o fim do read. Sempre será verificada a regra do contido *pela esquerda* por primeiro, pois o programa irá ignorar não apenas o read que não preencheu nenhum dos requerimentos, mas também todos os reads que tenham seus valores iniciais maiores do que os reads analisados, evitando assim verificações desnecessárias. A Figura 22 ilustra de forma simplificada o processo de busca por reads válidos e como o programa se comporta ao encontrar reads não pertencentes.

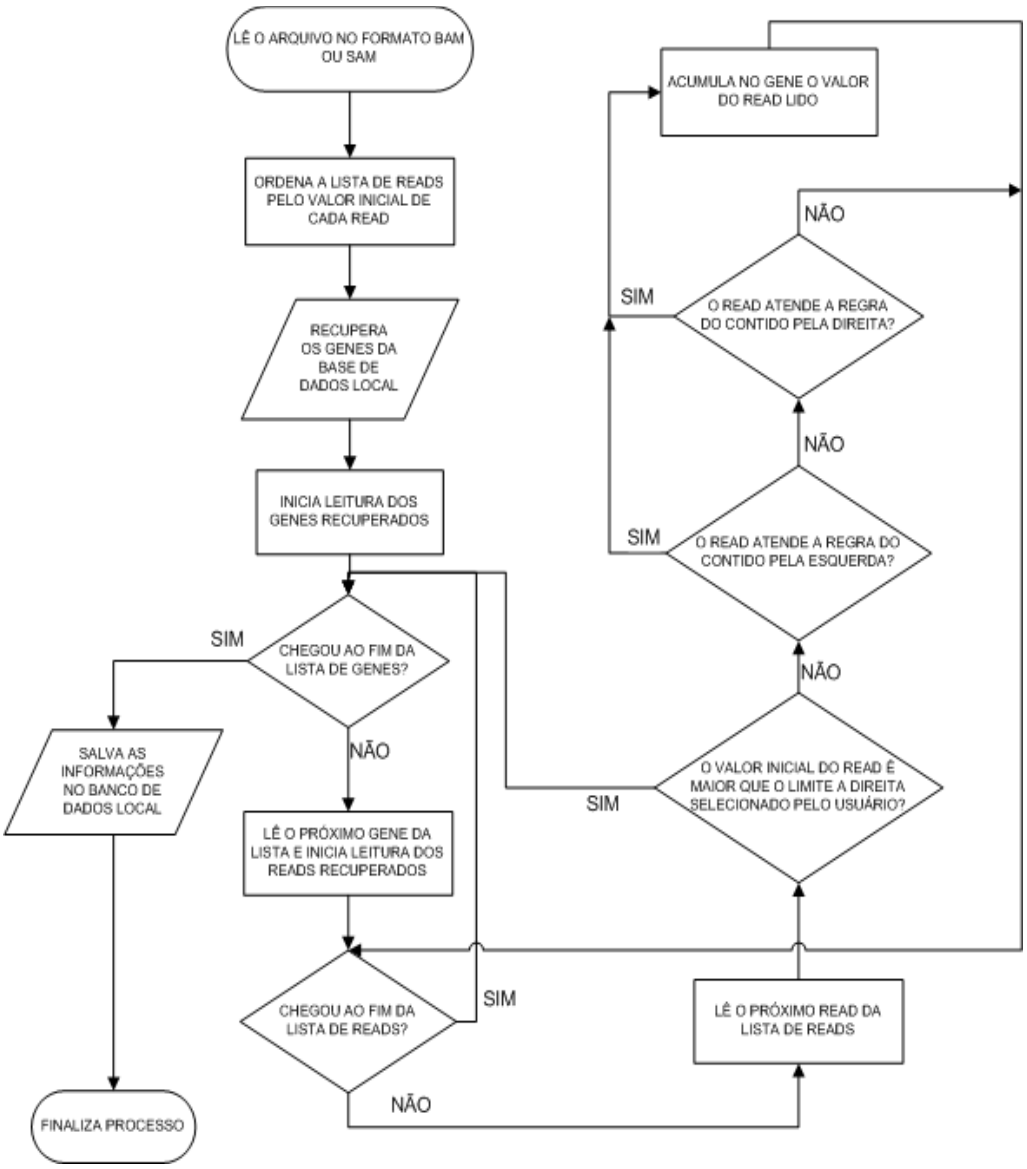
Figura 22 - Exemplo do processo de validação do Reads. FONTE: o autor (2011).



A Figura 23 exibe em forma de fluxograma como o sistema reage na leitura de cada *Read* respeitando a ordem de validação, fazendo assim que o número de

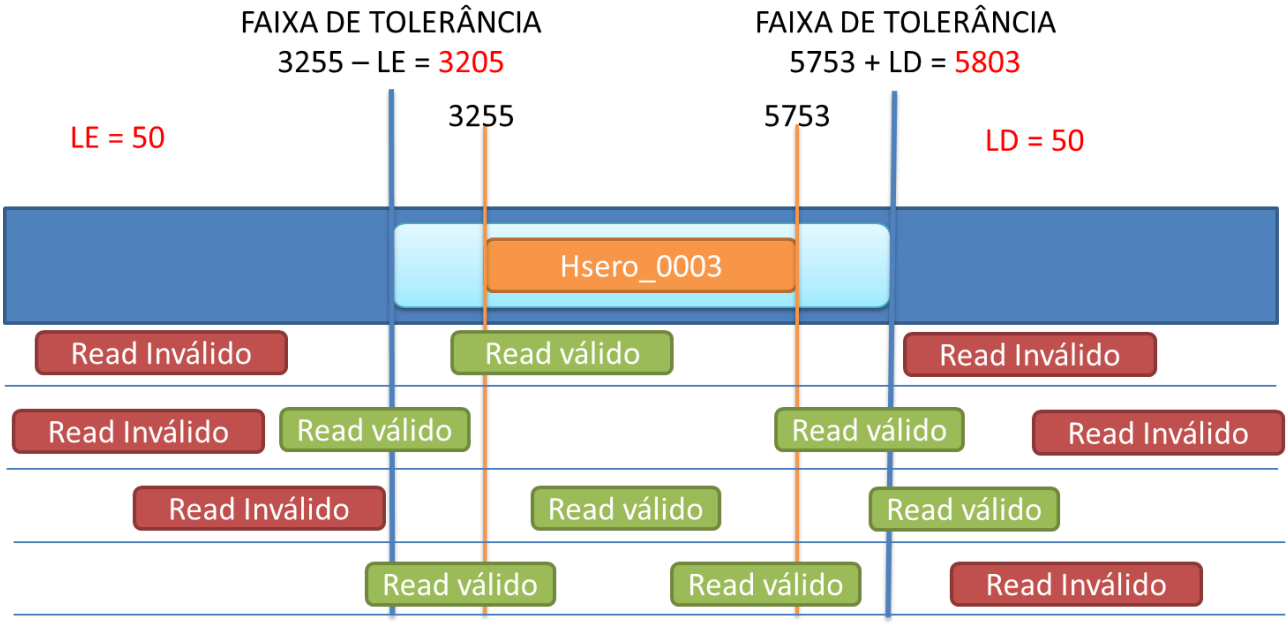
iterações entre os genes do organismo e os *Reads* do arquivo seja sempre baixo. Simplificando, a Figura 23 nos demonstra que ao cadastrar uma nova amostra, o software ordena os dados lidos da amostra pelos valores iniciais dos *Reads*, pesquisa no banco de dados os genes do organismo relacionado a ela e em seguida pesquisa na base de dados locais pelos genes do organismo correspondente a amostra, iniciando em seguida uma iteração pelos genes lidos, iterando também cada gene com os *Reads* recuperados do arquivo, até que uma das condições de parada seja satisfeita, podendo ser esta a chegada ao fim da lista de *Reads* ou se o *Read* lido não satisfazer a regra do contido pela direita. Para cada *Read* lido e que atenda ao menos a uma das regras especificadas, o sistema atribui seu valor ao montante de *Reads* já relacionados ao gene analisado. Este ciclo se repete até o sistema chegar ao fim da lista de genes do organismo selecionado. Após o término da iteração, a amostra é salva na base de dados local contendo todas as informações dos genes e seus valores de expressão de acordo com os *Reads* a eles relacionados.

Figura 23 - Exemplo do fluxo para validações de Reads lidos de arquivos no formato SAM e BAM. FONTE: o autor (2011).



Analisando o fluxograma da Figura 23, entende-se melhor como o sistema valida cada read individualmente, tratando os valores de tolerância á esquerda e á direita. A imagem da Figura 24 não demonstra o procedimento para exclusão de reads em massa, conforme descrito no fluxograma da Figura 23, mas sim o comportamento pontual para cada read.

Figura 24 - Exemplo de *Reads* válidos e inválidos de acordo com as regras de validações de *Reads*. FONTE: o autor (2011).



4.4.2.2. Teste De Desempenho Relativo

Os testes de desempenho do sistema ocorreram em computadores pessoais (*notebooks* e *netbooks*) de propriedade de alguns dos profissionais da área de bioquímica da Universidade Federal do Paraná, para garantir que mesmo computadores com um nível de processamento padrão aos encontrados em lojas o software pudesse ser executado de forma aceitável aos parâmetros de desempenho estipulado na projeção do sistema.

A Tabela 4 exibe o tempo de processamento em minutos para a tarefa mais demorada do sistema, que é carregar a nova amostra em memória, filtrar as informações do arquivo BAM ou SAM e salvar estes dados no banco de dados local. A amostra utilizada para os testes possuía inicialmente um montante de aproximadamente 14 milhões de Reads mapeados.

Tabela 4 - Comparativo do tempo de processamento para o cadastro de uma amostra com 14 milhões de Reads em diferentes computadores com hardware e sistemas operacionais distintos. FONTE: o autor (2012).

Processador	Memória	Sistema Operacional	Tempo (Minutos)
Athlon NEO mv-40 1,6 Ghz	2.0 GB	Windows Vista 32 Bits.	10
Intel Atom N270 1,6 Ghz	2.0 GB	Windows 7 Starter Edition 32 Bits.	13
AMD Turion 64 X2 2.6 GHz	3.0 GB	Debian 5.0 32 Bits.	4
Intel Core I5 2.26 GHz	8.0 GB	Windows 7 Home Edition 64 Bits.	8

É possível verificar através dos dados exibidos na tabela 3 que o desempenho do programa é condicionado não apenas ao processador e a memória disponível, mas também ao sistema operacional utilizado, principalmente quando verificado a enorme diferença no tempo de processamento entre um computador contendo o sistema operacional *Linux* e outro com um maior poder de processamento e memória RAM, mas com sistema *Windows*, pois o *Linux* tende a trabalhar mais agilmente com manipulação de arquivos em disco.

4.4.2.3. Testes De Desempenho Comparativo E Opções Disponíveis

Diferente do teste de desempenho relativo, onde o foco é mostrar o comportamento da solução em diferentes computadores, o teste de desempenho comparativo visa exibir as vantagens e desvantagens do uso da ferramenta como alternativa a outras ferramentas disponíveis, sendo elas:

4.4.2.3.1. Gta X Planilhas eletrônicas

Foram executadas comparações de tarefas simples com os dados obtidos pelo Rna-Seq e executadas pelo bioquímico a fim de obter os resultados esperados e estas comparações podem ser visualizadas na Tabela 5.

Tabela 5 - Comparativo de desempenho computacional entre o sistema GTA e a ferramenta padrão Excel, onde os tempos são exibidos em minutos. FONTE: o autor (2012).

	Planilhas eletrônicas	GTA
Cadastro do organismo (Herbaspirillum seropedicae)	- Não possui, os valores devem ser inseridos manualmente.	1,5 min. Trabalha diretamente com os arquivos no formato Genbank.
Registro das amostras	10 min. (deve ser preenchido manualmente e requer a conversão dos dados BAM para arquivo de texto tabulado, exigindo tempo extra não contabilizado aqui).	8 min Recebe diretamente os arquivos no formato BAM Ordenado.
Comparação da expressão gênica entre duas amostras	30 min. Deve ser preparado manualmente.	0,1min. Utiliza as normalizações com os dados do organismo utilizado durante o cadastro das amostras.
Geração de gráficos	0,1 min. Deve ser preparado manualmente.	0,05 min. Os dados para geração já estão disponíveis para vários tipos de gráficos.
Exportação e envio dos dados	- Não possui opção, seria	0,1 min Executa a exportação e o

	necessário salvar o arquivo e usar um programa de terceiros.	envio tanto no formato PDF como em CSV ou PNG.

Conforme exibido na Tabela 5, o *Genetic Transcript Analyzer* disponibiliza recursos não encontrados no Excel, o que dificulta uma real comparação, mas enfatiza a vantagem de se utilizar o sistema para obter os dados necessários. A maior vantagem do sistema GTA sobre o Microsoft Excel é a característica de trabalhar diretamente nos arquivos em formato BAM Ordenado, não exigindo assim um pré-processamento destes dados. Verifica-se também a enorme diferença de tempo quando a tarefa é a comparação entre as amostras, pois o programa GTA já contem todos os cálculos necessários já processador internamente, podendo otimizar este processo em até 180 vezes. Uma das desvantagens é a impossibilidade de executar comparações customizadas, como permite o Microsoft Excel, devido aos dados estarem sempre disponíveis em seu formato simplificado, neste caso arquivo texto.

4.4.2.3.2. Gta X Clc Dna Workbench

Uma das principais vantagens que o *CLC DNA WorkBench* possui sobre o sistema GTA é a possibilidade de obter os dados diretamente de uma das plataformas da nova geração, sem a necessidade de conversões nestes dados, agilizando ainda mais o processo de comparações. Por outro lado o *Genetic Transcript Analyzer* possui uma extrema facilidade em seu uso sem a necessidade de conhecimentos avançados em computação ou no próprio sistema.

4.4.2.3.3. Gta X Bioconductor

O Bioconductor consegue exportar diversos tipos de arquivos de sequências, incluindo o formato fasta, fastq, BAM, gff, bed e wig, entre muitos outros, enquanto na sua primeira concepção o GTA pode trabalhar apenas com os arquivos no formato BAM e SAM. Quanto à disposição das informações, apesar do GTA possuir menos

opções, as apresentadas são de fácil entendimento, manipulação e exportação, não necessitando que o usuário insira linhas de códigos ou scripts pré-processados para obter as informações desejadas. Ressalta-se também que por ser desenvolvido sempre com o acompanhamento dos profissionais da bioquímica, o GTA tornou-se uma ferramenta personalizada, que cumpre a promessa de exibir as informações de forma simples e eficiente.

5. CONCLUSÃO

O desenvolvimento do GTA mostrou que com poucos recursos e investimentos financeiros é possível criar ferramentas simples para auxiliar no trabalho de análise de dados e comparações. Todo o projeto foi baseado na necessidade de oferecer uma alternativa para que os bioquímicos pudessem executar suas comparações sem perder horas de trabalho preparando as amostras ou terem de recorrer a ferramentas comerciais, como por exemplo, o *CLC Dna Workbench* ou complexas como o Bioconductor. Em nenhum momento a ferramenta GTA visou substituir um destes dois programas em todas as suas tarefas, mesmo pelo conhecimento de que estes softwares estão consolidados no mercado e possuem alguns anos de melhorias e atualizações que os tornam cada dia mais promissores e funcionais. Ao contrario, o GTA teve desde sua concepção a missão de facilitar o acesso a dados complexos sem a necessidade de criação de scripts ou linhas de código complexas, tornando a experiência de análise de amostras algo mais agradável e proveitoso, sempre oferecendo simplicidade e confiabilidade nos dados analisados. Esta é a principal contribuição desta pesquisa.

O maior desafio na implementação da ferramenta foi a leitura rápida dos arquivos em formato BAM, que no inicio chegou a levar até 6 horas para ser concluída em uma amostra com 14 milhões de *Reads* mapeados, mas que após varias modificações no código e o auxilio da ferramenta *samtools*, hoje não passa de alguns minutos, conforme exibido no capítulo 4. Foram também executados inúmeros testes comparativos para que os dados exibidos no GTA não estivessem em nenhum momento relatando informações imprecisas ou tendenciosas, que pudessem invalidar o uso da ferramenta .

Até o momento da criação deste documento, estão sendo trabalhadas no programa novas ferramentas matemáticas que agregar novas informações as amostras analisadas. A cada nova melhoria da ferramenta, será fornecida aos usuários uma atualização contendo os pacotes necessários. Com base nas comparações executadas entre o GTA e as ferramentas disponibilizadas no laboratório, foi concluído que:

- A ferramenta GTA cumpre com a proposta inicial de ser uma alternativa a ferramentas já existentes.
- A ferramenta é de fácil utilização e adequação.
- O programa não exige de seus usuários conhecimentos de computação avançados

O programa possui grandes possibilidades e facilidades para agregação de novas funcionalidades, tornando-o de alta portabilidade para novas versões.

6. TRABALHOS FUTUROS

Esperamos que futuramente a ferramenta seja utilizada como base para desenvolvimentos que visem auxiliar no trabalho de análise ou até mesmo seja trabalhada para se tornar uma ferramenta ainda mais poderosa, desenvolvendo módulos adicionais que possibilitem a substituição de outras ferramentas e não mais ser visualizado como uma alternativa, mas sim como uma opção consolidada. Dentre algumas melhorias e implementações esperadas, destacam-se as seguintes:

- Criação de uma interface Web, podendo ser acessada de qualquer computador pessoal com acesso á internet.
- Leitura de outros formatos de arquivos, como por exemplo, fasta, fastq, gff, bed e wig.
- Leitura de outros formatos de arquivos com informações de Organismos.
- Inserção de novos cálculos, para agregar ao software módulos de análises estatísticas.
- Codificação do sistema em outras linguagens de programação.

7. REFERÊNCIAS

Apache foundation (2012). Disponível em http://www.apache.org. Acessado em setembro de 2011.

Applied ByoSystems, **A Theorical understanding of 2 base color codes and its application to annotation, error detection and error correction**. Applied Byosystems. 2008.

Anthony J.F. Griffiths, jeffrey H. Miller, David T. Suzuki, Richard C. Lewontin, William M. Gelbart, **An introduction to genetic analysis**, 7ª edição, Freeman, 1999.

Adams, M. et al. **Complementary DNAsequencing: expressed sequence tags and human genome project**. Science, 252, 1651–1656. 1991.

Amabis, J M ; Martho, G R. **Biologia** , Editora Moderna. 2004.

A.C. Pinto, H.P. Melo-Barbosa, A. Miyoshi, A. Silva e V. Azevedo. **Application of RNA-Seq to reveal the transcript profile in bacteria**. 2011.

Baldani, J.I., Baldani, V.L.D., Seldin, L. e Dõbereiner, J. 1986. **Characterization of Herbaspirillum seropedicae gen. nov., sp. nov., a root-associated nitrogen-fixing bacterium**. Int. J. Syst. Bacteriol.36:86-93.

BIOPERL (2012). Disponível em http://www.bioperl.org. Acessado em janeiro de 2012.

CLCBIO (2012). Disponível em http://www.clcbio.com/index.php?id=27. Acessado em janeiro de 2012.

Cloonan, N. et al. **Rna-mate: a recursive mapping strategy for high-throughput rna-sequencing data**. Bioinformatics. 2009.

Guizelini, Dieval. **Banco de dados biológico no modelo relacional para mineração de dados em genomas completos de procariotos disponibilizados pelo ncbi genbank**. Curitiba, 2010. 149 f. Dissertação (mestrado em bioinformática). Departamento de bioinformatica. Universidade Federal do Paraná.

Genbank Flat File Format Disponível em http://www.cs.sunysb.edu/~skiena/648/presentations/genbank.html#intro. Acessado em fevereiro de 2012.

Goncalves A, Tikhonov A, Brazma A and Kapushesky M. **A pipeline for RNA-Seq data processing and quality assessment**. Bioinformatics 27: 867-869. 2011.

Java (2011). Disponível em http://www.java.com. Acessado em janeiro de 2011.

Hibernate (2011). Disponível http://www.hibernate.org. Acessado em julho de 2011.

Jiang, H. andWong,W. **Statistical inferences for isoform expression in RNA-Seq. Bioinformatics**, 25, 1026–1032. 2009,

Kent, W.J. et al. **The human genome browser at UCSC**. Genome Res., 12,996–1006. 2002.

King, Gavin; Christian, Bauer. **Hibernate In Action (Second ed.),** Manning Publications, pp. 400, ISBN 193239415X. 2004.

Langmead, B. et al. **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. Genome Biol., 10, R25. 2009.

Lemmer, E. R., Friedman, S.L. & Llovet, J.M. **Molecular diagnosis of chronic liver disease and hepatocellular carcinoma: the potential of gene expression profiling**. Semin Liver Dis, 26: 373-384. 2006.

Li, H. et al. **Mapping short DNA sequencing Reads and calling variants using mapping quality scores**. Genome Res., 18, 1851–1858. 2008.

Moore, J.B., S.P. SHIAU, AND L.J. REITZER. **Alterations of highly conserved residues in the regulatory domain of nitro-gen regulator I (NtrC) of Escherichia coli**. 1993.

Mortazavi A, Williams BA, Mccue K, Schaeffer L, ET AL. 2008. **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. Nat. Methods 5: 621-628.

M. Morgan. **An introduction to Rsamtools**. Disponível http://prs.ism.ac.jp/bioc/2.6/bioc/html/rsamtools.html. Acessado em outubro de 2011.

Nagalakshmi U, Wang Z, Waern K, Shou C, et al. 2008. **The transcriptional landscape of the yeast genome defined by RNA sequencing**. Science 320: 1344-1349.

NCBI (2011). National Center for Biotechnology Information. Disponível em http://www.ncbi.nlm.nih.gov. Acessado em Novembro de 2011.

Nicolae, M., Mangul, S., M, I., and Zelikovsky, A. (2010). **Estimation of Alternative Splicing isoform Frequencies from RNA-Seq Data**. In Proceedings of the 10th International Conference on Algorithms in Bioinformatics, pages 202-214. Springer-Verlag.

Nils Home et al. 2009. **The Sequence Alignment/Map format and SAMtools**. Bioinformatics (2009) 25 (16):2078-2079.

Oracle (2011). Disponivel em http://www.oracle.com. Acessado em janeiro de 2011.

Pearson B.M., Gaskin D.J, Segers R.P., Wells J.M., Nuijten P.J., Van Vliet AH. **The complete genome sequence of Campylobacter jejuni strain 8111 (NCTC11828)**. J Bacteriol 2007;189:8402–3.

Samtools (2011). Disponível em http://samtools.sourceforge.net/SAM-1.3.pdf. Acessado em Novembro de 2011.

Shrimp (2011). Disponível em http://www.shrimp.com. Acessado em agosto de 2011.

The SAM Format Specification Working Group. **The SAM format specification v1.4-r985**. setembro de 2011.

Trygve Reenskaug. 1979. **Models - Views - Controllers. Technical note, Xerox PARC**. Disponível em http://heim.ifi.uio.no/~trygver/mvc/index.html. Acessado em novembro de 2011.

Wang, Z., Gerstein, M., and Snyder, M. (2009). **RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews. Genetics**, 10(1), 57-63.

Winkler, W. C., F. J. Grundy, B. A. Murphy, and T. M. Henkin. 2001. **The GA motif: an RNA element common to bacterial antitermination systems, rRNA, and eukaryotic RNAs**. RNA 7:1165–1172.

YONG, Chu Shao. **Banco de dados: organização sistemas e administração. São Paulo**: Atlas, 1983.