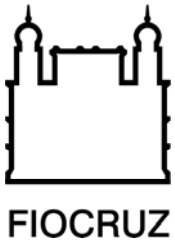


FUNDAÇÃO OSWALDO CRUZ
INSTITUTO GONÇALO MONIZ

Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa

DISSERTAÇÃO DE MESTRADO

ATUALIZAÇÃO E APRIMORAMENTO DO BANCO DE DADOS E DA FERRAMENTA DE GENOTIPAGEM DO HTLV-1 E DESENVOLVIMENTO DE FERRAMENTAS DE TIPAGEM PARA O HTLV E DE GENOTIPAGEM E FILOTIPAGEM PARA O HTLV-1 E HTLV-2

MURILO FREIRE OLIVEIRA ARAÚJO

Salvador - Bahia
2016

FUNDAÇÃO OSWALDO CRUZ
INSTITUTO GONÇALO MONIZ

Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa

ATUALIZAÇÃO E APRIMORAMENTO DO BANCO DE DADOS E DA FERRAMENTA DE GENOTIPAGEM DO HTLV-1 E DESENVOLVIMENTO DE FERRAMENTAS DE TIPAGEM PARA O HTLV E DE GENOTIPAGEM E FILOTIPAGEM PARA O HTLV-1 E HTLV-2

MURILO FREIRE OLIVEIRA ARAÚJO

Orientador: Prof. Dr. Túlio de Oliveira
Coorientador: Prof. Dr. Luiz Carlos Júnior Alcântara

Dissertação apresentada ao Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa para a obtenção do grau de Mestre.

Salvador
2016

DEDICATÓRIA

Dedico este trabalho, em primeiro lugar, à minha mãe, Doracy, que sempre me apontou para a direção correta e acreditou no meu potencial.

Dedico à minha esposa, Rosiene, pela paciência, compreensão e apoio durante a execução de todo este projeto.

Dedico também a meu filho Francisco, por ocupar uma fatia enorme de tempo da minha vida e por isso me mostrar que sempre devo ser mais produtivo, em qualquer ocasião que seja.

Dedico também aos meus professores, que consolidaram a base que hoje permite que eu desfrute da oportunidade de cursar proveitosamente um curso de pós-graduação em uma instituição de renome como a Fiocruz.

AGRADECIMENTOS

Agradeço ao meu orientador e ao meu coorientador, pela orientação, auxílio e paciência na execução deste trabalho.

Agradeço aos meus colegas do LHGB, principalmente ao Vagner e à Inês, por trabalharem comigo neste projeto.

Agradeço aos meus colegas da Seção de Tecnologia da Informação, pela paciência, suporte e por cobrir minhas ausências durante este período.

Agradeço também aos meus colegas de curso, que muito me auxiliaram durante as disciplinas, me dando explicações (às vezes miniaulas) e sanando minhas dúvidas.

Agradeço ainda às funcionárias da Coordenação de Pós-Graduação, que sempre se mostraram presentes na resolução das questões relacionadas ao curso.

Agradeço aos funcionários da Biblioteca do Instituto Gonçalo Moniz pela presteza e apoio em todas as situações em que necessitei dos seus serviços.

Por último, agradeço ao Instituto Gonçalo Moniz pela oportunidade de vivenciar um curso de pós-graduação nesta instituição de tanto prestígio perante a sociedade.

"Não sabendo que era impossível, foi lá e fez"

Mark Twain

ARAÚJO, Murilo Freire Oliveira. Atualização e aprimoramento do banco de dados e da ferramenta de genotipagem do HTLV-1 e desenvolvimento de ferramentas de tipagem para o HTLV e de genotipagem e filotipagem para o HTLV-1 e HTLV-2. 62 f. il. Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) – Fundação Oswaldo Cruz, Instituto Gonçalo Moniz, Salvador, 2016.

RESUMO

INTRODUÇÃO: O gerenciamento e a análise de dados biológicos através de métodos computacionais modernos necessitam que, em alguns casos, os pesquisadores submetam seus dados para diversos *softwares* distintos. A integração entre as ferramentas de bioinformática pode atenuar a problemática relacionada ao fluxo de dados biológicos entre várias aplicações diferentes, tornando mais transparente para o usuário o processo de obtenção de tipos específicos de informação, como a identificação automática de vírus e seus subtipos. **OBJETIVO:** Desenvolver ferramentas de bioinformática para a tipagem, genotipagem e de filotipagem para o *Human T Lymphotropic Virus* (HTLV) tipos 1, 2 ,3 e 4. **MATERIAL E MÉTODOS:** Primeiro analisou-se e modificou-se a ferramenta *REGA Genotype Tool* adicionando suporte para a genotipagem do HTLV tipos 1, 2 3 e 4. A segunda etapa aprimorou o *HTLV-1 Molecular Epidemiology Database*, reconstruindo seu banco de dados e as páginas *web* da aplicação. Adicionou-se a funcionalidade de *login*, *download* automático de dados e de auditoria e manutenção dos dados. O banco de dados passou a incluir sequências do HTLV-2, HTLV-3 e HTLV-4. Na terceira etapa, as ferramentas foram integradas através da construção de um *Servlet* no *REGA Genotype Tool* e de páginas específicas na aplicação de banco de dados capazes de realizar requisições e recuperar informações acerca das análises filogenéticas. **RESULTADOS:** A funcionalidade de genotipagem do HTLV-1 foi migrada para a nova versão do *REGA Genotype Tool e* adicionou-se a capacidade de genotipar os demais tipos do HTLV. Esta ferramenta foi otimizada visando um melhor desempenho e facilidades na inclusão de novos organismos no futuro. O Banco de Dados Público do HTLV-1 foi aprimorado: seu esquema foi normalizado e ganhou novas tabelas; As páginas foram refeitas utilizando padrões modernos de tecnologia *web;* Foram acrescidas as funcionalidades de *login*, *download* de sequências e auditoria e manutenção dos dados; E este passou a conter também sequências dos tipos 2, 3 e 4 do HTLV. Ocorreu a integração entre estas duas ferramentas, permitindo que o resultado de uma genotipagem seja recuperado pelo Banco de Dados e este possa realizar os tratamentos de arquivos e as consultas necessários para identificação dos tipos de vírus e seus subtipos. **CONCLUSÃO:** Foram desenvolvidas ferramentas para auxílio nos estudos referentes aos quatro tipos do HTLV, tanto em sua análise filogenética quanto provendo um repositório *online* de sequências e seus dados biológicos, geográficos, epidemiológicos e clínicos associados, o que contribuiu para a celeridade nas pesquisas relacionadas uma vez que o processo de análise dos dados do HTLV é simplificado através de aplicações confiáveis e otimizadas.

Palavras-Chave: HTLV, Bioinformática, Genotipagem, Banco de Dados.

ARAÚJO, Murilo Freire Oliveira. Update and improvement on database and genotype tool for the HTLV and development of typing tools for HTLV and of genotyping and phylotyping tools for HTLV-1 and HTLV-2. 60 f. il. Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) – Fundação Oswaldo Cruz, Instituto Gonçalo Moniz, Salvador, 2016.

ABSTRACT

INTRODUCTION: The management and analysis of biological data using modern computational methods needs, in some cases, that researchers submit their data to different software. The integration of bioinformatics tools can alleviate the problem related to the flow of biological data among several applications, making transparent to the user the process of obtaining specific types of information, such as the automatic identification of viruses and their subtypes. **OBJECTIVE:** Development of bioinformatics software application for typing, genotyping and phylotyping tools for the Human T Lymphotropic Virus (HTLV) types 1, 2 ,3 e 4. **MATERIAL AND METHODS:** In First, we analyzed and modified the REGA Genotype Tool by adding support for HTLV types 1, 2, 3 and 4. The second stage enhanced the HTLV-1 Molecular Epidemiology Database, rebuilding its database and the web pages of the application. We added the login functionality, automatic data download, audit and maintenance of data. The database now includes HTLV-2, HTLV-3 and HTLV-4 sequences. In the third stage, these tools were integrated by building a Servlet in REGA Genotype Tool and specific pages in the database application capable of performing requests and retrieve information about the phylogenetic analyzes. **RESULTS:** Genotyping feature of HTLV-1 has been migrated to the new version of REGA Genotype Tool and added the ability to genotype other types of HTLV. This tool has been optimized to provide a better performance and to facilitate the inclusion of new organisms in the future. The HTLV-1 Public Database has been improved: its scheme was normalized and gained new tables; the pages were refactored using modern standards of web technology; we added login, download sequences and auditing and maintenance of data functionalities; And it became able to also contain sequences of types 2, 3 and 4 of HTLV. The integration of these two tools occurred, allowing that the result of genotyping be recovered by the database and it can perform files treatments and queries necessary to the identification of the viruses' types and their subtypes. **CONCLUSION:** We developed tools to help in studies related to four types of the HTLV, both in its phylogenetic analysis as in providing an online repository of sequences and its respective biological, geographic, epidemiologic and clinical data, contributing to celerity in researches related, since the process of data analysis of HTLV were simplified through trustable and optimized applications.

Keywords: HTLV, Bioinformatics, Genotyping, Database.

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

Entre os objetivos da bioinformática, destaca-se o gerenciamento e a análise de dados biológicos através de métodos computacionais modernos. A utilização de bancos de dados é de grande valia para este fim, uma vez que estes podem ser utilizados para armazenamento e consulta de informações sobre genoma, como sequências de DNA, síntese de RNA e geração de proteínas. Deste modo, técnicas de processamento de dados são críticas para o desenvolvimento destes bancos de dados, fornecendo o arcabouço necessário para o projeto, acesso e gerência dos dados (LIFSCHITZ, 2016).

As ferramentas filogenéticas, por sua vez, são recursos utilizados no campo da virologia para estudar a evolução viral, traçar a origem de epidemias, estabelecer o modo de transmissão, pesquisar a ocorrência de resistência a medicamentos ou determinar a origem do vírus nos diferentes compartimentos corporais. Este processo envolve a construção de árvores filogenéticas, sua visualização e interpretação. A partir destas informações, são construídos os filotipos, que podem incorporar informações sobre região geográfica, data da isolação, hospedeiro, rota de transmissão, entre outros, permitindo assim a identificação de grupos monofiléticos, numa reconstrução filogenética, associados a dados geográficos, temporais e parâmetros epidemiológicos e clínicos (CHEVENET et al., 2013).

Todavia, torna-se necessário algumas vezes que para a obtenção de determinadas informações os pesquisadores submetam seus dados para diversas aplicações diferentes, em um processo de entrada (*input*), processamento, análise e saída (*output*), em alguns casos através de diversos *software*s distintos. Por conseguinte, configurações, parâmetros e comandos devem ser conhecidos pelo pesquisador, a fim de permitir o fluxo de informações entre as diversas ferramentas.

Ainda, existem situações em que são necessárias conversões de formato de arquivos, o que pode exigir trabalho adicional do pesquisador. A integração entre as ferramentas de bioinformática, ou seja, a automação da comunicação entre estas, pode atenuar esta problemática através da abstração do fluxo de dados, tornando mais transparente para o usuário o processo de obtenção de tipos específicos de informação, como os filotipos.

Como exemplo de método para processamento destes tipos de dados e obtenção destas informações, é possível descrever: a) Obtenção de sequências através de sequenciadores ou bases de dados biológicas b) Alinhamento destas sequências c) Tratamento dos dados com um editor de alinhamento, caso necessário; d) Criação de arquivos texto com o resultado do alinhamento para possibilitar a utilização de aplicativos de reconstrução filogenética para o processamento das árvores; e) E submissão do arquivo contendo a árvore resultante a um visualizador para construção da árvore gráfica. Ainda, para se obter os filotipos, é necessário que se crie uma estrutura de dados que relacione o resultado da reconstrução filogenética com suas informações associadas, como dados geográficos, clínicos, epidemiológicos, entre outros, e que permita a visualização organizada desta estrutura por parte do pesquisador.

Deste modo, com a intenção de tornar mais simplificada a tipagem, a subtipagem e a construção dos filotipos em estudos sobre o HTLV, utilizaremos a estrutura e os dados do *HTLV-1 Molecular Epidemiology Database*, desenvolvido pelo nosso grupo de pesquisa (ARAUJO et al., 2012), e a ferramenta de subtipagem *Rega Subtyping Tool* (ALCANTARA et al., 2009), para desenvolver ferramentas de tipagem, genotipagem e de filotipagem viral para o HTLV, o que inclui adicionar tanto no banco de dados quanto na ferramenta de genotipagem suporte para o HTLV-2, 3 e 4. Desta maneira, além de simplificar a reconstrução filogenética, estas ferramentas proporcionarão aos pesquisadores uma maneira mais simples de classificar sequências do HTLV, integrando duas aplicações que realizam processamento, armazenamento e consulta destes dados; permitindo criar relações entre os resultados da genotipagem e os dados geográficos, epidemiológicos e clínicos das amostras, além de gerar tabelas relacionando características comuns das sequências referência, como carga proviral, localidade, quadro clínico do paciente que forneceu a amostra, entre outros. Em consonância com isto, adicionaremos novas sequências deste vírus no Banco de Dados Público do HTLV e na ferramenta de genotipagem, incluindo no primeiro as sequências a partir de 2009 do HTLV-1 e todas do HTLV-2, 3 e 4, e, no último, sequências referência destes vírus, além de modificações na aplicação para adaptá-la à execução destas novas análises. Adicionalmente, criaremos no banco de dados um mecanismo de *download* automático de sequências a partir do *GenBank*, que facilitará o processo de atualização e inclusão de novas sequências na base de dados do HTLV.

Por último, serão implementados instrumentos de controle e auditoria sobre estes dados, a fim de se determinar sua origem e sua tramitação na aplicação, como informações sobre o usuário que realizou o *download*, data de inclusão no banco de dados, entre outras informações. Ao final, espera-se que, além da simplificação do processo relacionado à obtenção dos filotipos do HTLV, os recursos computacionais desenvolvidos para as análises do HTLV-1 estarão estendidos também para os demais tipos deste mesmo vírus.

2 OBJETIVOS

2.1 GERAL

Desenvolver ferramentas de bioinformática para a tipagem, genotipagem e de filotipagem para o HTLV (HTLV-1, HTLV-2, HTLV-3 e HTLV-4).

2.2 ESPECÍFICOS

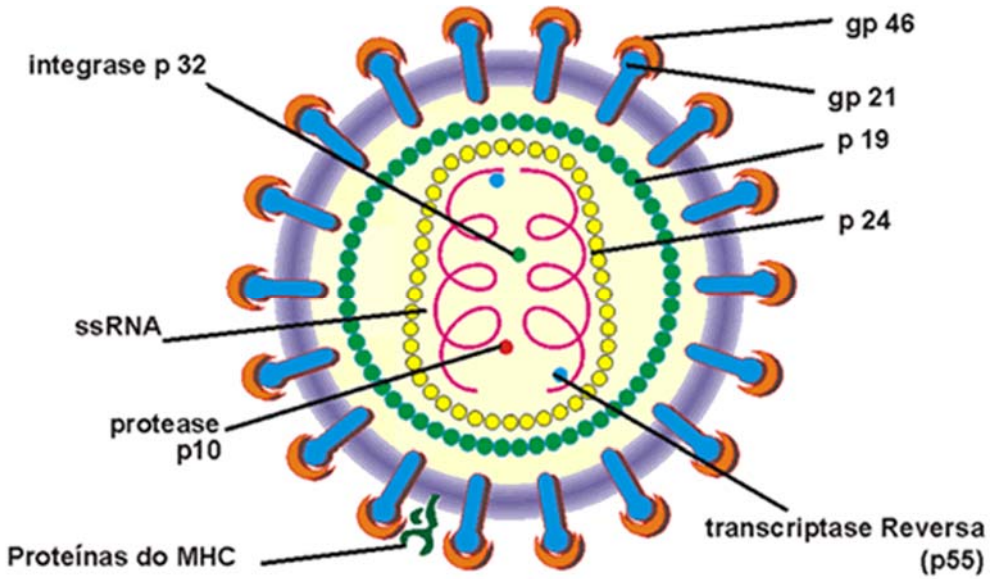
- Adicionar no *REGA Subtyping Tool* suporte para a genotipagem de sequências do HTLV-1 e 2;
- Adicionar no *REGA Subtyping Tool* suporte para a tipagem de sequências do HTLV-3 e 4;
- Adicionar no Banco de Dados Público do HTLV suporte para sequências do HTLV-2, HTLV-3 e HTLV-4;
- Desenvolver uma ferramenta para automatizar a obtenção de dados a partir do *Genbank*;
- Implementar no Banco de Dados Público do HTLV-1 ferramentas para manutenção e auditoria dos seus dados;
- Integrar o banco de dados *HTLV-1 Molecular Epidemiology Database* e a ferramenta de genotipagem *REGA Subtyping Tool*.

3 REVISÃO DE LITERATURA

3.1 HTLV

3.1.1 Origem e Caracterização

O Vírus Linfotrópico de Células T Humanas (HTLV-1) foi o primeiro retrovírus humano descrito, isolado por Poiesz e colaboradores em 1980, a partir da investigação de um paciente com linfoma cutâneo de células T. Apesar disto, uma forma distinta de leucemia com características clínicas e com morfologia celular especiais foi definida anteriormente em pacientes no Japão, nomeada de Leucemia de células T do Adulto (ATLL) (UCHIYAMA et al., 1977). Em 1980, os soros destes pacientes foram analisados, sendo positivos para anticorpos anti-HTLV-1 fornecendo evidências para a ligação do *HTLV-1* às células T malignas da ATLL (GALLO, 1981).



Fonte: adaptado do site http://www.htlv.com.br.

Figura 1. Modelo esquemático da estrutura do HTLV- 1

O HTLV-1 pode ser classificado em sete subtipos e seis subgrupos, de acordo às diferenças nas sequências do gene env e LTR do DNA proviral (Quadro 1).

Quadro 1 - Subtipos do HTLV-1.

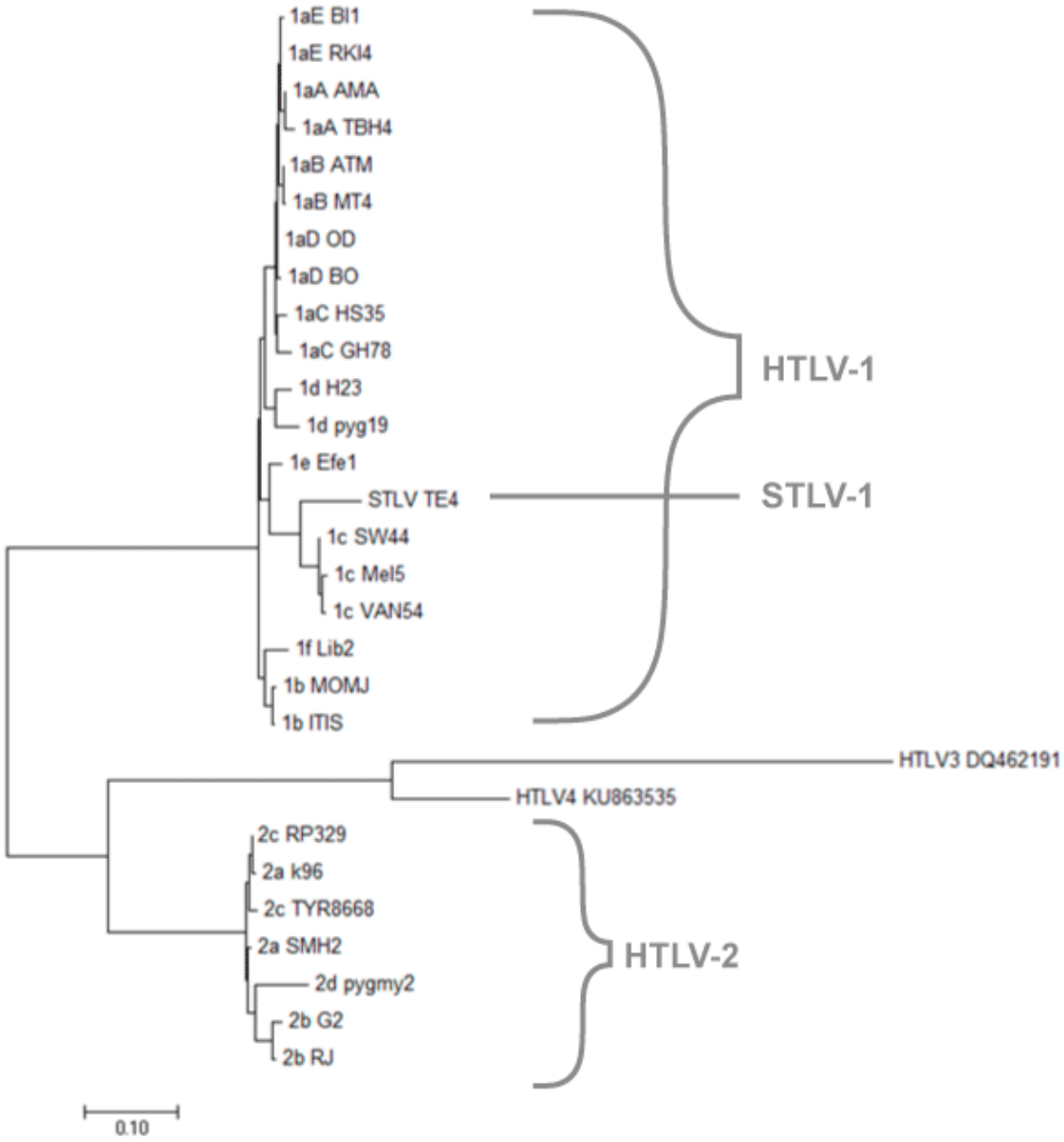
Subtipo	Subgrupo	Descrição	Referência
A	A –Transcontinental B – Japonês C – Oeste Africano D – Norte Africano E – Negro do Peru F – Espanha/África	Cosmopolita	(SEIKI; HATTORI; YOSHIDA, 1982) (VAN DOOREN et al., 1998) (TREVIÑO et al., 2014)
B	-	Central Africano	(GESSIAN et al., 1991; BASTIAN et al., 1993)
C	-	Melanésia	(GESSIAN et al., 1991; BASTIAN et al., 1993)
D	-	Isolado de pigmeus em Camarões e no Gabão	(CHEN et al., 1995; MAHIEUX et al., 1997)
E	-	Isolado de pigmeus na República Democrática do Congo	(SALEMI et al., 1998)
F	-	Isolado de um indivíduo do Gabão	(SALEMI et al., 1998)
G	-	Recentemente descrito como um novo subtipo em Camarões, na África Central	(WOLFE et al., 2005)

Fonte: (SEIKI; HATTORI; YOSHIDA, 1982), (VAN DOOREN et al., 1998), (TREVIÑO et al., 2014), (GESSIAN et al., 1991; BASTIAN et al., 1993), (CHEN et al., 1995; MAHIEUX et al., 1997), SALEMI et al., 1998), (WOLFE et al., 2005).

Em 1982, Kalyanaraman e colaboradores isolaram o HTLV-2 a partir de um paciente com uma forma atípica de Leucemia de células T pilosas, a tricoleucemia e notaram que este vírus apresentava diferenças antigênicas em relação ao *HTLV-1* (KALYANARAMAN et al., 1982). Este vírus apresenta menor divergência genética quando comparado ao HTLV-1 e isto pode ser parcialmente explicado pelo fato de o HTLV-2 ter sido isolado somente de duas espécies, o *Pan paniscus* (Bonobo) e o *Homo sapiens*. O HTLV-1, por sua vez, já foi isolado a partir de mais de 20 diferentes primatas, com múltiplos episódios de transmissão interespécies (SLATTERY et al., 1999). O HTLV-2 é caracterizado em três subtipos genéticos distintos: a, b e d, compartilhando cerca de 90% de similaridade de nucleotídeos, diferenciando-se através de análises nas regiões LTR e *env*.

O HTLV-1 e o HTLV-2 originaram-se independentemente e estão relacionados ao STLV (Vírus Linfotrópico da Célula T de Símios) tipo 1 e 2, respectivamente, apresentando inclusive maior similaridade com o STLV do que entre si.

Recentemente, foram descritos em indivíduos de Camarões, na África Central, dois novos tipos de HTLV: o HTLV-3 e HTLV-4 (WOLFE et al., 2005; CALATTINI et al., 2005). O HTLV-3 parece ser originado do STLV-3, entretanto não foi identificado um equivalente do STLV para o HTLV-4. A Figura 3 mostra uma relação evolutiva entre os diferentes tipos e subtipos do HTLV.

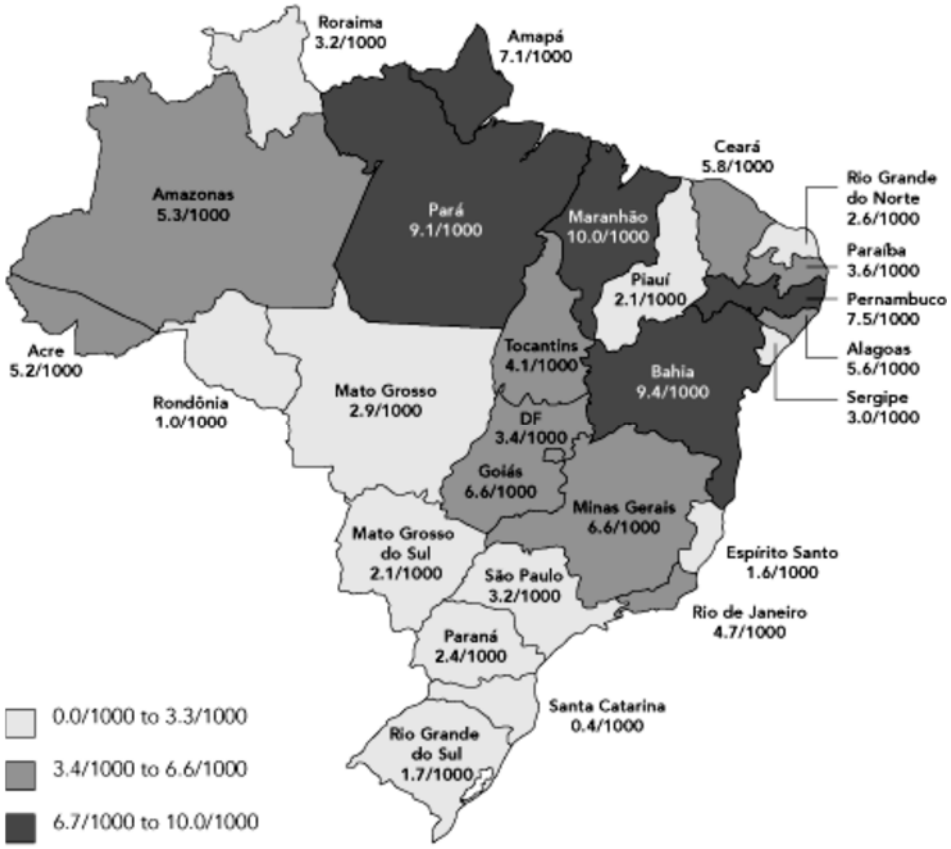


A figura mostra os tipos do HTLV agrupados, inclusive com o STLV (amostra STLV TE4, STLV-1) se agrupando junto com o HTLV-1. Fonte: o autor.

Figura 2. Árvore filogenética contendo sequências do HTLV-1, HTLV-2, HTLV-3 e HTLV-4

3.1.1 Epidemiologia

O Brasil constitui-se de uma área endêmica para o HTLV-1 (CATALAN-SOARES et al., 2004). Apesar disso, há variação no nível de prevalência nas áreas distintas do país. Os dados mostrados na Figura 2 foram compilados a partir de amostras originadas de bancos de sangue ou de amostras de grupos específicos (gestantes, pacientes de clínicas de doenças sexualmente transmissíveis, coinfectados) que não representam necessariamente a população geral (revisado por PROIETTI et al., 2005).



Fonte: Adaptado de CATALAN-SOARES et al., 2004.

Figura 3. Prevalência de HTLV-1 entre doadores de sangue em capitais de 26 estados brasileiros e no Distrito Federal.

Até o momento, apenas um estudo apresenta dados de base populacional e foi realizado por Dourado e colaboradores (2003) na cidade de Salvador - Bahia, uma das capitais com a maior prevalência de HTLV-1 no Brasil, em que demonstra-se uma

soroprevalência de 1,8%, estimando-se que existam cerca de 40 a 50 mil indivíduos infectados na cidade.

O HTLV-2 tem seus subtipos a e b são distribuídos mundialmente, podendo ser encontrados em usuários de drogas ilegais injetáveis (UDI) nos Estados Unidos, na Europa, no Vietnam e em ameríndios e tribos africanas (GESSAIN et al., 1995), enquanto o subtipo d foi identificado em 1998, no Congo (GESSAIN et al., 1995; VANDAMME et al., 1998).

Amostras do HTLV-3 e do HTLV-4 foram localizadas somente em indivíduos de Camarões, na África Central.

3.1.2 Doenças associadas

Entre as enfermidades associadas ao HTLV-1, encontra-se a paraparesia espástica tropical/mielopatia associada ao HTLV (TSP/HAM), caracterizada como uma doença neurológica crônico-degenerativa que atinge o sistema nervoso central causando, principalmente, o aumento da espasticidade dos membros inferiores (GESSAIN et al., 1985, OSAME et al., 1986). Em torno de 0.3-2% das pessoas infectadas apresentam esta patologia (EDWARDS et al., 2011), sendo afetadas mais mulheres do que homens e com a maioria dos indivíduos recebendo diagnósticos tardios por volta de 40 a 50 anos de idade (GESSAIN et al., 1985; OSAME et al., 1986).

As pesquisas em relação à determinação da doença em um indivíduo infectado ainda não são conclusivas, porém acredita-se na possível influência do modo de transmissão, da carga proviral, tipo e magnitude da resposta imune do hospedeiro contra os antígenos do HTLV-1, além de fatores genéticos individuais como polimorfismos em genes de *HLA* (*Human Lecocitary Antigen*) e genes envolvidos na resposta imune (MARTINS; STANCIOLI, 2006). Além da TSP/HAM, o HTLV-1 também é o agente etiológico da Leucemia/Linfoma de Células T do Adulto (ATLL), uma neoplasia de linfócitos T maduros, que ocorre devido a uma expansão monoclonal dos linfócitos T infectados (YOSHIDA et al., 1982), dos indivíduos infectados pelo HTLV-1, cerca de 3-5% desenvolvem ATLL (EDWARDS et al., 2011). O HTLV-1 também é associado a outras doenças inflamatórias, como a dermatite infecciosa associada ao HTLV-1 (LA GRENADE et al., 1998; GONÇALVES et al.,

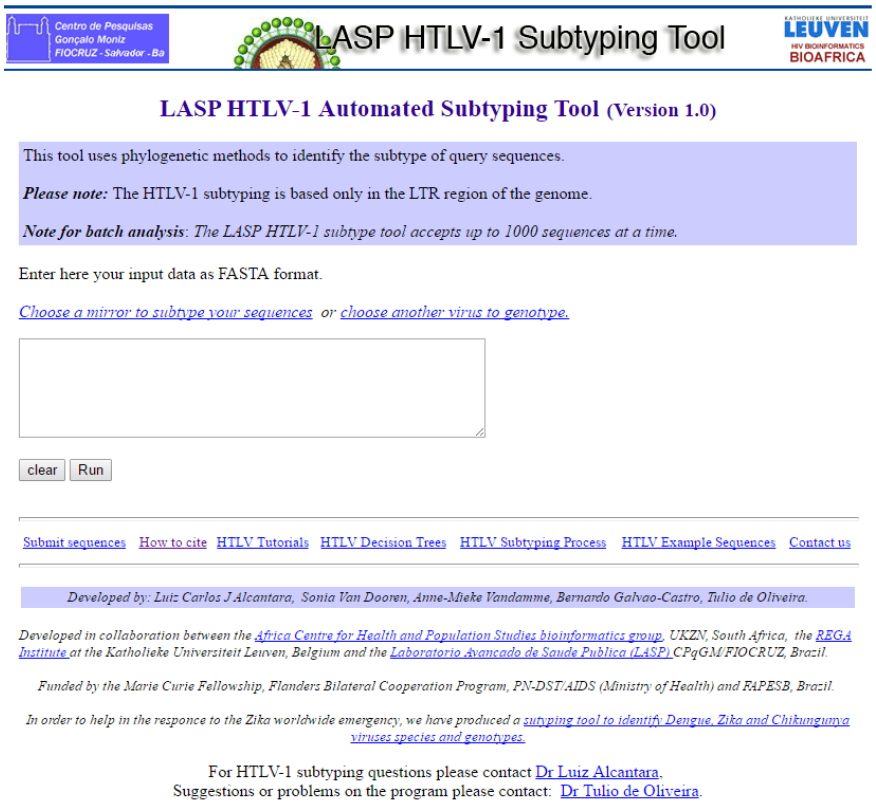
2003), uveíte associada ao HTLV (HAU) (MOCHIZUKI et al., 1992), além de doenças reumáticas como síndrome de Sjögren e artrite reumatóide (MCCALLUM et al., 1997, NISHIOKA, 1996) e ao aumento da prevalência de algumas parasitoses (PORTO et al., 2001). Até o presente momento, somente o HTLV-1 está comprovadamente associado a doenças humanas.

Assim como o HTLV-1, o HTLV-2 é transmitido através de sangue contaminado, intercurso sexual e de mãe para filho através da amamentação (LEE et al., 1989; SOARES et al, 2003). Embora este não possuísse, inicialmente, uma clara associação a doenças, trabalhos têm relatado que sua infecção pode estar associada a desordens neurológicas e a um aumento na taxa de doenças infecciosas e linfoproliferativas (BIGLIONE et al., 2003; BISWAS et al., 2010; MURPHY et al., 1997; SILVA et al., 2002).

Atualmente, não se sabe se o HTLV-3 e o HTLV-4 podem ser transmitidos entre seres humanos e se são capazes de desenvolver patologias nos hospedeiros (GESSAIN et al., 2013; WOLFE et al., 2005).

3.2 *HTLV-1 SUBTYPING TOOL*

A primeira versão da ferramenta de bioinformática para realizar a tipagem e subtipagem do HTLV-1 proposta pelo nosso grupo de pesquisa foi desenvolvida em 2009 (ALCANTARA et al., 2009) e se chama *LASP HTLV-1 Subtyping Tool*. Ela está publicada no endereço eletrônico http://www.bioafrica.net/rega-genotype/html/subtypinghtlv.html, foi desenvolvida utilizando a linguagem de programação *Java* em conjunto com *scripts* em *php* e é composta por um *framework* - conjunto de bibliotecas de *software* - e uma interface *web* que permitem ao usuário a submissão de sequências em formato *FASTA*(LIPMAN; PEARSON, 1985) para análise. Este *framework* também deu origem às ferramentas de subtipagem para os vírus HIV-1 (Vírus da Imunodeficiência Humana Adquirida tipo 1), HIV-2 (Vírus da Imunodeficiência Humana Adquirida tipo 2), HBV (Vírus da Hepatite B), HCV (Vírus da Hepatite C), HTLV-1, HHV-8 (Vírus da Herpes Humano tipo 8) e HPV (Vírus do Papiloma Humano).



Fonte: adaptado de http://www.bioafrica.net/rega-genotype/html/subtypinghtlv.html, acesso em 24/11/2016

Figura 4. Primeira versão da ferramenta para subtipagem do HTLV-1

O método utilizado para realizar a subtipagem consiste em comparar as sequências submetidas pelo usuário com um conjunto de sequências referência cuidadosamente selecionadas. No primeiro momento, as sequências são submetidas a uma análise utilizando o software *BLAST*, com a intenção de identificar suas diferentes regiões genômicas. Em seguida, cada uma das sequências do usuário é alinhada ao arquivo de referência, que contém amostras de todos os subtipos do organismo em estudo.

O próximo passo é a construção de uma árvore filogenética utilizando o método de distância Tamura-Nei ou o HKY, utilizando uma taxa de substituição heterogênea entre os sítios (distribuição gama discreta). Neste momento, as sequências são divididas em segmentos menores e são analisadas utilizando uma técnica de janela deslizante, em que 400 nucleotídeos são avaliados contra segmentos do mesmo tamanho, em incrementos de 20 nucleotídeos. Por fim, o método de *Bootstrap* fornece o suporte estatístico para a inclusão de uma sequência em um determinado subtipo.

Para reduzir o risco de classificações equivocadas e de falsos positivos, os valores de corte foram definidos como >70% para as análises de *bootstrap* e >90% para o *bootscaning*, o que tornou possível analisar grandes volumes de dados - até mil sequências por vez - e determinar genótipos conhecidos com um elevado grau de confiança.

3.3 BANCOS DE DADOS BIOLÓGICOS

Através do avanço das tecnologias de obtenção de dados biológicos é notável o aumento no volume de sequências de nucleotídeos e proteínas armazenadas. Deste modo, para armazenar informações relacionadas a diversos domínios do conhecimento e tendo em vista requisitos específicos de cada grupo de pesquisa – volume dos dados, nível de detalhamento, entre outros – surgem diversos bancos de dados, cada qual projetado para atender a suas próprias demandas.

De acordo com Prosdocimi (2002), os bancos de dados biológicos podem ser classificados como primários e secundários. Um banco de dados primário é constituído pela deposição direta de sequências de nucleotídeos, aminoácidos ou estruturas proteicas, sem qualquer processamento ou análise, como por exemplo, o *GenBank do National Center for Biotechnology Information* (NCBI) /*National Institutes of Health (NIH)*, *European Bioinformatics Institute* (EBI) *European Molecular Biology Laboratory* (EMBL) e o *DNA Data Bank of Japan* (DDBJ) que constituem o *International Nucleotide Sequence Database Collaboration* (INSDC). Por secundários, entende-se aqueles que derivam dos primários, ou seja, foram formados usando as informações depositadas nos bancos primários (PROSDOCIMI et al., 2002). Como exemplo, podemos o citar o *HTLV-1 Molecular Epidemiology Database,* construído através das sequências contidas no *GenBank* associadas a informações clínicas, filogenéticas e epidemiológicas obtidas através da mineração dos dados dos artigos citados nas sequências do *GenBank.*

NAVATHE (2005) destaca algumas características especiais relacionadas ao gerenciamento de dados biológicos, dentre as quais:

- Mudanças rápidas e constantes nos esquemas dos bancos de dados obrigam seus projetos a serem flexíveis - facilidade para mudanças - e extensíveis - facilidade para implementação de novas funcionalidades. Em alguns casos, os

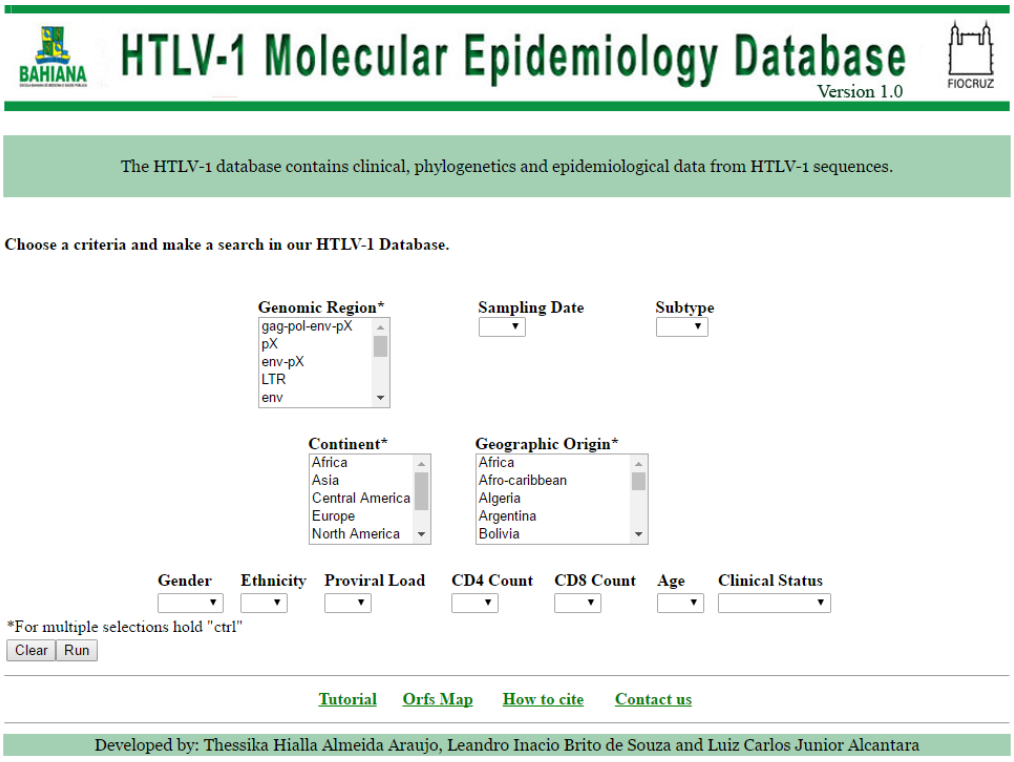
bancos precisam ser refeitos periodicamente na intenção de melhor representar novos conjuntos de dados e hierarquias.

- A maioria dos usuários não necessita de acesso para escrita nos bancos de dados, apenas a realização de consultas é adequada para grande parte dos acessos. Por deixar a escrita reservada a um grupo especial de usuários, chamados curadores, restringe-se a atualização dos dados a este grupo. Também pode-se adicionar erros e inconsistências no banco de dados no processo de atualização e manutenção deste sistema de informação se os curadores não forem profissionais capacitados para realizar estes tipos de ações no banco de dados.
- Grande parte dos biólogos provavelmente não possui conhecimento sobre o projeto do esquema ou sobre a estrutura interna do banco de dados, deste modo, a interface da aplicação do banco de dados deve fornecer as informações dentro de um contexto que seja compreensível para o usuário, ou seja, dentro de sua área de conhecimento, e abstrair requisitos técnicos da base de dados.

Assim, entende-se que um banco de dados biológico deve ao mesmo tempo atender requisitos técnicos, se mantendo flexível e extensível, e permitir o uso - execução de consultas e manutenção de dados - por indivíduos que não conheçam a arquitetura dos seus esquemas e suas estruturas de funcionamento interno.

3.4 HTLV-1 MOLECULAR EPIDEMIOLOGY DATABASE

O *HTLV-1 Molecular Epidemiology Database* é um banco de dados secundário que mantém sequências de nucleotídeos do *GenBank* associadas a informações geográficas, clínicas e epidemiológicas tendo como objetivo principal fornecer subsídios para pesquisas clínicas, relacionadas ao comportamento evolutivo viral, genótipo-fenótipo e desenvolvimento de vacinas. O projeto foi inicialmente desenvolvido por um estudante de mestrado da Fiocruz (ARAUJO et al., 2012) e tem como elemento de destaque a manutenção de informações não indexadas extraídas de outras fontes, como os artigos em que as sequências foram publicadas além de dados obtidos diretamente a partir de seus autores, ou seja, dados que não constam nas anotações das sequências do *GenBank*.



Fonte: Adaptado de http://htlv1db.bahia.fiocruz.br

Figura 5. Página inicial do HTLV-1 Molecular Epidemiology Database

A ferramenta de bioinformática apresenta uma interface *web* que permite aos usuários executarem consultas aos dados e exportarem as informações nos formatos *FASTA* e *csv*. O armazenamento das sequências é feito utilizando o banco de dados *MySQL*, a lógica da aplicação *web* foi desenvolvida utilizando a linguagem de programação *php* e o *layout* e a aparência das páginas foram obtidos através do uso da linguagem de folha de estilo *CSS*. A ferramenta foi desenvolvida no sistema operacional *Linux* está disponível *online* utilizando o servidor *web Apache.*

Na página inicial é apresentado ao usuário um formulário contendo diversos campos de filtro relacionados aos dados das sequências, como subtipo, região genômica, dados clínicos e epidemiológicos, tais como região geográfica, estado clínico, contagem de CD4+ e CD8+, além de informações sobre os autores das publicações relacionadas às sequências. Após a execução da busca, as informações das sequências são apresentadas em uma tabela, permitindo que o usuário visualize o resultado e exporte as sequências.

Esta aplicação é pública, tem acesso gratuito e se encontra publicada no endereço eletrônico http://htlv1db.bahia.fiocruz.br.

3.5 SEGURANÇA DA INFORMAÇÃO

A informação pode ser conceituada como qualquer dado valioso para um indivíduo ou organização. Deste modo, a segurança da informação é caracterizada como conjunto de medidas que visa proteger um sistema de informações da negação de serviço a um usuário autorizado, ao mesmo tempo em que impede a intrusão e modificação desautorizada de dados ou informações (DSIC, 2016).

Quanto ao desenvolvimento de *software*, sua especificação pode ser dividida em requisitos funcionais e não funcionais, sendo os primeiros relacionados diretamente às funcionalidades da aplicação, ou seja, as regras de negócio, e os últimos ligados a aspectos como usabilidade, performance, qualidade, entre outros (SOMMERVILLE, 2011). É justamente nos requisitos não funcionais que se encontram atributos ligados à segurança da informação, delimitando escopo para acesso aos dados, permissões de usuários, tratamento de erros, rotinas de recuperação, requisitos de disponibilidade, entre outros.

Quando pensamos neste conceito sobre a ótica das aplicações, a linguagem de programação utilizada para a implementação de um *software* ou serviço traz intrinsecamente suas próprias características de uso, configuração e boas práticas relacionadas à segurança da informação. A linguagem *php*, por exemplo, pode ser incluída em um servidor *web* como um módulo ou pode ser executada separadamente, como binário *CGI* (linha de comandos através de um terminal ou *scripts*) e permite que um usuário possa acessar arquivos, executar comandos, abrir conexões de rede, entre diversas outras tarefas (Manual do PHP, 1997).

Dadas as diferentes maneiras de utilizar o *php*, existem várias opções de configuração para controlar o seu comportamento e justamente esta flexibilidade permite o uso da linguagem para vários propósitos, mas ao mesmo tempo, também significa que existem combinações dessas opções e configurações do servidor que resultam em uma instalação insegura. Deste modo, o conhecimento sobre os recursos da linguagem aliado à experiência do programador serve como elemento atenuador das possíveis falhas associadas à segurança da informação.

4 MATERIAL E MÉTODOS

O trabalho realizado pode ser dividido em três etapas: a atualização da ferramenta *REGA Subtyping Tool*, o aprimoramento do banco de dados público do *HTLV-1* e a integração destas ferramentas.

4.1 ATUALIZAÇÃO DA FERRAMENTA *REGA SUBTYPING TOOL*

A primeira etapa teve início com a análise do código fonte da ferramenta *Rega Subtyping Tool*. Esta, diferente das duas versões anteriores escritas em php e *Java*, foi desenvolvida utilizando somente a linguagem de programação *Java* e utilizou dois *frameworks* de *software*: O *Rega-Genotype* (https://github.com/rega-cev/rega-genotype), que realiza a reconstrução filogenética através da integração de diversas ferramentas como o *ClustalW*, o *Blast* e o *Paup** aliadas a funcionalidades em *Java*, documentos *XML* e arquivos de configuração, proporcionando uma ferramenta genérica o bastante para realizar a análise de diversos organismos diferentes; e o *framework* de arquitetura de *software Java Web Toolkit*, composto por um conjunto de bibliotecas escritas em linguagem *Java* que visam aumentar a produtividade da construção de aplicações *web* através do uso de classes abstratas que facilitam o emprego das tecnologias *HTLM*, *CSS* e *Javascript* para a construção das páginas da aplicação.

Assim, o *Rega Subtyping Tool* é uma aplicação *web*, ou seja, desenvolvida para permitir seu uso através de uma rede de computadores utilizando a arquitetura cliente/servidor. Basicamente, seu código fonte divide-se em um núcleo (*core*) e uma camada *web*, sendo o primeiro responsável por executar a maior parte do processamento, inclusive encapsulando as chamadas para diversas outras ferramentas que são usadas internamente - através de linha de comandos - durante o processo de análise das sequências submetidas pelo usuário. Pode-se dizer que o processo de análise de uma sequência, neste caso, é o conjunto ordenado de chamadas a aplicações e a coleta e tratamento de seus resultados, tudo isto controlado e parametrizado pelo núcleo da ferramenta de genotipagem.

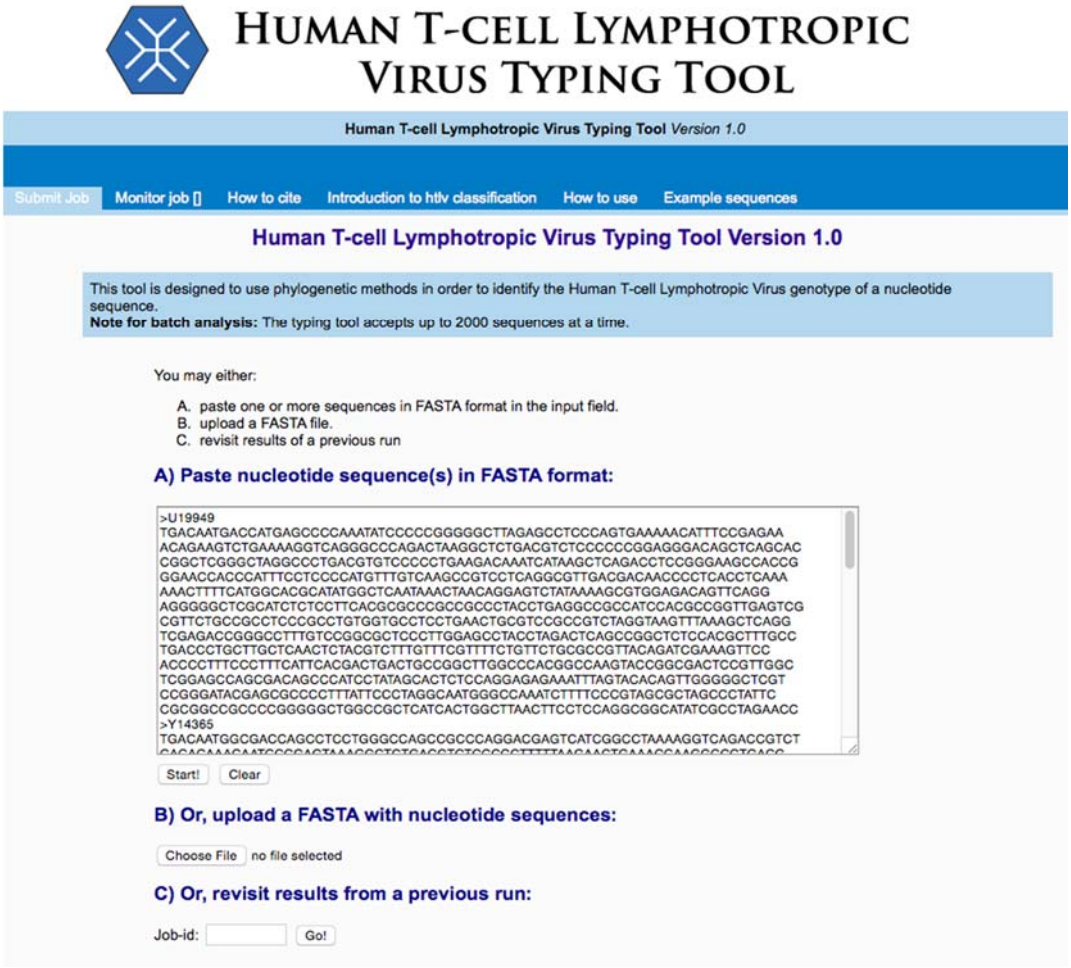
Cada tipo de organismo que a ferramenta é capaz de analisar possui um diretório próprio no qual são armazenados seus parâmetros de análise específicos,

inclusive, um arquivo contendo suas sequências referência de tipos conhecidos destes organismos. A camada *web*, por sua vez, contém dados em arquivos *XML* capazes de manter páginas específicas para os diversos tipos de organismos, o que serve para determinar que rotinas serão chamadas para a análise das sequências submetidas na ferramenta. Como exemplo, pode-se citar que se o usuário acessa a página de genotipagem do *HTLV*, os dados de sequência que este submeter serão enviados para as rotinas de análise específicas deste organismo.

Nossa análise mostrou que a funcionalidade de genotipagem do *HTLV-1* não havia sido migrada para sua versão mais recente desta ferramenta, encontrando-se parcialmente implementada. Uma vez que as funcionalidades referentes ao *HIV* estavam completas e funcionais, este módulo foi utilizado em conjunto com o que havia relacionado ao *HTLV* para realizar a implementação do módulo *HTLV*. Foi elaborado um novo arquivo referência contendo sequências com subtipos conhecidos deste vírus para inclusão na ferramenta, item este indispensável para o processo de genotipagem.

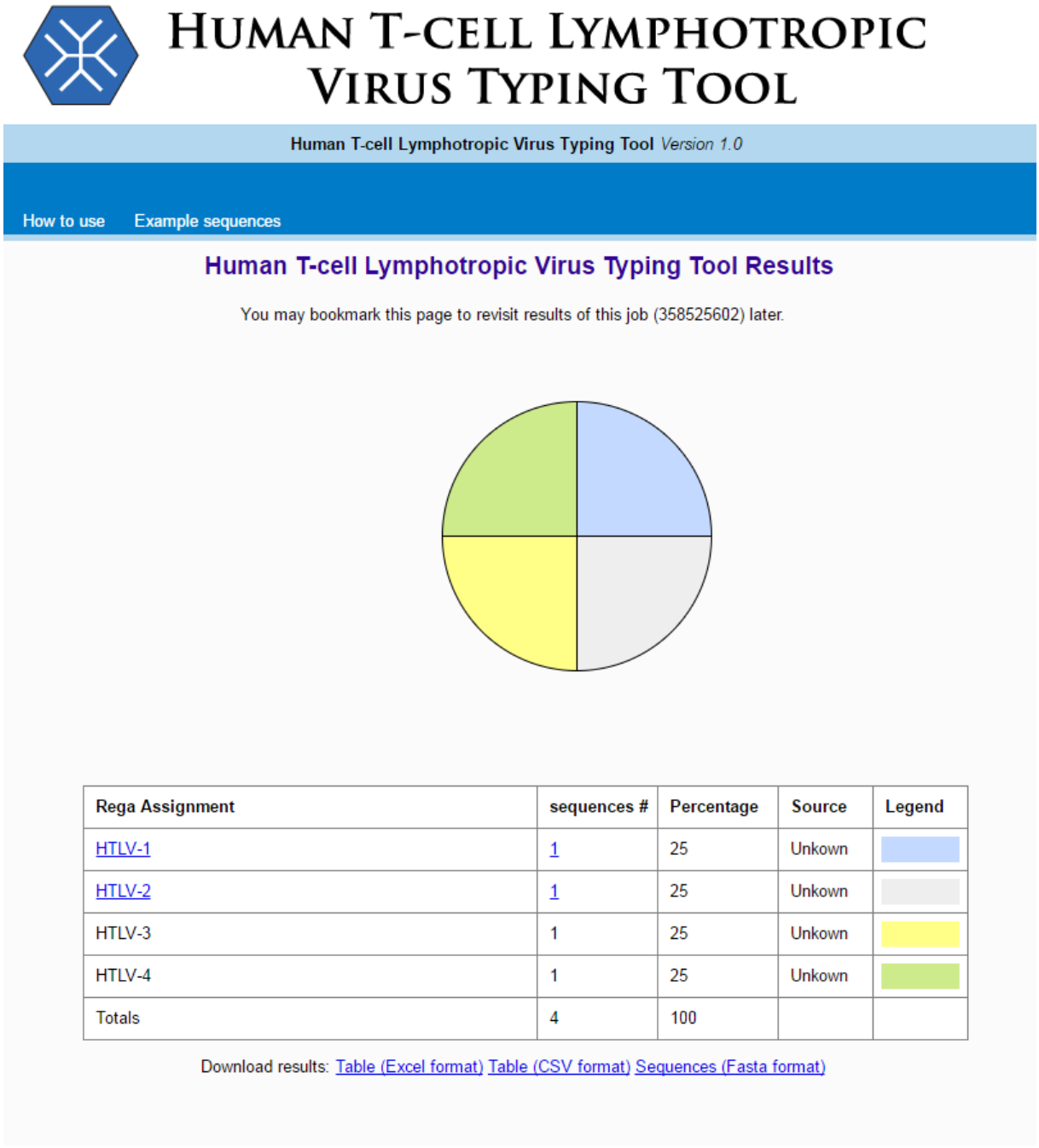
Em um segundo momento, intercedemos em parceria com o grupo de pesquisa da África do Sul, o qual o Dr. Túlio de Oliveira participa como pesquisador e coordenador, para realizar novas modificações na ferramenta. O código fonte foi simplificado, facilitando a inclusão de novos organismos, inclusive o HTLV-2, que teve sequências incluídas no novo arquivo referência, permitindo agora que a mesma ferramenta realize a tipagem do HTLV-1, 2, 3 e 4. A ferramenta foi então renomeada para *Human T-cell Lymphotropic Virus Typing Tool, versão 1.0* (Figura 6).

Assim adicionamos sequências referência dos tipos 1, 2, 3 e 4 do *HTLV*, para permitir que a ferramenta execute a análise filogenética de todos os tipos relacionados ao *HTLV*. O fluxo interno do sistema foi modificado uma vez que estes últimos dois tipos não apresentam subtipos, encerrando-se o processamento após a detecção de seus tipos. Por fim, foram obtidas três ferramentas: uma para realizar a tipagem do *HTLV* e duas outras para genotipar o *HTLV-1* e o *HTLV-2*.



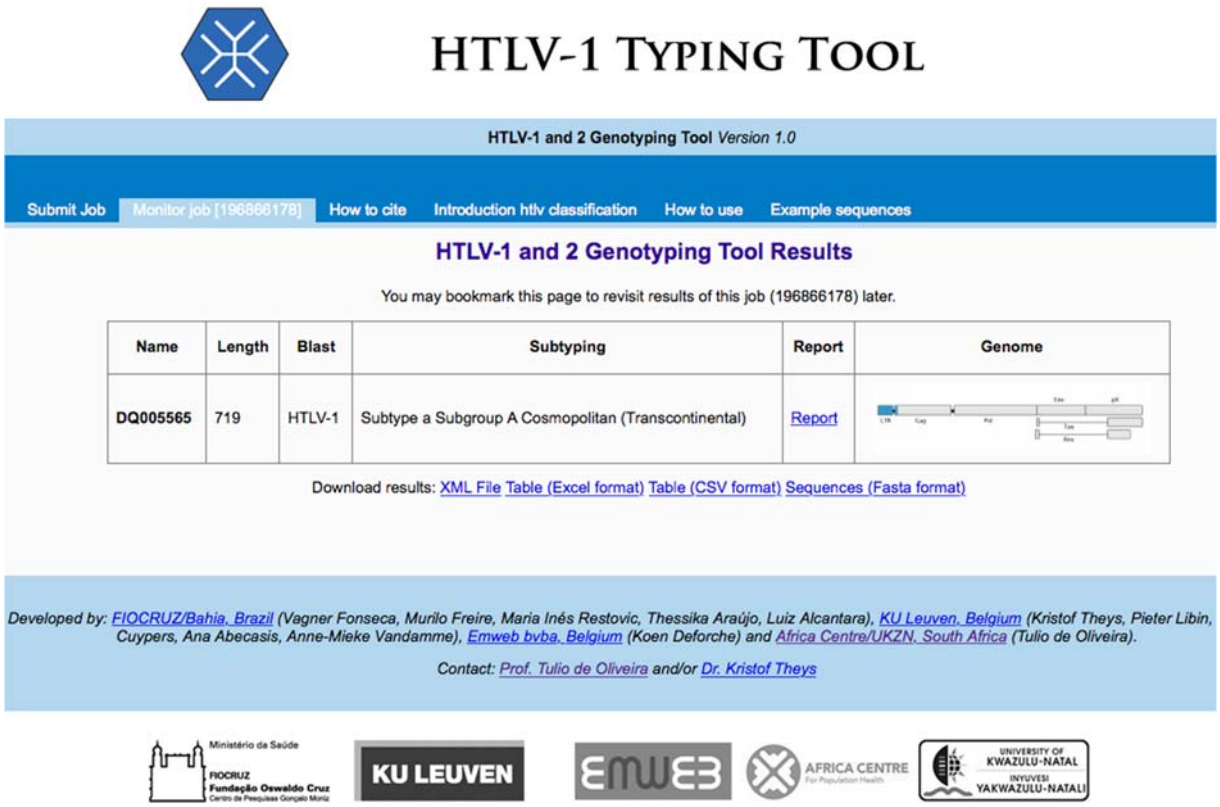
Esta página permite que 2000 sequências de HTLV sejam submetidas por sessão. Fonte: adaptado de http://bioafrica2.mrc.ac.za/rega-genotype/typingtool/htlv

Figura 6. Página inicial da ferramenta para tipagem do HTLV



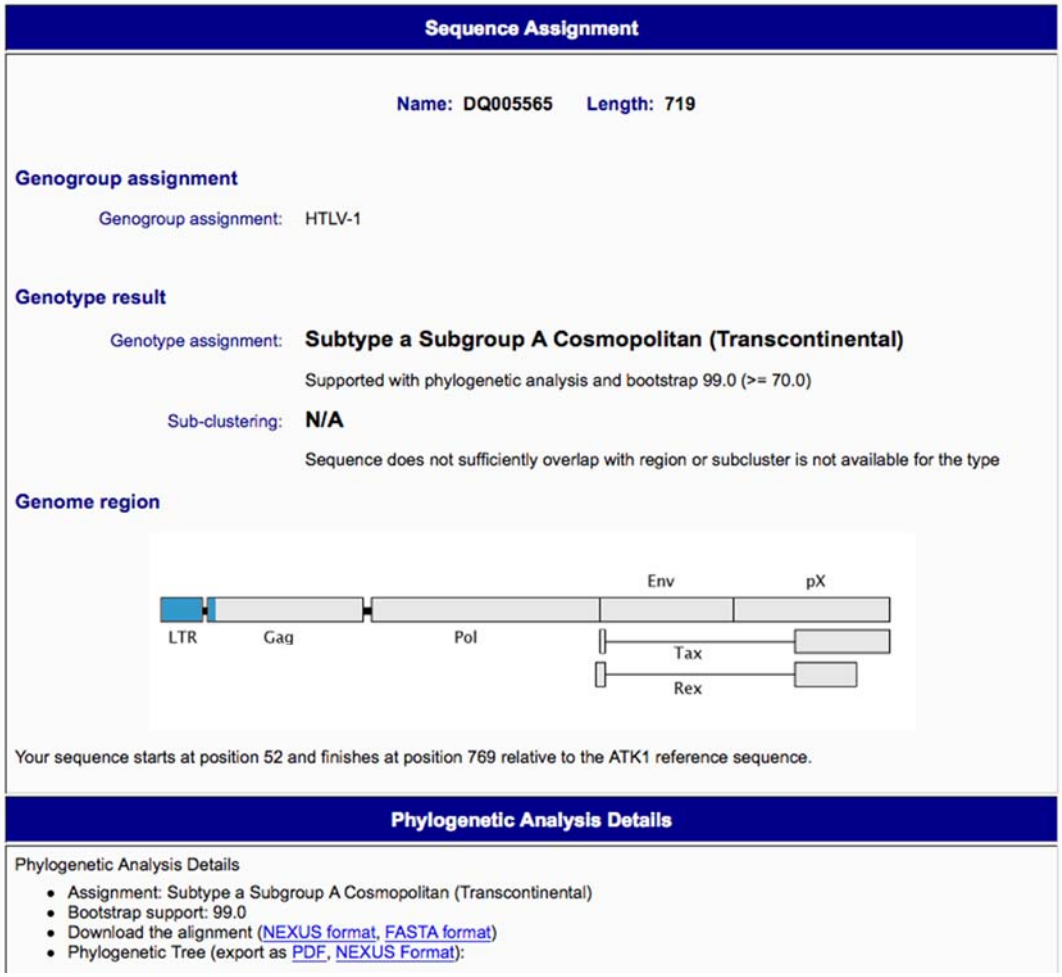
Resultados da identificação de sequências dos quatro tipos do HTLV. Os tipos 1 e 2 apresentam *links* para suas respectivas ferramentas de subtipagem (Figura 8). Fonte: adaptado de http://bioafrica2.mrc.ac.za/rega-genotype/typingtool/htlv

Figura 7. Página de resultados da ferramenta para tipagem do HTLV



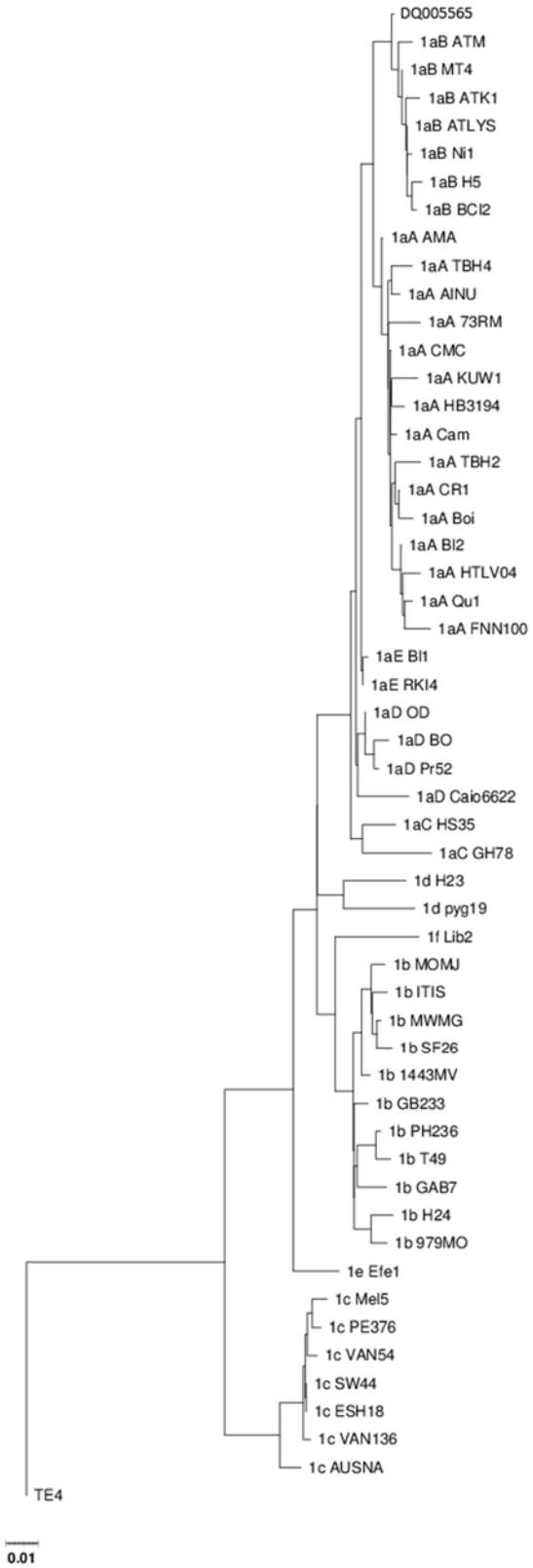
Página de resultados contendo o nome da sequência, o seu tamanho, resultado de subtipagem e posição no genoma. Foi obtida acessando o *link* de resultados detalhados da página mostrada na figura anterior (Figura 7). Fonte: Adaptado de http://bioafrica2.mrc.ac.za/rega-genotype/typingtool/htlv1

Figura 8. Visualização do resultado inicial da genotipagem na ferramenta.



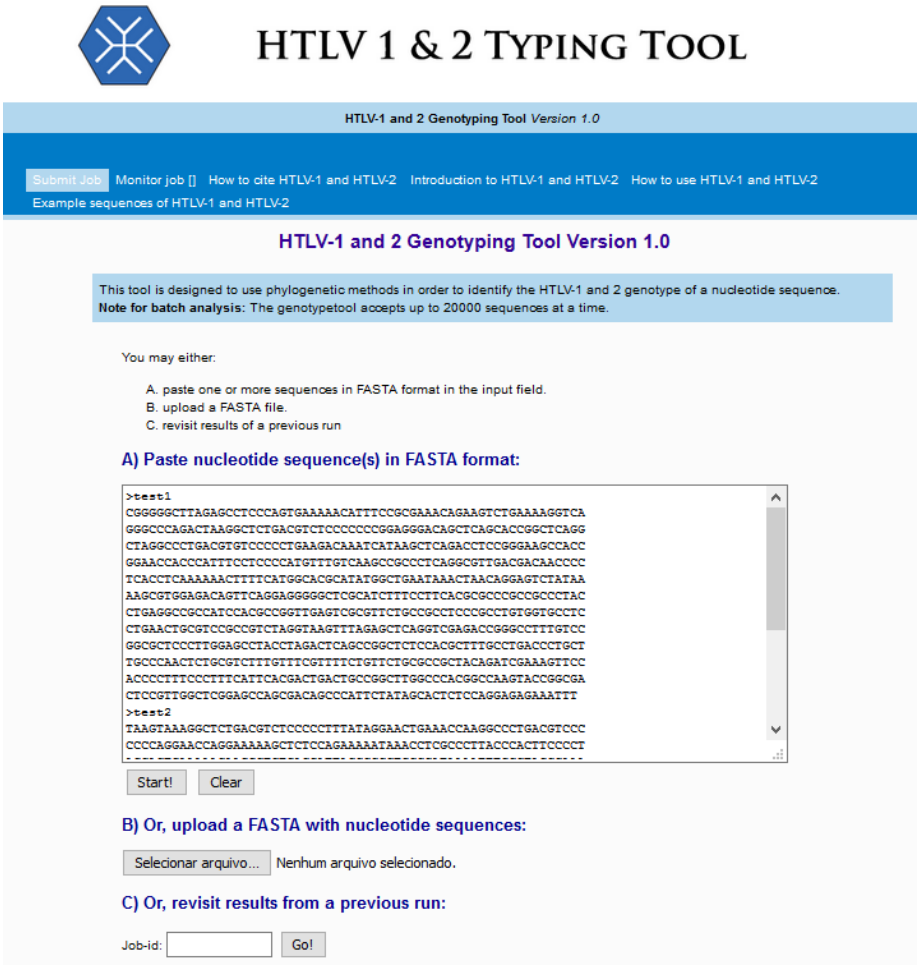
Resultado e detalhes filogenéticos da análise, incluindo suporte de *bootstrap* e árvore construída (Figura 10). Fonte: adaptado de http://bioafrica2.mrc.ac.za/rega-genotype/typingtool/htlv

Figura 9. Página de resultado detalhado da análise de genotipagem



A sequência analisada (DQ005565) foi posicionada no topo da árvore filogenética para melhor interpretação dos resultados. Fonte: adaptado de http://bioafrica2.mrc.ac.za/rega-genotype/typingtool/htlv

Figura 10. Árvore filogenética reconstruída durante a análise da sequência do usuário.



Após a tipagem, a sequência submetida pelo usuário é direcionada para a ferramenta de genotipagem do HTLV-1 ou do HTLV-2. Fonte: Adaptado de http://bioafrica2.mrc.ac.za/rega-genotype/typingtool/htlv1

Figura 11. Formulário de entrada de dados da ferramenta de genotipagem para o HTLV-1 e HTLV-2.

Diversas partes do aplicativo foram otimizadas e foram acrescidas também funcionalidades para permitir acesso aos arquivos gerados nos resultados da genotipagem, através de requisições *web* utilizando a tecnologia *Servlets* do *Java*, o que possibilita que outras aplicações possam obter e utilizar estes resultados para os mais diversos fins. Nestes arquivos estão contidos alinhamentos das sequências dos usuários com as sequências referência, valores de *bootstrap*, reconstruções de árvores filogenéticas, *logs* das ferramentas utilizadas na genotipagem, como o *PAUP**, entre diversas outras informações.

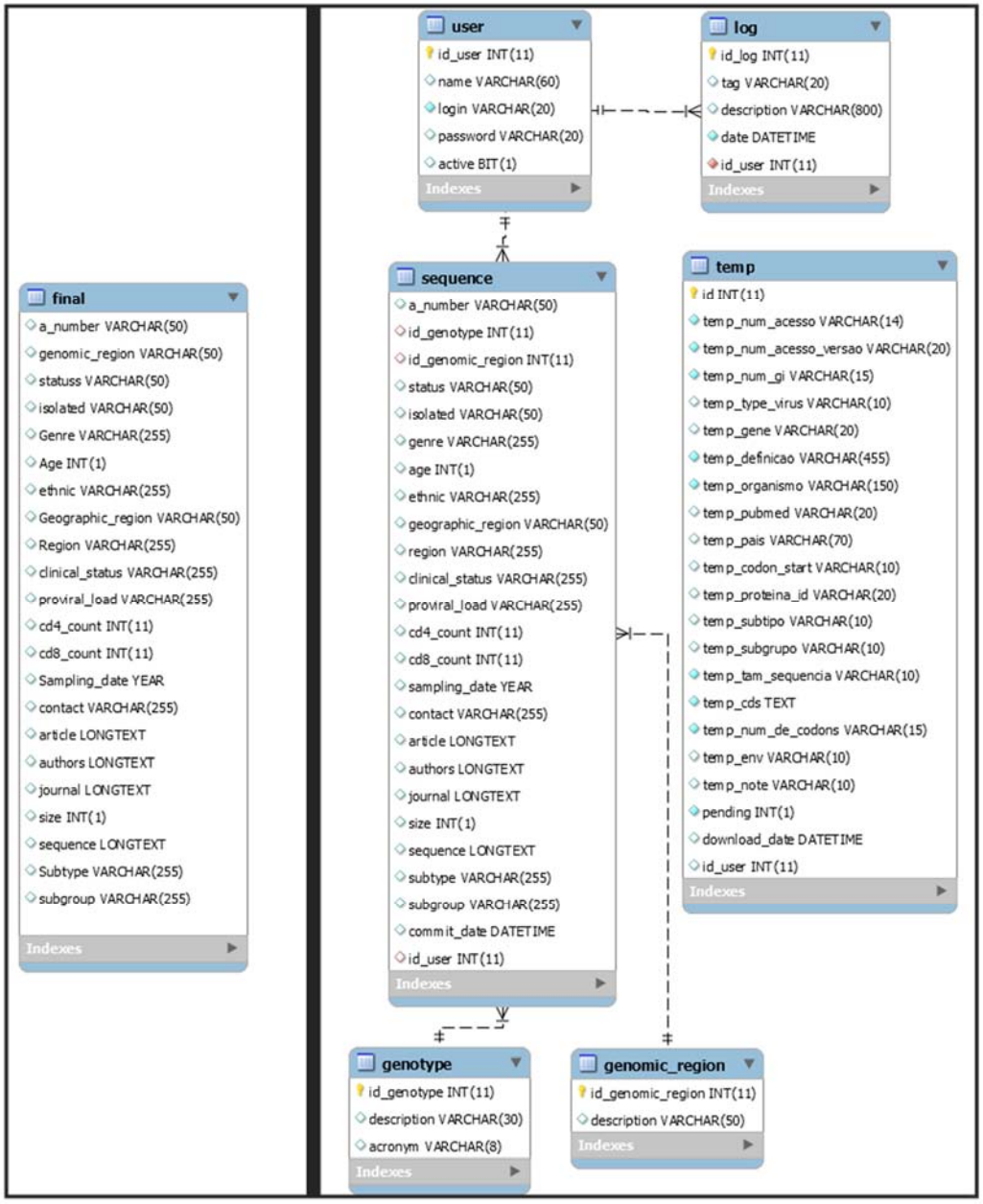
A interface requer três parâmetros: o JOB ID, número este gerado automaticamente, cada vez que um usuário procede uma análise na ferramenta; o JOB DIR, um parâmetro de configuração de versão do aplicativo; e o nome do arquivo

requerido, como por exemplo, result.xml, retornando para o solicitante o conteúdo deste arquivo. Os arquivos *XML* já têm o conteúdo estruturado, dada a sua natureza, entretanto cabe ao solicitante tratar os arquivos de dados não estruturados, como alinhamentos e *logs*, quando da resposta da requisição pelo servidor.

4.2 APRIMORAMENTO DO BANCO DE DADOS PÚBLICO DO *HTLV-1*

A segunda etapa do trabalho envolveu o aprimoramento do *HTLV-1 Molecular Epidemiology Database* e teve início com a obtenção da cópia do banco de dados e do código fonte, desenvolvidos em *MySQL* e *php*, respectivamente. O esquema do banco de dados continha uma única tabela, denominada final, que armazenava os dados compilados sobre o HTLV-1. O primeiro passo foi realizar a manutenção no banco de dados, visando a normalização dos seus dados. Este processo resultou na criação de 3 tabelas: *genotype*, *genomic_region* e *sequence*, esta última sendo a principal tabela do esquema. Em seguida, configuramos as chaves primárias e estrangeiras das tabelas e elaboramos *scripts* em linguagem *SQL* para realizar a migração dos dados para a nova estrutura. Uma vez que a ferramenta de genotipagem passou a contemplar também o HTLV-2, 3 e 4, decidiu-se por incluir estes vírus também no banco de dados, definindo-se assim a tabela *genotype* como diferenciador entre sequências dos tipos de *HTLV*.

Adicionalmente, criou-se as tabelas *user*, para armazenar os dados de acesso dos curadores do banco de dados; *log*, para armazenar os eventos do sistema, ou seja, o histórico de ações dos usuários; e *temp*, para armazenar os dados temporários de *downloads* de sequências a partir do *GenBank*. Ao final, o esquema conteve seis tabelas.



À esquerda, o banco de dados original; em seguida, o resultado após a normalização. Fonte: O autor.

Figura 12. Processo de normalização do banco de dados.

O código fonte da aplicação, em *php*, apresentou uma arquitetura de programação estruturada, ou seja, a lógica da aplicação se encontrava distribuída nos vários arquivos que compunham a aplicação.

A intervenção iniciou-se na confecção de um relatório classificando as pendências e as necessidades de melhoria na ferramenta, que foi estabelecida sobre três perspectivas: Primeira, o desempenho, teve como foco a avaliação das consultas *SQL*, verificando se havia campos não utilizados nos resultados e se era necessária

a criação de índices, um recurso do banco de dados que ordena os dados em seu armazenamento lógico ou físico para aumentar a velocidade das consultas; Na segunda perspectiva avaliou-se a segurança da aplicação, testando vulnerabilidades conhecidas nos *softwares* escritos em *php*, como o uso de *SQL Injection*, alteração de parâmetros das requisições *web* e tratamento de mensagens de erro; Por último, a perspectiva voltada para facilidade de manutenção contemplou a organização e a utilização de padrões no código fonte, visando deixar a aplicação atualizada junto às mudanças na linguagem de programação e mais fácil de modificar por programadores no futuro.

Também foi contemplado o trabalho necessário para adaptar a aplicação ao novo esquema do banco de dados, modificado após a normalização, que contempla a reescrita das consultas *SQL* da aplicação, a mudança nos campos tipo *"combobox"* dos formulários, que carregam dados a partir do banco, reconstrução de rotinas e atividades menores, como pequenas correções de código.

A partir do relatório, foi elaborada uma lista contendo as modificações que deveriam ser executadas na ferramenta, classificadas de acordo com seu impacto, mostradas no Quadro 2.

Quadro 2. Resultado a análise do código fonte do banco de dados público

(continua)

	Procedimento	**Guia**
1	Substituir as short tags do php (<? por <?php).	Facilidade de manutenção
2	Substituir a conexão ao banco de dados pelo *PDO.*	Desempenho, segurança
3	Padronizar o *layout*, colocando a aplicação sobre um único template, contendo cabeçalho, rodapé e as principais bibliotecas. A partir deste modelo (*template)*, as páginas da aplicação serão carregadas.	Facilidade de manutenção
4	Aplicar o padrão *tableless* no *layout.*	Facilidade de manutenção, acessibilidade
5	Agrupar funções comuns do *php.*	Facilidade de manutenção
6	Resolver as funções depreciadas do *php.*	Segurança
7	Adaptar as consultas *SQL* para o banco de dados normalizado.	Desempenho
8	Os campos do tipo "*combobox*" devem ser carregados automaticamente a partir do banco de dados.	Facilidade de manutenção

(conclusão)

	Procedimento	**Guia**
9	Criar os índices no *MySQL* para campos chave de busca.	Desempenho
10	Desenvolver as funcionalidades de manutenção de usuários do sistema.	Segurança
11	Implementar nas páginas que tenham acesso restrito rotinas de restrição de acesso	Segurança
12	Implementar as funções de higienização e manutenção de dados obtidos através de *download* automático	Segurança, Desempenho
13	Implementar as funções de auditoria dos dados (*log* de ações dos usuários do sistema)	Segurança

Fonte: O autor.

Alguns dos itens, como a substituição das *short tags* do *php*, permitiram que as correções fossem executadas em lotes, através da execução de comandos sobre todo o código fonte de forma automática, entretanto as modificações mais complexas, como a substituição do método de conexão ao banco de dados pelo padrão *PDO*, demandaram que várias partes do código fossem reescritas. Neste caso, uma classe chamada Conexao.class.php é responsável por executar e enviar os resultados das consultas, recebendo como parâmetros informações sobre tabelas, campos e condições de busca e ordenamento, mantendo assim o código fonte mais próximo do paradigma da orientação a objetos, utilizando o conceito de encapsulamento na execução das consultas à base de dados. Outros itens, como os de número 3 e 4 da Quadro 2, foram resolvidos aplicando o *framework Bootstrap*, um conjunto de bibliotecas e rotinas de *HTML*, *CSS* e *Javascript* que contempla tecnologias modernas de construção de páginas *web*, como o padrão *tableless*.

Alguns dos itens da lista foram agrupados em uma seção administrativa da ferramenta, com acesso restrito, e que permite aos curadores efetuar ações além da busca de informações de sequências. Basicamente, esta seção é constituída pelas funcionalidades de *download* de sequências, visualizador de *logs* de eventos e cadastro de usuários. Na primeira, a ferramenta possibilita uma interface para que o usuário obtenha sequências diretamente no *GenBank*, diferente do método anterior em que as sequências eram incluídas manualmente no banco de dados. Desta maneira, quando as sequências são obtidas, estas ficam marcadas como pendentes, sendo necessária a inspeção dos dados pelo administrador da aplicação. Se o

administrador constatar que os dados estão corretos, ele acessa a funcionalidade de gravação definitiva dos dados no sistema (*commit*) e estes passam a estar disponíveis para consulta pública; senão, ele marca os dados como "não conformidade": neste momento, o sistema altera as informações para que os dados não sejam mais apresentados como pendentes, porém os mantém ocultos no banco de dados, com a intenção de evitar o *download* futuro destes mesmos dados.

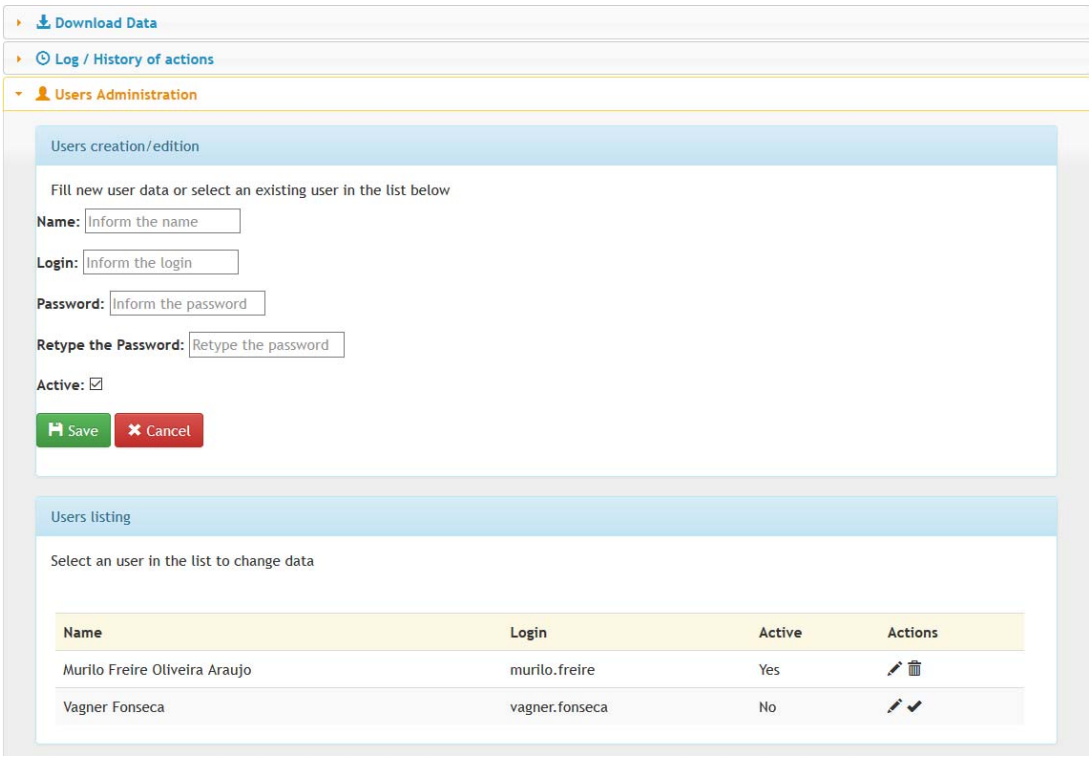
O visualizador do *log* de eventos é uma parte da funcionalidade de auditoria dos dados, que proporciona o controle por parte dos curadores sobre as ações executadas no sistema, como *download* de dados, acesso à área administrativa da ferramenta, edição de sequências, entre outras. A cada ocorrência de um destes eventos, é gerada uma estrutura de dados contendo o usuário relacionado, a data e hora do evento, uma *tag*, que consiste em um texto curto identificador daquele tipo de evento (por exemplo, "LOGIN"), e dados específicos sobre o evento, como campos alterados ou consultas realizadas. Em cada parte do código fonte em que estas ações precisam ser registradas é inserido um código como o exemplo a seguir, que se refere ao login no sistema:

"registerLog(CONST_LOG_LOGIN, "User id: " . $id_user . ", Login: " . $login);" Neste exemplo, é chamada a função registerLog e são passados três parâmetros: a constante CONST_LOG_LOGIN, que define o evento como um acesso ao sistema; a variável $id_user, que contém o identificador único de cada usuário no banco de dados, obtido quando efetuado um acesso válido à aplicação; e a variável $login, que contém o nome que identifica o usuário no sistema, relacionada ao identificador anterior. Esta função ainda permite mais um parâmetro, um tipo texto que pode ser utilizado para registrar alguma informação específica relacionada ao evento.

O visualizador do *log* permite que os curadores acompanhem o histórico de ações através da visualização destes registros.

Quanto ao cadastro de usuários, esta funcionalidade permite inserir os administradores da ferramenta, ou seja, os usuários que terão acesso à parte com acesso restrito desta. O cadastro é feito através de um formulário em que constam campos referentes ao nome, *login*, senha e ativação (o usuário pode estar cadastrado, mas se não estiver ativo, não pode acessar o sistema). Também é disponibilizado o mecanismo de mudança de senha para estes usuários. Na tela, é mostrado ao usuário um formulário contendo os campos do cadastro de usuários e abaixo uma lista de

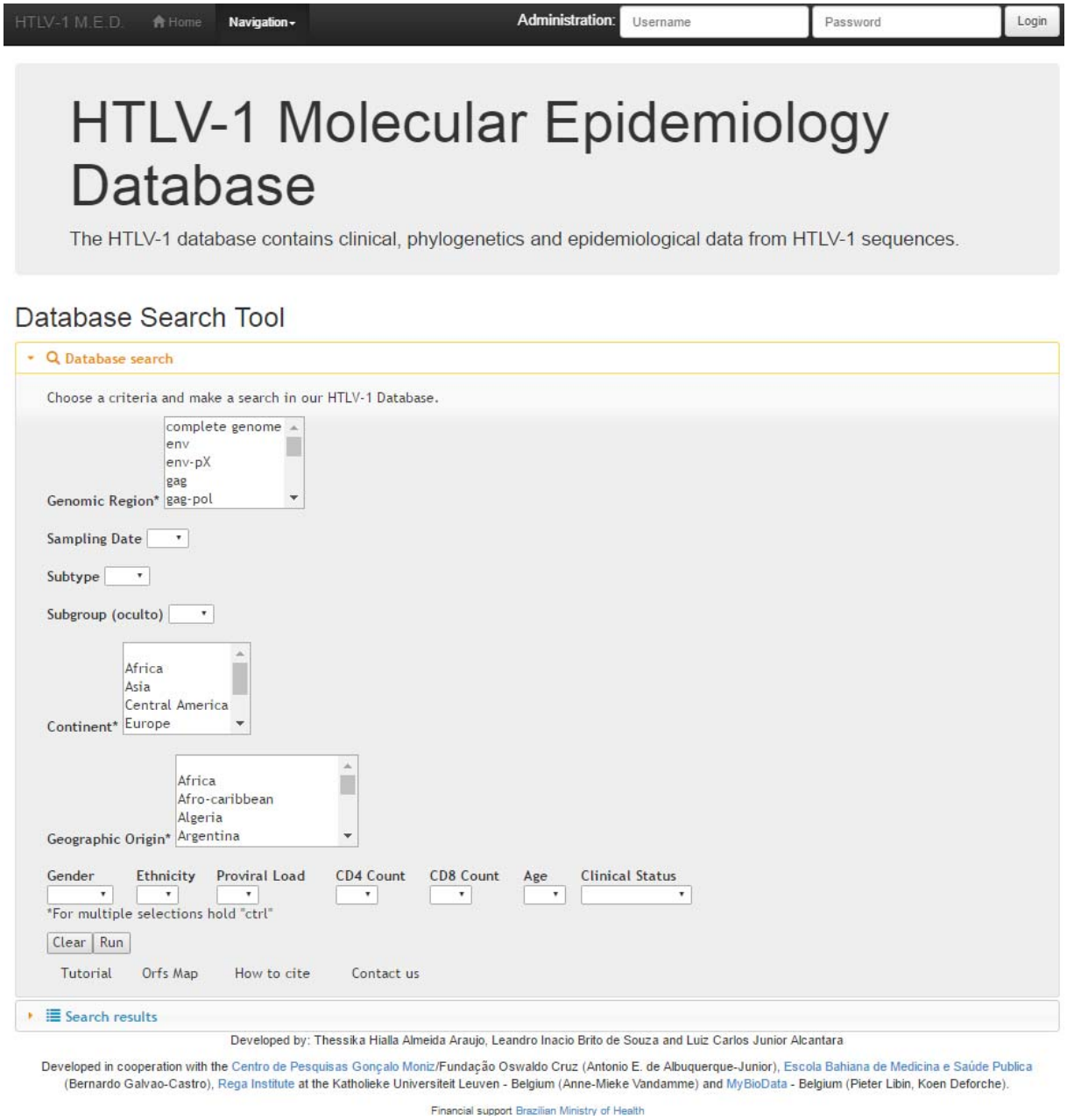
todos os usuários cadastrados, o que possibilita executar as funções de edição e exclusão lógica de usuários, conforme mostrado na Figura 13.



Fonte: O autor.

Figura 13. Tela para cadastro e edição de dados de usuários do Banco de Dados

A interface também permite que o usuário inclua ou altere informações das sequências, tarefa que antes era executada diretamente no banco de dados, através da edição diretamente na tabela do esquema. Com essa funcionalidade, um usuário pode incluir, atualizar ou corrigir as informações das sequências pela própria ferramenta, reduzindo assim o risco de isto ocasionar inconsistências no banco de dados.



Fonte: O autor.

Figura 14. Formulário de busca após a evolução do Banco de Dados do HTLV.

4.3 INTEGRAÇÃO DAS FERRAMENTAS

Por último, a terceira etapa do projeto abordou a integração entre a ferramenta de genotipagem e o banco de dados público do HTLV. Observado que estas ferramentas foram desenvolvidas em linguagens de programação diferentes – *Java* e *php*, respectivamente – e estão hospedadas em servidores distintos, a integração não poderia ser realizada pelo código fonte. Para solucionar este problema, optou-se por

usar um padrão de comunicação que fosse comum para as duas linguagens, ou seja, que permitisse o envio de mensagens de uma aplicação para a outra. Assim, optou-se pelo uso do padrão de intercâmbio de dados *JSON* (http://www.json.org/), associado ao envio de mensagens pelo protocolo *HTTP* utilizando requisições *web* sobre os métodos *GET* e *POST*. Desde modo, pode-se criar um fluxo de informações entre as ferramentas ainda que estas estejam em servidores distintos.

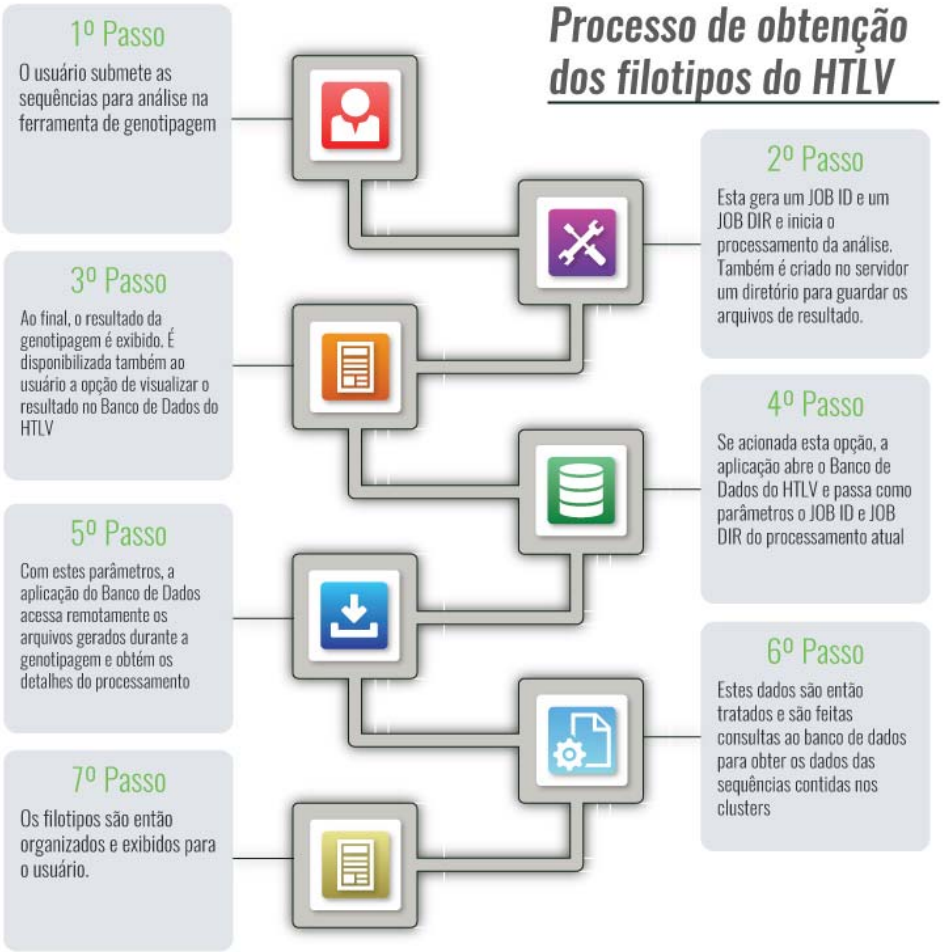
Para construir a interface de comunicação, foi usado um recurso que a ferramenta de genotipagem já dispunha: o JOB ID, um número único gerado a cada análise e que permite ao usuário revisitar um resultado de análise passado, além de servir também, internamente no servidor, para referenciar um diretório no qual são armazenados os arquivos gerados durante o processo de genotipagem, como alinhamentos, *logs,* árvores de reconstrução filogenéticas, entre outros. Entretanto, não é possível ter acesso direto remotamente a estes arquivos uma vez que nenhum destes diretórios é acessível fora do servidor em que a aplicação está sendo executada. Desta maneira, estabelecemos um *Servlet* que realiza esta função de disponibilizar os arquivos mediante uma requisição *web*, através de uma url como a do exemplo a seguir:

"http://bioafrica2.mrc.ac.za/regagenotypealpha/get_job?job_id=943225394&file=result.xml", em que o parâmetro "job_id" se refere ao processamento de número 943225394 e o parâmetro "file" referencia o arquivo de resultado result.xml.

Do ponto de vista do Banco de Dados do *HTLV*, no momento em que o usuário seleciona a funcionaliade "*Analyse in HTLV Molecular Epidemiology Database*" na ferramenta de genotipagem, esta abre o banco de dados em uma nova janela do navegador, chamando o script PHP genotype_result_reader.php e passando como parâmetro o JOB ID do processamento atual. Esta página por sua vez, executa determinadas requisições web para obter os dados necessários para montar os filotipos. A primeira destas, requisita o arquivo result.xml e para isto, considerando que este está estruturado em XML, é utilizada a função nativa do PHP simplexml_load_file, que retorna o conteúdo do arquivo como grupos de vetores aninhados. Esta estrutura de dados contém diversas informações importantes, como o genótipo, o subtipo, o subgrupo e o nome dos arquivos de log e alinhamento de cada uma das sequências processadas.

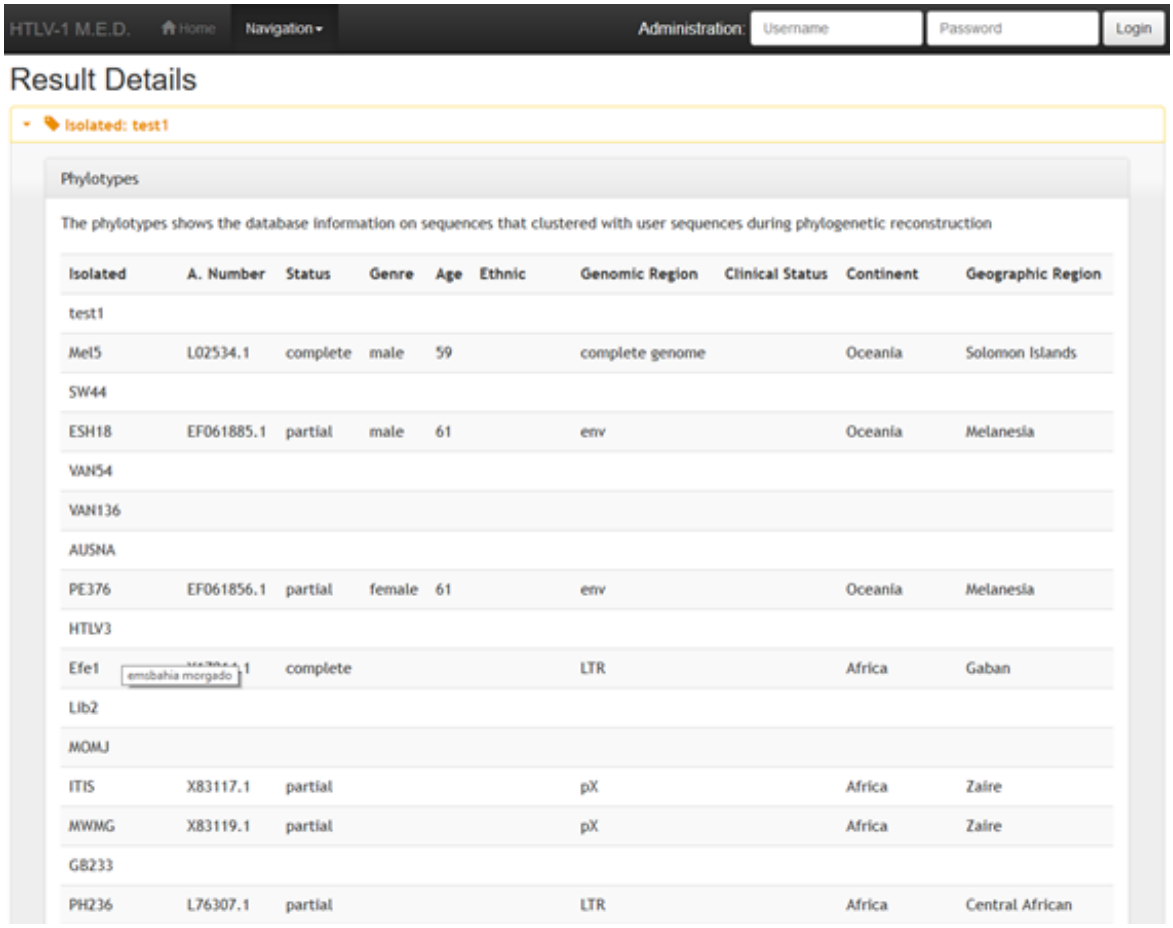
A primeira corrida por este arquivo monta uma tabela contendo somente as sequências submetidas pelo usuário para a genotipagem. Em seguida, no

detalhamento, para cada uma destas sequências será efetuada uma nova requisição a fim de se obter seu arquivo de *log* do PAUP*, cujo nome se encontra no campo "phylo-minor". Este log não é estruturado logo é necessário realizar operações de tratamento de *strings* para se obter os nomes das sequências que se agruparam quando da formação do *cluster*. A partir deste ponto, para cada sequência do *cluster*, é feita uma consulta no banco de dados: se a sequência for localizada, seus dados são exibidos; se não, apenas o nome do isolado é mostrado permanecendo os demais campos em branco. Ao final, também é exibido o arquivo de *log* na íntegra para o usuário. Este processo é repetido para todas as sequências constantes no arquivo XML.



Fonte: o autor, 2016.

Figura 15. Fluxo de integração das ferramentas de tipagem e subtipagem com o banco de dados do HTLV.



A primeira sequência da tabela é a que foi submetida pelo usuário e as demais são os resultados da reconstrução filogenética (*cluster*) associadas aos dados contidos no Banco de Dados do HTLV.
Fonte: O autor.

Figura 16. Resultado da integração das ferramentas, demonstrando as sequências anotadas pela região, continente e região geográfica.

Para averiguar a conformidade da ferramenta, duas funcionalidades foram auditadas: a obtenção automática de sequências a partir do *GenBank* e a reconstrução dos filotipos na ferramenta de banco de dados.

4.3.1 Auditoria da obtenção automática de dados

O procedimento de teste teve início com a obtenção das sequências do HTLV-1 e 2 a partir de 2009, uma vez que a ferramenta já continha as sequências anteriores. Para isto, foram importados dados a partir do *GenBank* utilizando o mecanismo de obtenção automática, mantendo-se estas novas sequências como pendentes, o que serviu como medida de controle destes dados. Em seguida, realizou-se uma consulta

semelhante no próprio *GenBank*, a fim de se averiguar se houve diferença na obtenção das sequências. O resultado do teste mostrou que a ferramenta foi capaz de obter os mesmos dados que uma busca realizada diretamente no *GenBank* (através de sua interface *web*). Foram importadas 3884 sequências do HTLV-1, 763 do HTLV-2, 8 do HTLV-3 e 4 do HTLV-4 para o banco de dados.

O próximo passo foi avaliar se todas as sequências constantes no arquivo referência da ferramenta de genotipagem se encontravam no banco de dados do *HTLV-1*. Esta análise foi efetuada comparando-se os campos *accession number* e *isolated* do banco de dados com o arquivo referência da ferramenta de genotipagem e utilizando o *GenBank* para eventuais dúvidas, uma vez que o número do isolado pode não ser único. Esta medida foi realizada para assegurar que todos os táxons das reconstruções filogenéticas, pertencentes ao arquivo referência, estarão presentes nas consultas do banco de dados quando da recuperação de dados do banco de dados e construção dos filotipos.

Ao final, para avaliar os resultados, foram sorteadas sequências do banco de dados e estas foram submetidas à ferramenta de genotipagem, para construção dos filotipos. Após o processamento, já no banco de dados, os resultados foram avaliados mais uma vez utilizando os dados do *GenBank* (através de consultas manuais) e verificando as informações diretamente no banco de dados MySQL.

5 RESULTADOS

5.1 ATUALIZAÇÃO DA FERRAMENTA DE GENOTIPAGEM *REGA SUBTYPING TOOL*

Como resultado da atualização da ferramenta *REGA Subtyping Tool* obtivemos a funcionalidade para genotipar o HTLV-1, disponível somente nas versões anteriores desta ferramenta e agora migrada para a nova versão. Uma segunda ferramenta foi configurada para realizar a genotipagem e a subtipagem do HTLV-2, e, por último, uma terceira ferramenta foi criada para realizar a tipagem dos tipos 3 e 4 do HTLV. O núcleo do *framework REGA GENOTYPE* também foi modificado visando ganhos em performance e a simplificação do processo de inclusão de novos vírus no futuro. Adicionalmente, foi criado um mecanismo através do qual é possível que outras aplicações acessem os resultados da genotipagem através de requisições *web* do tipo *POST* e assim possam visualizar com mais detalhes os arquivos resultantes do processamento das análises.

Este acesso utiliza uma tecnologia nativa do *Java*, os *Servlets*, e provê uma interface padronizada para este tipo de requisição, que é capaz de responder tanto com dados estruturados (em *XML*, por exemplo) quanto não estruturados (como arquivos textos de *log*). Por fim, a ferramenta foi renomeada para *HTLV-1 & 2 Genotype Tool*, e sua versão foi estabelecida em 1.0 (Figura 11). Em sua versão atual, esta foi novamente renomeada para *Human T-cell Lymphotropic Virus Typing Tool Version 1.0* (Figura 6)

5.2 APRIMORAMENTO DO *HTLV-1 MOLECULAR EPIDEMIOLOGY DATABASE*

Como resultados acerca do aprimoramento do banco de dados público do *HTLV-1*, podemos citar inicialmente a manutenção em seu esquema de banco de dados e arquivos de código fonte. Disto, resultou um banco de dados mais robusto, estruturado e que atende mais plenamente às formas normais além de uma aplicação que utiliza padrões mais modernos de construção de páginas *web*, como *CSS* e *Tableless,* além da reformulação na maneira como o conteúdo (*layout* e

funcionalidades) é distribuído pelas páginas, ficando itens comuns de *layout*, como cabeçalhos e rodapés, padronizados em um só arquivo. Assim, obteve-se uma estrutura mais leve, organizada e fácil de manter no futuro.

Aliado a isto, a ferramenta passou a contemplar também os dados de sequências do *HTLV-2*, *3* e *4*. Isto possibilita que os recursos computacionais – como a busca, o *download* automático e a edição dos dados das sequências – utilizados para o estudo do *HTLV-1* também possam ser empregados na pesquisa dos demais tipos deste vírus, agregando novas ferramentas para o campo de estudo do HTLV. Em uma versão futura, o banco de dados será renomeado para *HTLV Molecular Epidemiology Database version 2*.

Do ponto de vista do usuário final, este obteve uma interface mais organizada e capaz de se ajustar melhor a diferentes resoluções de tela, recurso este útil quando do acesso por dispositivos móveis. As funcionalidades principais da ferramenta, entretanto, como busca e *download* de sequências em formato fasta, foram mantidas em posições próximas às da versão anterior, para a conveniência dos usuários frequentes da aplicação que já estão acostumados com a distribuição das funcionalidades pelo *layout*.

Em relação à manutenção da aplicação, a adoção do padrão *PDO* possibilitou que se aplicasse o encapsulamento nas consultas ao banco de dados, reduzindo a complexidade em toda página da aplicação que utiliza este recurso. Além disto, o agrupamento das funcionalidades repetidas em funções permitiu que a manutenção de algumas das funções seja realizada apenas em uma parte do código e se reflita para toda a aplicação. Por último, a implantação da tecnologia *AJAX*[1] reduziu a quantidade de dados enviados através das requisições *web*, melhorando assim o desempenho de um modo geral da aplicação.

Os administradores da ferramenta, por sua vez, passaram a dispor de mecanismos de auxílio para a manutenção dos dados. Os dados podem ser obtidos automaticamente a partir do *GenBank* e passam pela análise dos curadores do banco de dados antes de serem disponibilizados publicamente. Além disto, a edição das informações é feita na própria ferramenta e tornou-se possível acompanhar esse fluxo de ações através dos registros de *log.*

[1] *Asynchronous Javascript and XML*, tecnologia capaz de realizar uma requisição *web* assíncrona, ou seja, obtendo novos dados do servidor ao mesmo tempo em que mantém as demais funcionalidades da aplicação disponíveis para o usuário.

5.3 INTEGRAÇÃO DAS FERRAMENTAS

A troca de informações entre as ferramentas foi implementada. O *Human T-cell Lymphotropic Virus Typing Tool* foi dotado de um mecanismo que disponibiliza o conteúdo dos arquivos resultantes das análises – tipagem e subtipagem - através de respostas a requisições *web*, permitindo assim acesso a estes dados por outras aplicações. No *HTLV-1 Molecular Epidemiology Database* foi implementada uma funcionalidade capaz de executar estas requisições *web*, receber as respostas e tratar os dados recebidos relacionando estes com o conteúdo presente em seu próprio banco de dados. Desta maneira, são construídos os filotipos através do relacionamento entre as informações obtidas acerca da reconstrução filogenética e das informações complementares destas sequências recuperadas do banco de dados. Ao final, os filotipos são organizados em uma tabela para cada sequência submetida pelo usuário e são exibidos em uma tela específica na interface do banco de dados. No total, 3884 sequências do HTLV-1, 763 do HTLV-2, 8 do HTLV-3 e 4 do HTLV-4 foram analisadas pelas ferramentas e passaram a fazer parte do banco de dados.

6 DISCUSSÃO

A manutenção, aprimoramento e integração das ferramentas expostas neste trabalho demandou que algumas decisões de projeto de *software* fossem tomadas, tendo em vista os mais diversos aspectos do desenvolvimento de aplicações. Algumas situações apresentavam mais de uma solução, cada qual apresentando suas vantagens e desvantagens em relação aos requisitos funcionais e não funcionais do projeto.

A análise das tabelas do banco de dados público do *HTLV-1* revelou que a normalização de alguns dos atributos traria benefícios à ferramenta ao mesmo tempo em que a aplicação das formas normais a outros poderia prejudicar requisitos funcionais, como o tratamento dos dados obtidos por *download* automático a partir do *Genbank*.

Em relação à primeira forma normal, há dois atributos multivalorados, *authors* e *contact*, que armazenam informações dos autores dos artigos em que as sequências foram obtidas e seus contatos, respectivamente. Estes campos permaneceram inalterados, pois têm função acessória na aplicação, ou seja, não são critérios de pesquisa e a mudança para a primeira forma normal acarretaria uma carga de trabalho adicional por parte do usuário que efetuar o *download* de dados, tendo que tratá-los para adequação à nova arquitetura. Como não havia chaves compostas no banco de dados, este se encontrava automaticamente na segunda forma normal.

Por último, para realizar a adequação à terceira forma normal, foram modificados alguns campos que participavam diretamente nos mecanismos de filtro e busca da ferramenta, o que resultou na divisão da única tabela que constava no banco de dados – nomeada “final” – em três outras: genomic_region, genotype e sequence. Além destas, outras três foram criadas: uma para armazenar os administradores do sistema, outra para armazenar os dados temporários de *download* do *Genbank* e a última para registro das ações dos usuários (*logs*), denominadas *user*, *temp* e *log* respectivamente.

Um ganho que deve ser mencionado é a criação da tabela *genotype*, que diferencia os genótipos do HTLV, permitindo assim que sejam adicionados dados de qualquer genótipo deste vírus, servindo ainda como elemento de filtro nas consultas através do formulário de busca da ferramenta.

A utilização de uma tabela em contrapartida à utilização de um campo (atributo) na tabela *sequence* contribui para a normalização do banco de dados além de reduzir o risco de erros de digitação, o que poderia tornar sequências não acessíveis através da interface de pesquisa deste aplicativo. Também é preciso mencionar que a decisão por esta configuração permite que todas as funcionalidades já existentes do Banco de Dados Público do HTLV – *download* de dados, edição de informações de sequências, *logs* de eventos – possam também ser utilizados para os demais genótipos do HTLV.

id_genotype	description	acronym
1	Human T Lymphotropic Type 1	HTLV-1
2	Human T Lymphotropic Type 2	HTLV-2
3	Human T Lymphotropic Type 3	HTLV-3
4	Human T Lymphotropic Type 4	HTLV-4

Fonte: O Autor.

Figura 157. Estrutura e dados da tabela entitulada *genotype.*

Em relação à manutenção no código fonte do banco de dados do *HTLV-1*, foi necessário atualizar a tecnologia de conexão ao banco de dados pelo *php*, pois a mesma está marcada como depreciada na documentação da linguagem e será removida na versão 7 do *php* (“Manual do PHP”, 1997). Havia duas opções disponíveis: o *mysqli* e o *PDO*. Apesar do *mysqli* apresentar uma versão melhorada da tecnologia até então utilizada (o “*i*” vem de “*improved*”, melhorada), sendo necessárias neste caso muito menos intervenções no código para realizar a migração, o padrão *PDO* demonstra nativamente uma semelhança maior com o paradigma da orientação a objetos, conforme observado em (Manual do PHP, 1997). Deste modo, optou-se por realizar uma intervenção maior na aplicação visando usufruir dos benefícios que a orientação a objetos pode proporcionar, como melhorias na arquitetura, organização e manutenção do código, uma vez que esta abordagem – orientada a objetos – é mais atual e se encontra presente em grande parte das modernas linguagens de programação. Em todo caso, salientamos que a decisão por uma tecnologia também é relacionada com aspectos subjetivos, ficando assim, parte dela a critério da preferência da equipe de desenvolvimento, que optou por tornar o

código fonte mais próximo de padrões mais modernos de desenvolvimento de *software*.

Quadro 3. Funções depreciadas do PHP encontradas no *HTLV-1 Molecular Epidemiology Database*.

Função depreciada	Descrição	Correção
mysql_connect	Executa a conexão ao banco de dados Mysql	O manual sugere a substituição por mysqli_connect ou por pdo_mysql. Optou-se pelo PDO por ter uma implementação mais próxima do paradigma da orientação a objetos.
Split	Dada uma string e um caracter separador, divide esta *string* cortando nas ocorrências do caracter separador, retornando um vetor como resultado.	Substituído por explode.
mysql_query	Executa uma consulta ao banco de dados e retorna um objeto do tipo *Resultset*	Substituído pelo PDO
mysql_fetch_array	Converte um objeto do tipo *Resultset* para um vetor chave/valor	Devido a substituição do mecanismo de execução de consultas pelo PDO, o *Resultset* não precisa de conversão de formato, tornando esse comando desnecessário.

Estas funções foram marcadas como depreciadas na versão 5 do PHP e serão removidas na versão 7. Fonte: O Autor.

Em relação à funcionalidade de *log* do sistema, implementamos rotinas genéricas que podem ser incluídas em pontos-chave da aplicação para realizar o registro dos eventos, como acesso ao sistema, inclusão ou edição dos dados, *download*, entre outros. Desta maneira, apenas uma linha de código é parametrizada e inserida nestes pontos, causando um impacto mínimo no entendimento do fluxo do sistema, quando seu código fonte for lido por um programador. Assim, criamos o conceito de evento: um registro no banco de dados que contém o usuário que executou a ação, a data e hora deste evento, um campo com o tipo do evento (*login* no sistema, *commit* de dados, etc.) e outro campo descritivo, que armazenará as informações específicas do evento. Isto tornou a implementação desta funcionalidade aderente ao benefício do encapsulamento, ou seja, disponibilização de uma interface

simplificada entre rotinas do sistema, diretiva essa contida no paradigma da orientação a objetos.

Sobre o uso do padrão *tableless* na construção das páginas do banco de dados público do *HTLV-1*, apesar de não ser necessário para o funcionamento da ferramenta, sua adoção acarretou em ganhos relacionados à manutenção e acessibilidade da aplicação. Na versão anterior foi utilizado o recurso de tabelas do HTLM para estruturar as páginas, ou seja, para definir seu formato e aparência, enquanto com a adoção do *tableless* desassociamos estas tabelas da parte estrutural do site, utilizando-as apenas para exibição de dados tabulados, tornando o layout do site mais simples de ler e modificar. Ao mesmo tempo, isto facilita para que *softwares* de leitura de tela possam reconhecer com mais eficiência os elementos que compõe as páginas da aplicação, melhorando a usabilidade para indivíduos que necessitem deste tipo de auxílio. Cabe ressaltar que ainda devem existir funcionalidades que necessitem de modificações para tornar a aplicação ainda mais acessível, mas estas intervenções serão contempladas em manutenções futuras.

A integração entre as ferramentas se deu basicamente utilizando os *Servlets* da linguagem *Java* associados às funções de requisição de conteúdo *web* do *php*. Inicialmente, a implementação desse tipo de comunicação é associada ao uso da tecnologia de *webservices*, um modo de comunicação entre sistemas que envolve o envio de mensagens através de um protocolo comum a estas, como o *SOAP* (WORLD WIDE WEB CONSORTIUM, 1994), que encapsula as mensagens através de um documento *XML*. Entretanto, para viabilizar a comunicação entre os sistemas citados, adotamos uma solução mais simples: o uso dos *Servlets*, o que permitiu requisições *web* simples através dos métodos GET e POST, eliminando a complexidade acidental[2] ocasionada pelo uso dos *webservices* em *Java*. Ao final, ainda que as ferramentas não estejam integradas fisicamente, ou seja, não compartilham os mesmos arquivos, a comunicação automatizada possibilitou o fluxo de dados entre as duas atendendo assim o requisito de associar informações de genotipagem com as informações do banco de dados do *HTLV-1*. Um dos benefícios que pode ser citado quando utilizada esta abordagem é a criação de interface que permite a outras aplicações também

[2] Complexidade adicionada ao projeto de *software* que não é essencial para a resolução do problema, mas que decorre de uma decisão sobre a solução utilizada na aplicação. Pode ser entendida como um efeito colateral de uma tecnologia.

possam fazer uso destes dados no futuro, uma vez que a utilização de requisições *web* é um recurso presente em muitas linguagens de programação.

7 CONCLUSÃO

O desenvolvimento de aplicações (*software*) é sempre de grande importância para a bioinformática, pois fornece meios para o tratamento, processamento e análise dos dados biológicos. Neste trabalho, desenvolvemos ferramentas para auxílio nos estudos referentes aos quatro tipos do HTLV, tanto em sua análise filogenética quanto provendo um repositório *online* de sequências e seus dados biológicos, geográficos, epidemiológicos e clínicos associados.

Para a comunidade científica, a disponibilização destes recursos traz celeridade à pesquisa uma vez que o processo de análise dos dados do HTLV é simplificado através de ferramentas confiáveis e otimizadas. Nos dois anos desde que estas ferramentas vêm sendo publicadas, já receberam mais de três mil acessos e seus artigos foram citados noventa vezes (ALCANTARA et al., 2009) e 7 vezes (ARAUJO et al., 2012). Esperamos assim, que a disponibilização de novas ferramentas, bem como sua integração, produzam mais citações e também mais acessos da comunidade científica, contribuindo de maneira relevante para o desenvolvimento de pesquisas relacionadas ao HTLV.

REFERÊNCIAS

ALCANTARA, L. C. J. et al. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. **Nucleic Acids Research**, v. 37, n. Web Server, p. W634–W642, 1 jul. 2009.

ARAUJO, T. H. A. et al. A public HTLV-1 molecular epidemiology database for sequence management and data mining. **PLoS ONE**, v. 7, n. 9, p. 7–10, 2012.

BASTIAN, I. et al. Isolation of a human T-lymphotropic virus type I strain from Australian aboriginals. **Journal of Virology**, v. 67, n. 2, p. 843–51, fev. 1993.

BIGLIONE, M. M. et al. A cluster of human T-cell lymphotropic virus type I-associated myelopathy/tropical spastic paraparesis in Jujuy, Argentina. **Journal of Acquired immune Deficiency Syndromes**, v. 32, n. 4, p. 441–445, abr. 2003.

BISWAS, H. H. et al. Increased all-cause and cancer mortality in HTLV-II infection. **Journal of Acquired Immune Deficiency Syndromes** , v. 54, n. 3, p. 290–296, jul. 2010.

CALATTINI, S. et al. Discovery of a new human T-cell lymphotropic virus (HTLV-3) in Central Africa. **Retrovirology**, v. 2, n. 1, p. 30, 2005.

CATALAN-SOARES, B. et al. Heterogeneous geographic distribution of human T-cell lymphotropic viruses I and II (HTLV-I/II): serological screening prevalence rates in blood donors from large urban areas in Brazil. **Cadernos de Saude Publica**, v. 21, n. 3, p. 926–931, 2004.

CHEN, J. et al. HTLV type I isolated from a Pygmy in Cameroon is related to but distinct from the known central African type. **AIDS Research and Human Retroviruses**, v. 11, n. 12, p. 1529–1531, dez. 1995.

CHEVENET, F. et al. Searching for virus phylotypes. **Bioinformatics**, v. 29, n. 5, p. 561–570, 2013.

DSIC. **Segurança da Informação**. Disponível em: <http://dsic.planalto.gov.br/seguranca-da-informacao>. Acesso em: 20 nov. 2015.

GALLO, R. C.; DE-THÉ, G. B.; ITO, Y. Kyoto Workshop on Some Specific Recent Advances in Human Tumor Virology. **Cancer Research** , v. 41, n. 11 Part 1, p. 4738–4739, 1 nov. 1981.

GESSAIN, A. et al. Isolation and molecular characterization of a human T-cell lymphotropic virus type II (HTLV-II), subtype B, from a healthy Pygmy living in a remote area of Cameroon: an ancient origin for HTLV-II in Africa. **Proceedings of the National Academy of Sciences of the United States of America**, v. 92, n. 9, p. 4041–4045, abr. 1995.

GESSAIN, A. et al. HTLV-3/4 and simian foamy retroviruses in humans: discovery, epidemiology, cross-species transmission and molecular virology. **Virology**, v. 435, n. 1, p. 187–199, jan. 2013.

GESSIAN, A. et al. Highly divergent molecular variants of human T-lymphotropic virus type I from isolated populations in Papua New Guinea and the Solomon Islands. **Proceedings of the National Academy of Sciences of the United States of America**, v. 88, n. 17, p. 7694–7698, 1 set. 1991.

KALYANARAMAN, V. S. et al. A new subtype of human T-cell leukemia virus (HTLV-II) associated with a T-cell variant of hairy cell leukemia. **Science**, v. 218, n. 4572, p. 571–573, 5 nov. 1982.

LEE, H. et al. High rate of HTLV-II infection in seropositive i.v. drug abusers in New Orleans. **Science**, v. 244, n. 4903, p. 471–475, abr. 1989.

LIFSCHITZ, S. Algumas pesquisas em banco de dados e bioinformática Workshop de Biologia Computacional. In: CONGRESSO DA SOCIEDADE DE BIOLOGIA COMPUTACIONAL, 26., 2016. Anais… Campo Grande, 2016.

LIPMAN, D. J.; PEARSON, W. R. Rapid and sensitive protein similarity searches. **Science,** v. 227, n. 4693, p. 1435–1441, 22 mar. 1985.

MAHIEUX, R. et al. Molecular epidemiology of 58 new African human T-cell leukemia virus type 1 (HTLV-1) strains: identification of a new and distinct HTLV-1 molecular subtype in Central Africa and in Pygmies. **Journal of Virology**, v. 71, n. 2, p. 1317–1333, fev. 1997.

MANUAL do PHP. Disponível em: <https://secure.php.net/manual/pt_BR/index.php>. Acesso em: 20 nov. 2015.

MURPHY, E. L. et al. Increased prevalence of infectious diseases and other adverse outcomes in human T lymphotropic virus types I- and II-infected blood donors. Retrovirus Epidemiology Donor Study (REDS) Study Group. **The Journal of Infectious Diseases**, v. 176, n. 6, p. 1468–1475, dez. 1997.

NAVATHE, R. E. S. B. **Sistemas de Bancos de Dados**. 4. ed. [s.l.] Pearson, 2005.

POIESZ, B. J. et al. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. **Proceedings of the National Academy of Sciences of the United States of America**, v. 77, n. 12, p. 7415–7419, dez. 1980.

PROSDOCIMI, F. et al. Bioinformática: manual do usuário. **Biotecnologia Ciência e Desenvolvimentos**, v. 29, p. 12–25, 2002.

SALEMI, M. et al. Two new human T-lymphotropic virus type I phylogenetic subtypes in seroindeterminates, a Mbuti pygmy and a Gabonese, have closest relatives among African STLV-I strains. **Virology**, v. 246, n. 2, p. 277–287, 5 jul. 1998.

SEIKI, M.; HATTORI, S.; YOSHIDA, M. Human adult T-cell leukemia virus: molecular cloning of the provirus DNA and the unique terminal structure. **Proceedings of the National Academy of Sciences of the United States of America**, v. 79, n. 22, p. 6899–6902, nov. 1982.

SILVA, E. A. et al. HTLV-II infection associated with a chronic neurodegenerative disease: clinical and molecular analysis. **Journal of Medical virology**, v. 66, n. 2, p. 253–257, fev. 2002.

SLATTERY, J. P.; FRANCHINI, G.; GESSAIN, A. Genomic evolution, patterns of global dissemination, and interspecies transmission of human and simian T-cell leukemia/lymphotropic viruses. **Genome Research**, v. 9, n. 6, p. 525–540, jun. 1999.

SOARES, B. C.; PROIETTI, A. B. DE F. C.; PROIETTI, F. A. HTLV-I/II and blood donors: determinants associated with seropositivity in a low risk population. **Revista de Saude Publica**, v. 37, n. 4, p. 470–476, ago. 2003.

SOMMERVILLE, I. **Engenharia de Software**. 6. ed. São Paulo: Pearson/ Addison Wesley, 2011.

TREVIÑO, A. et al. Molecular epidemiology and clinical features of human T cell lymphotropic virus type 1 infection in Spain. **AIDS Research and Human Retroviruses**, v. 30, n. 9, p. 856–862, set. 2014.

UCHIYAMA, T. et al. Adult T-cell leukemia: clinical and hematologic features of 16 cases. **Blood**, v. 50, n. 3, p. 481–492, 1 set. 1977.

VAN DOOREN, S. et al. Evidence for a post-Columbian introduction of human T-cell lymphotropic virus [type I] [corrected] in Latin America. **The Journal of General Virology**, v. 79, Pt 11, p. 2695–2708, nov. 1998.

VANDAMME, A. M. et al. African origin of human T-lymphotropic virus type 2 (HTLV-2) supported by a potential new HTLV-2d subtype in Congolese Bambuti Efe Pygmies. **Journal of Virology**, v. 72, n. 5, p. 4327–4340, maio 1998.

WOLFE, N. D. et al. Emergence of unique primate T-lymphotropic viruses among central African bushmeat hunters. **Proceedings of the National Academy of Sciences**, v. 102, n. 22, p. 7994–7999, 31 maio 2005.

WORLD WIDE WEB CONSORTIUM. **W3C**. Disponível em: <https://www.w3.org/>. Acesso em: 20 jan. 2016.

APÊNDICE A

Comandos executados no MySQL para migrar os dados do HTLV-1 *Molecular Epidemiology Database* para forma normalizada.

```
-- criando as tabelas:
CREATE TABLE `genomic_region` (
  `id_genomic_region` int(11) NOT NULL AUTO_INCREMENT,
  `description` varchar(50) DEFAULT NULL,
  PRIMARY KEY (`id_genomic_region`)
) ENGINE=InnoDB AUTO_INCREMENT=14 DEFAULT CHARSET=latin1;

CREATE TABLE `genotype` (
  `id_genotype` int(11) NOT NULL AUTO_INCREMENT,
  `description` varchar(30) DEFAULT NULL,
  `acronym` varchar(8) DEFAULT NULL,
  PRIMARY KEY (`id_genotype`)
) ENGINE=InnoDB AUTO_INCREMENT=5 DEFAULT CHARSET=latin1;

CREATE TABLE `log` (
  `id_log` int(11) NOT NULL AUTO_INCREMENT,
  `tag` varchar(20) DEFAULT NULL,
  `description` varchar(800) DEFAULT NULL,
  `date` datetime NOT NULL,
  `id_user` int(11) NOT NULL,
  PRIMARY KEY (`id_log`),
  KEY `FK_LOG_USER_idx` (`id_user`),
  CONSTRAINT `FK_LOG_USER` FOREIGN KEY (`id_user`) REFERENCES `user`
(`id_user`) ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB AUTO_INCREMENT=14 DEFAULT CHARSET=latin1;

CREATE TABLE `temp` (
```

```
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `temp_num_acesso` varchar(14) NOT NULL,
  `temp_num_acesso_versao` varchar(20) NOT NULL,
  `temp_num_gi` varchar(15) NOT NULL,
  `temp_type_virus` varchar(10) DEFAULT NULL,
  `temp_gene` varchar(20) DEFAULT NULL,
  `temp_definicao` varchar(455) NOT NULL,
  `temp_organismo` varchar(150) NOT NULL,
  `temp_pubmed` varchar(20) DEFAULT NULL,
  `temp_pais` varchar(70) DEFAULT NULL,
  `temp_codon_start` varchar(10) DEFAULT NULL,
  `temp_proteina_id` varchar(20) DEFAULT NULL,
  `temp_subtipo` varchar(10) DEFAULT NULL,
  `temp_subgrupo` varchar(10) DEFAULT NULL,
  `temp_tam_sequencia` varchar(10) NOT NULL,
  `temp_cds` text NOT NULL,
  `temp_num_de_codons` varchar(15) NOT NULL,
  `temp_env` varchar(10) DEFAULT NULL,
  `temp_note` varchar(10) DEFAULT NULL,
  `pending` int(1) NOT NULL DEFAULT '1' COMMENT '0-Commited; 1-Pending; 2-Excluded',
  `download_date` datetime DEFAULT NULL,
  `id_user` int(11) DEFAULT NULL,
  PRIMARY KEY (`id`),
  UNIQUE KEY `temp_num_gi` (`temp_num_gi`),
  KEY `temp_num_acesso_versao` (`temp_num_acesso_versao`),
  KEY `temp_num_acesso`
(`temp_num_acesso`,`temp_num_acesso_versao`,`temp_num_gi`),
  KEY `temp_type_virus` (`temp_type_virus`,`temp_gene`),
  KEY `temp_gene` (`temp_gene`),
  KEY `temp_pais` (`temp_pais`),
  KEY `temp_subtipo` (`temp_subtipo`)
) ENGINE=InnoDB AUTO_INCREMENT=7196 DEFAULT CHARSET=latin1;
```

```
CREATE TABLE `user` (
  `id_user` int(11) NOT NULL,
  `name` varchar(60) DEFAULT NULL,
  `login` varchar(20) NOT NULL,
  `password` varchar(20) DEFAULT NULL,
  `active` bit(1) DEFAULT NULL,
  PRIMARY KEY (`id_user`),
  UNIQUE KEY `login_UNIQUE` (`login`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1;
```

-- Renomeando a tabela 'final' para 'sequence':

```
ALTER TABLE `finalhtlv`.`final` RENAME TO  `finalhtlv`.`sequence` ;
```

-- Alterando a coluna 'genomic_region' da tabela 'sequence' para que esta contenha o identificador - chave primária - do seu registro correspondente na tabela 'genomic_region':

```
UPDATE sequence AS seq, genomic_region AS genr
SET seq.genomic_region = genr.id_genomic_region
WHERE seq.genomic_region = genr.description;
```

-- Renomeando a coluna 'genomic_region' da tabela 'sequence' para 'id_genomic_region':

```
ALTER TABLE `finalhtlv`.`sequence` CHANGE COLUMN `genomic_region`
`id_genomic_region` VARCHAR(50) NULL DEFAULT NULL;
```

-- Alterando o tipo de dado da coluna 'id_genomic_region' de varchar para número inteiro:

```
ALTER TABLE `finalhtlv`.`sequence` CHANGE COLUMN `id_genomic_region`
`id_genomic_region` INT(11) NULL DEFAULT NULL;
```

-- Adicionando a coluna 'id_genotype' na tabela 'sequence':

```
ALTER TABLE `finalhtlv`.`sequence` ADD COLUMN `id_genotype` INT NULL AFTER
`a_number`;
```

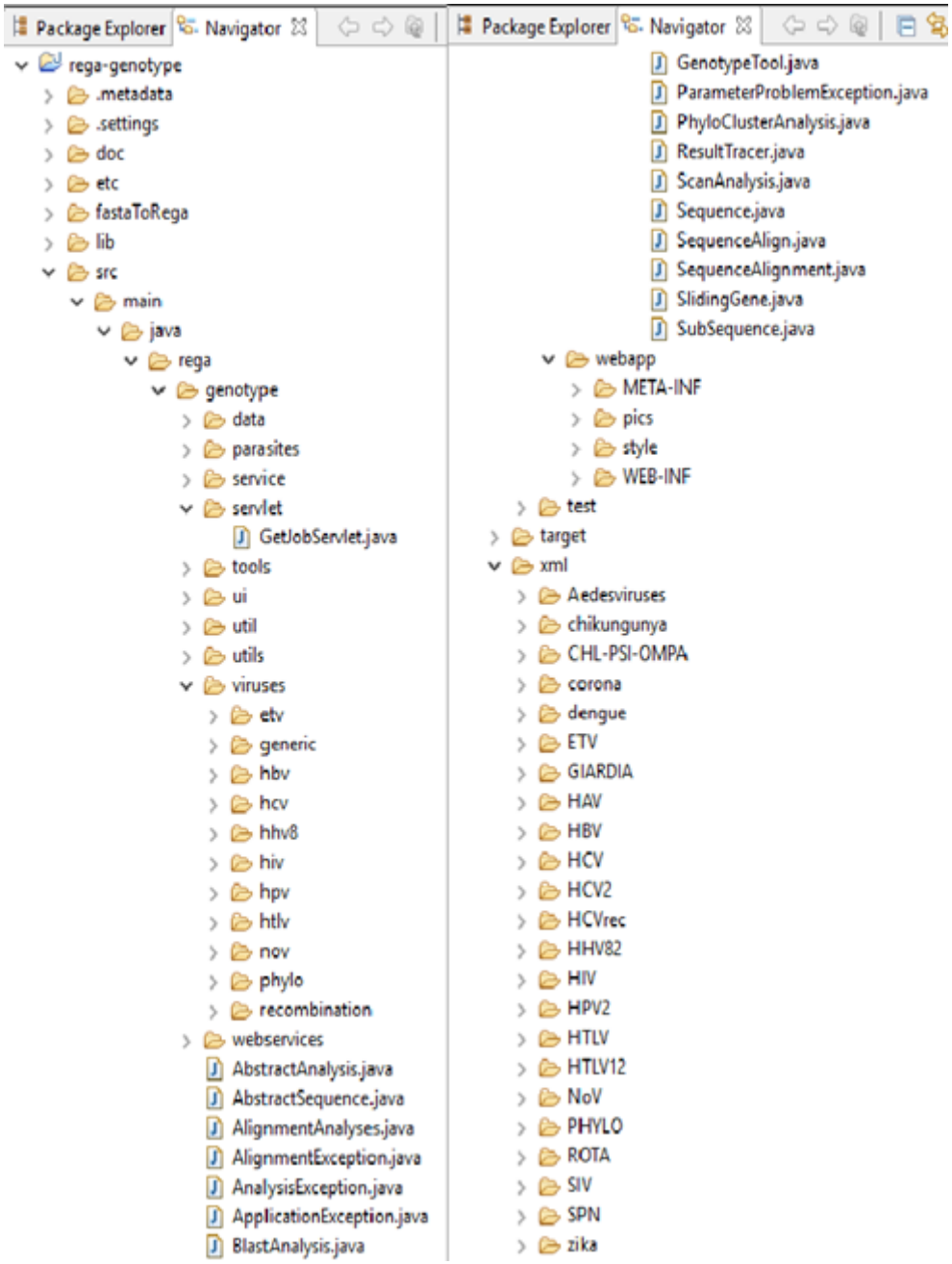
-- Alterando todos os valores para 1, o valor da chave primária do HTLV-1. Como -- todos os dados que existiam eram relacionados ao HTLV-1, executou-se um update sem where:

UPDATE `finalhtlv`.`sequence` set id_genotype = 1;

APÊNDICE B

Hierarquia do HTLV-1 & 2 *Genotype Tool* vista no *Eclipse*

Figura 18. Hierarquia do *REGA Genotype Tool* vista a partir do *Eclipse*.



No diretório "viruses" estão contidos os parâmetros de cada vírus e em "*xml*" estão as sequências referência e os respectivos arquivos de *layout* de páginas. Fonte: O autor.

GLOSSÁRIO

Encapsulamento. Um dos conceitos do paradigma da orientação a objetos, preconiza que se forneçam interfaces para comunicação entre as unidades de *software* de uma aplicação, sem que as mesmas tenham acesso à implementação interna das outras unidades. Estas unidades são tratadas como “caixas-pretas” em que a unidade solicitante não enxerga os mecanismos do processamento de outra unidade, ou seja, acessa somente entradas e saídas de dados.

Métodos GET e POST. São dois dos oito métodos de resposta de um servidor *web,* definidos no protocolo *HTTP*. Dada uma *URL*, é através do método que o servidor decide o que fazer no momento da requisição de um recurso.

Normalização do banco de dados. Consiste em uma série de diretrizes, que aplicadas ao banco de dados, visam tornar o armazenamento dos dados consistente e o acesso a estes otimizado. NAVATHE (2005) definem quatro formas normais para bancos de dados: 1ª Forma Normal, 2ª Forma Normal, 3ª Forma normal e Forma Normal de Boyce-Codd. Comumente diz-se normalizado um banco de dados que atende a algumas das formas normais.

Padrão Tableless. Padrão para construção de *layouts* para páginas *web*.Seu objetivo é coibir a utilização de *framesets* e tabelas como elementos estruturantes do conteúdo das páginas, tornando-as mais simples (internamente) e acessíveis.

Paradigma da Orientação a Objetos. Modelo de análise, projeto e desenvolvimento de *software* que trata os seus aspectos funcionais através da interação de unidades de *software*, chamadas objetos. Estes objetos são representações de elementos do domínio da aplicação, como por exemplo, um *software* de uma loja teria como objetos produtos, clientes, vendas, entre outros.

Protocolo HTTP. O protocolo HTTP (*Hypertext Transfer Protocol*) define padrões de comunicação nas requisições *web* de páginas da Internet. Este protocolo

é mantido pelo *World Wide Web Consortium* (WORLD WIDE WEB CONSORTIUM, 1994) e é a base para a comunicação na Internet.

Requisição *web.* Processo de requisição de dados e resposta entre uma máquina cliente e um servidor *web*.

Requisitos funcionais e não funcionais. Diz-se de requisitos funcionais àqueles relacionados diretamente às funcionalidades do *software*, como telas, menus, entradas e saídas, entre outros. Já os requisitos não funcionais se referem a propriedades que impactam no sistema, mas que não estão explicitadas diretamente nos requisitos, ou seja, se relacionam com a qualidade do *software*, como desempenho, segurança, disponibilidade e usabilidade (SOMMERVILLE, 2011).

SQL. A Structured Query Language é uma linguagem padronizada declarativa de consulta a bancos de dados relacionais. o *SQL ANSI* é o padrão comum, entretanto os vários sistemas gerenciadores de bancos de dados (SGBDs) podem implementar algumas modificações nesta linguagem para se adequarem a seus próprios requisitos.

XML. O *XML* (*eXtensible Markup Language*) é uma linguagem de marcação capaz de descrever diversos tipos de dados. Recomendada pela *W3C*, foi desenvolvida visando a facilidade compartilhamento de informações na Internet.