FUNDAÇÃO OSWALDO CRUZ

CENTRO DE PESQUISAS GONÇALO MONIZ

FIOCRUZ

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

DISSERTAÇÃO DE MESTRADO

DESENVOLVIMENTO DE UM BANCO DE DADOS (*HTLV-1 MOLECULAR EPIDEMIOLOGY DATABASE*) PARA *DATAMINING E DATA MANAGEMENT DE* SEQUÊNCIAS DO HTLV-1

THESSIKA HIALLA ALMEIDA ARAÚJO

Salvador – Brasil
2012

FUNDAÇÃO OSWALDO CRUZ

CENTRO DE PESQUISAS GONÇALO MONIZ

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

DESENVOLVIMENTO DE UM BANCO DE DADOS (*HTLV-1 MOLECULAR EPIDEMIOLOGY DATABASE*) PARA *DATAMINING E DATA MANAGEMENT DE* SEQUÊNCIAS DO HTLV-1

THESSIKA HIALLA ALMEIDA ARAÚJO

Orientador: Prof. Dr. Luiz Carlos Junior Alcântara

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

Salvador – Brasil
2012

Ficha Catalográfica elaborada pela Biblioteca do
Centro de Pesquisas Gonçalo Moniz / FIOCRUZ - Salvador - Bahia.

Araújo, Thessika Hialla Almeida
A662d Desenvolvimento de um banco de dados (HTLV-1 molecular epidemiology database) para datamining e data management de sequências do HTLV-1**.** [manuscrito] / Thessika Hialla Almeida Araújo. - 2012.
60 f.; 30 cm

Datilografado (fotocópia).

Dissertação (Mestrado) – Fundação Oswaldo Cruz, Centro de Pesquisas Gonçalo Moniz. Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa, 2012.
Orientador: Profº. Dr. Luiz Carlos Junior Alcântara, Laboratório de Patologia e Biointervenção - LPBI.

1. HTLV-1 2. Banco de dados. I.Título.

CDU 616.98:004

“Todo o conhecimento é uma resposta a uma pergunta.”

Gaston Bachelard

Dedico aos meus pais
por toda confiança e
amor incondicional.

AGRADECIMENTOS

A Deus pela minha vida, pela onipresente direção dos meus caminhos e por ter me colocado diante de tantas pessoas especiais.

Agradeço àqueles que me deram a vida e me ensinaram a amar, a lutar pelos meus sonhos, a ser ética e a sempre buscar a felicidade nas minhas decisões: meus queridos pais.

Agradeço às minhas irmãs por todo amor e em especial, a Léa por toda a torcida e grande amizade que nos une.

A todos os meus familiares e amigos por todo amor, apoio e carinho demonstrado.

Ao professor e orientador Dr. Luiz Carlos Junior Alcântara por todo conhecimento compartilhado, por acreditar no meu trabalho e por sempre incentivar o meu crescimento profissional.

Ao Dr. Bernardo Galvão pelo apoio constante, por todos valiosos ensinamentos e por ter sido um grande motivador na realização deste trabalho. Minha eterna gratidão.

Aos colaboradores deste projeto Dr. Pieter Libin, Dr. Koen Deforche e em especial a Leandro Souza pela importante e essencial participação no desenvolvimento e conclusão deste projeto.

A todos do LASP – FIOCRUZ pelo amparo e estímulo, e principalmente por tornar os meus dias muito mais leves e prazerosos no laboratório. Em especial ao meu grupo de pesquisa por todo apoio e carinho em todos os momentos.

À equipe do Centro de HTLV da Escola Bahiana de Medicina e Saúde Pública, em especial a Cláudio, Noilson, Sônia e Rodrigo por todo carinho e ajuda.

A todos os professores do Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa que participaram da minha formação acadêmica e profissional.

A todos que contribuíram direta ou indiretamente para a realização deste trabalho e para a minha formação profissional.

ARAÚJO, Thessika Hialla Almeida. Desenvolvimento de um banco de dados (*HTLV-1 Molecular Epidemiology Database*) para *datamining e data management de* sequências do HTLV-1. 60 f. il. Dissertação (Mestrado) – Fundação Oswaldo Cruz, Instituto de Pesquisas Gonçalo Moniz, Salvador, 2012.

RESUMO

As pesquisas biológicas geram uma grande quantidade de informações que devem ser armazenadas e gerenciadas, permitindo que os usuários tenham acesso a dados completos sobre o tema de interesse. O volume de dados não relacionados gerados nas pesquisas com HTLV-1 justifica a criação de um Banco de dados que contenha o maior número de informações sobre o vírus, seus aspectos epidemiológicos, para que possam estabelecer melhores relações sobre infecção, patogênese, origem e principalmente, evolução. Os dados foram obtidos a partir de pesquisa no GENBANK, em artigos relacionados e diretamente com os autores dos dados. O banco de dados foi desenvolvido utilizando o Apache Webserver 2.1.6 e o SGBD – MySQL. A *webpage* foi desenvolvida em HTML a escrita em PHP. Atualmente temos cadastradas 2435 sequências, sendo que 1968 (80,8%) representam diferentes isolados. Em relação ao status clínico, o banco de dados tem informação de 40,49% das sequências, no qual 43%, 18,69%, 32,7%, 5,61% são TSP/HAM, ATL, assintomático e outras doenças, respectivamente. Quanto ao gênero e idade tem-se informação de 15,4% e 10,56% respectivamente. O HTLV-1 *Molecular Epidemiology Database* está hospedado no servidor do Centro de Pesquisa Gonçalo Moniz/FIOCRUZ-BA com acesso em http://htlv1db.bahia.fiocruz.br/, sendo um repositório de sequências do HTLV-1 com informações clínicas, epidemiológicas e geográficas. Esta base de dados dará apoio às investigações clínicas e pesquisas para desenvolvimento de vacinas.

Palavras-chave: HTLV-1, banco de dados.

ARAÚJO, Thessika Hialla Almeida. Development of an HTLV-1 Molecular Epidemiology Database for Sequence Management and Data Mining**.** 60 f. il. Dissertação (Mestrado) – Fundação Oswaldo Cruz, Instituto de Pesquisas Gonçalo Moniz, Salvador, 2012.

ABSTRACT

Scientific development has generated a large amount of data that should be stored and managed in order for researchers to have access to complete data sets. Information generated from research on HTLV-1 warrants the design of databases to aggregate data from a range of epidemiological aspects. This database would support further research on HTLV-1 viral infections, pathogenesis, origins, and evolutionary dynamics. All data was obtained from publications available at GenBank or through contact with the authors. The database was developed using the Apache Webserver 2.1.6 and SGBD MySQL. The webpage interfaces were developed in HTML and sever-side scripting written in PHP. There are currently 2,435 registered sequences with 1,968 (80.8%) of those sequences representing different isolates. Of these sequences, 40.49% are related to clinical status (TSP/HAM, 43%, ATLL, 18.69%, asymptomatic, 32.7%, and other diseases, 5.61%). Further, 15.4% of sequences contain information on patient gender while 10.56% of sequences provide the age of the patient. The HTLV-1 Molecular Epidemiology Database is hosted on the Gonçalo Moniz/FIOCRUZ-BA research center server with access at http://htlv1db.bahia.fiocruz.br/. Here, we have developed a repository of HTLV-1 genetic sequences from clinical, epidemiological, and geographical studies. This database will support clinical research and vaccine development related to viral genotype.

Keys word: HTLV-1, database.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

3’	Região carboxi-terminal do ácido nucléico
5’	Região amino-terminal do ácido nucléico
ATL	Leucemia/linfoma de células T do adulto (*Adult T cell Leukemia*)
BD	Banco de Dados
BDB	Banco de dados biológicos
BLAST	Basic Local Alignment Search Tool
CD4+	*cluster of differentation 4*
CD8+	*cluster of differentation 8*
csv	*comma-separated values*
DDBJ	*DNA Data Bank of Japan*
DNA	Ácido desoxirribonucléico
EBI	*European Bioinformatics Institute*
EMBL	*European Molecular Biology Laboratory*
FIOCRUZ	Fundação Oswaldo Cruz
HCV	Vírus da hepatite C
HLA	*Human Lecocitary Antigen*
HIV	Vírus da Imunodeficiência Humana (*Human Imunodeficiency Vírus*)
HTLV-1	Vírus Linfotrópico de células T humanas tipo 1 (*Human T cell Lymphotropic vírus type 1*)
HTLV-2	Vírus Linfotrópico de células T humanas tipo 2 (*Human T cell Lymphotropic vírus type 2*)
HTLV-3	Vírus Linfotrópico de células T humanas tipo 3 (*Human T cell Lymphotropic vírus type 3*)
HTLV-4	Vírus Linfotrópico de células T humanas tipo 4

	(*Human T cell Lymphotropic vírus type 4*)
HAU	Uveíte associada ao HTLV
IDE	*Integrated Development Environment*
INSDC	*International Nucleotide Sequence Database Collaboration*
LTR	Extremidades em repetições longas (*Long Terminal Repeat*)
NCBI	*National Center for Biotechnology Information*
NIH	*National Institutes of Health*
nm	nanômetros
ORF	Fase de Leitura aberta (*Open Reading Frame*)
pb	Pares de bases
RNA	Ácido Ribonucléico
SGBD	Sistema de gerenciamento de bancos de dados
SQL	*Structured Query Language*
TSP/HAM	Paraparesia Espástica Tropical (*Tropical Spastic Paraparesis*)/
	Mielopatia Associada ao HTLV *(HTLV Associated Mielopathy)*
xml	*Extensible markup language*

SUMÁRIO

1 INTRODUÇÃO

As pesquisas realizadas em ciências biológicas geram uma grande quantidade de dados que necessitam ser armazenados e gerenciados, para atender as necessidades de múltiplos usuários. Em particular, o gerenciamento eficaz dos dados envolvidos faz com que as pesquisas em bancos de dados ganhem rumos e aplicações novas, fornecendo informações mais completas e atualizadas sobre o assunto de interesse.

No entanto, apesar da grande quantidade de informações, a velocidade na qual os usuários conseguem interpretá-las ainda é insatisfatória, exigindo mecanismos eficientes de armazenamento e análise. Neste contexto é cada vez mais necessária a utilização de ferramentas da bioinformática e de um Sistema de Gerenciamento de Banco de Dados (SGBD) específico (LIFSCHITZ, 2006).

A partir da descoberta do vírus linfotrópico de células-T humanas do tipo 1 (HTLV-1) na década de 80 (POIESZ *et al,* 1980) até os dias atuais, já se passaram mais de 30 anos. Durante este período foram realizadas inúmeras pesquisas e, consequentemente, gerou-se uma grande quantidade de dados com a finalidade de se encontrar as melhores respostas para perguntas relacionadas à patogênese, transmissão, polimorfismo gênico, epidemiologia, relação fenótipo-genótipo, distribuição geográfica e evolução viral.

A proposta deste trabalho foi reunir o máximo de informações relacionadas às sequências de nucleotídeos do HTLV-1, indexadas no *GenBank*, organizando-as em um banco de dados relacional com uma interface *web*, nomeado de *HTLV-1 Molecular Epidemiology Database* (Banco de Dados de Epidemiologia Molecular do HTLV-1).

2 REVISÃO BIBLIOGRÁFICA

2.1 HISTÓRICO DO HTLV

O 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. Entretanto, anos antes no sudoeste do Japão foi definida em pacientes, uma forma distinta de leucemia com características clínicas e com morfologia celular especiais, 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).

Em 1982, Kalyanaraman e colaboradores isolaram de um paciente com uma forma atípica de Leucemia de células T pilosas, o HTLV-2, e mais recentemente, os HTLV-3 e HTLV-4 foram descritos em indivíduos de Camarões, na África Central (WOLFE *et al,* 2005; CALATTINI *et al,* 2005).

2.2 AGENTE ETIOLÓGICO E ORGANIZAÇÃO GENÔMICA

O HTLV pertence à família *Retroviridae*, à subfamília *Orthoretrovirinae*, gênero *Deltaretrovirus*. Trata-se de um vírus envelopado, com diâmetro de aproximadamente 100 a 140 nanômetros (nm), possui um nucleocapsídio icosaédrico com cerca de 80 a 100 nm. O vírus possui um genoma composto por duas fitas simples de RNA (ácido ribonucléico) com polaridade positiva (BURKE, 1997) (Figura 1), formados pelos genes *gag, pro, pol, env* e região *pX*, além de ser flanqueado por duas regiões repetidas, chamadas *LTR* (*long terminal repeats),* importantes na integração do DNA proviral no DNA cromossômico do hospedeiro e na regulação transcricional do genoma do HTLV (GREEM & CHEN, 2001).

O HTLV-1, HTLV-2, HTLV-3 e HTLV-4 possuem o DNA proviral de 9032, 8952, 8553 e 8791 pares de bases (pb), respectivamente (SEIKI *et al,* 1983; TSUJIMOTO *et al,* 1988; CALATTINI *et al,* 2005 e WOLFE *et al,* 2005).

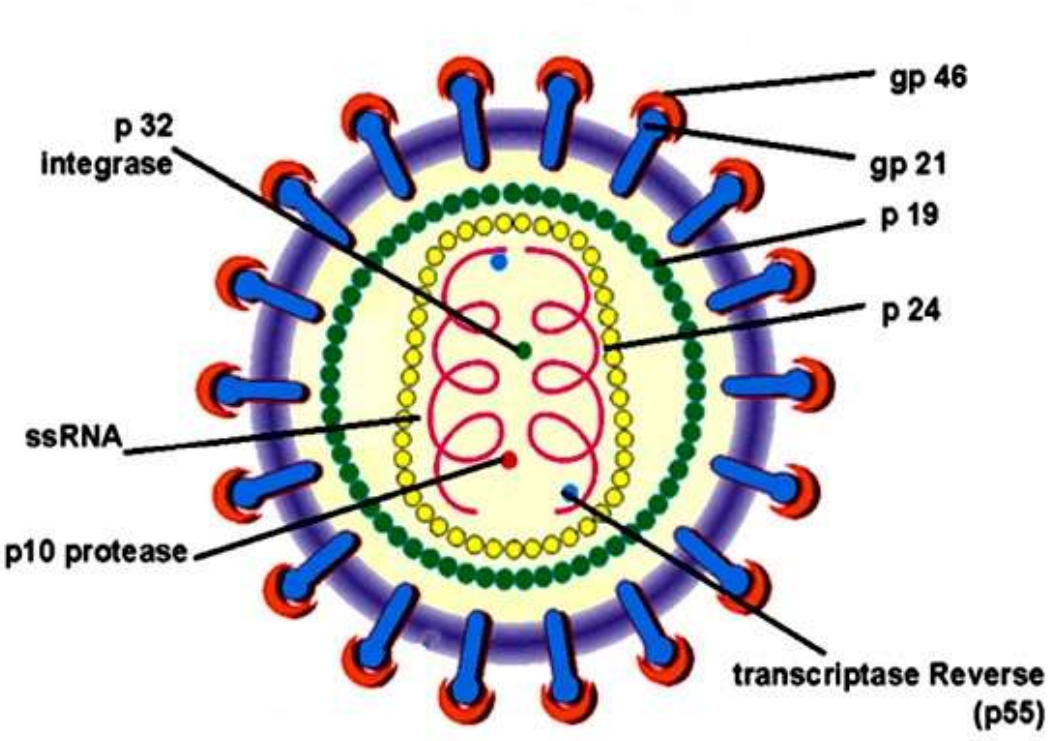


Figura 1. Estrutura do HTLV-1. (adaptado do site http://www.htlv.com.br)

O gene *gag* codifica as proteínas estruturais do capsídeo viral. A região é inicialmente traduzida como uma poliproteína precursora que é posteriormente clivada dando origem às proteínas da matriz (p19), do capsídeo (p24) e nucleocapsídio (p15).

O gene *pol* codifica enzimas virais transcriptase (TR), RNAseH e integrase (IN). A TR é responsável pela síntese do DNA viral a partir do seu genoma RNA, estando presente no cerne da partícula viral. A RNAse atua na remoção da fita RNA molde após a síntese da cadeia de DNA, degradando seletivamente o RNA da molécula híbrida DNA-RNA e, por fim, a IN, enzima responsável pela integração do DNA viral no genoma da célula hospedeira.

O gene *env* codifica as proteínas do envelope viral (ENV). A proteína precursora do ENV é clivada para gerar os produtos maduros, a glicoproteína de superfície (gp46 – SU) e uma proteína transmembrana (gp21- TM) (SHIMOTOHNO *et al,* 1985) ambas são importantes para a interação com a célula alvo e posterior infecção.

A região situada imediatamente antes da região *LTR* 3' (*long terminal repeat*), denominada *pX*, contêm 4 *Orfs* (*Open Reading Frames).* A *orf*-I do gene *pX* codifica a proteína p12 (KORALNIK *et al,* 1992; KORALNIK *et al,* 1993; FUKUMOTO *et al,* 2009). As proteínas p13 e p30 são codificadas pelo orf-II do gene *pX* (KORALNIK *et al,* 1992; CIMINALE *et al,* 1992) e as proteínas *Tax* (p40) e *Rex* (p27) responsáveis pela regulação da replicação viral, são codificadas pelas *Orfs* IV e III, respectivamente. As duas regiões *LTR* idênticas, localizadas nas extremidades do DNA viral, contêm regiões promotoras virais, bem como outros elementos regulatórios (Figura 2).

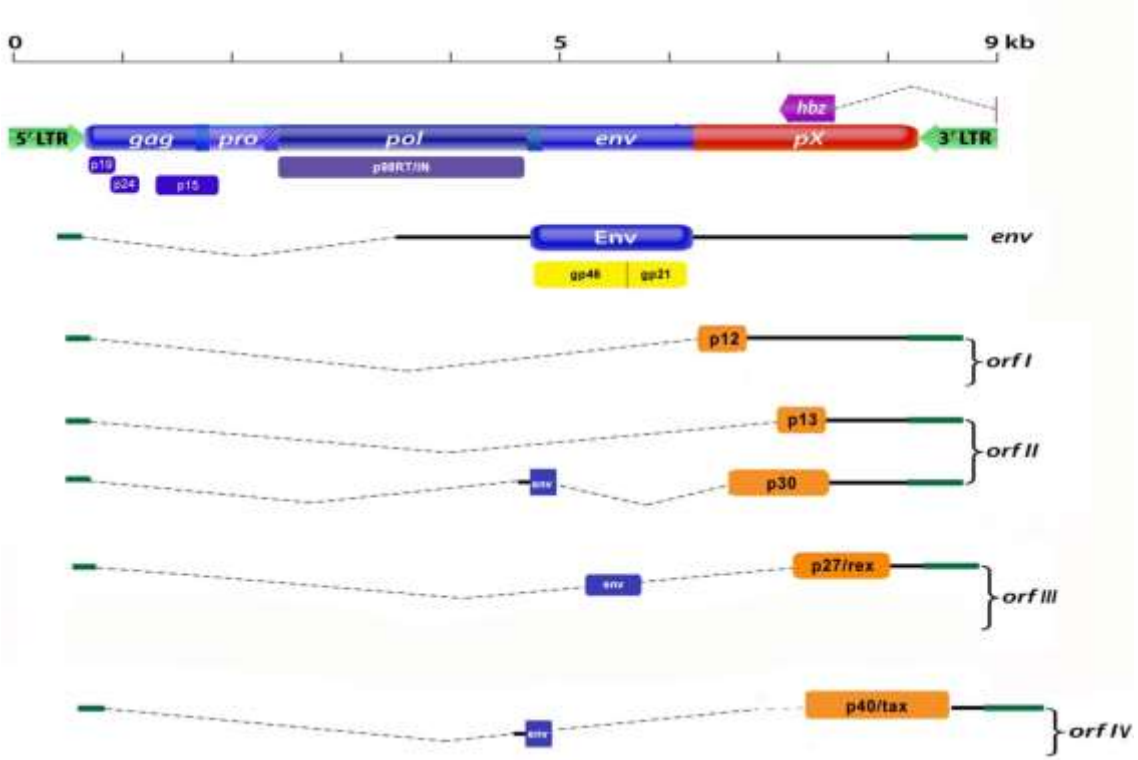


Figura 2. Estrutura genômica do HTLV-1 e seus transcritos (adaptado de Edwards *et al,* 2011).

2.3 EPIDEMIOLOGIA DO HTLV-1

2.3.1 Distribuição geográfica da infecção pelo HTLV-1

Estima-se que aproximadamente 15 a 20 milhões de pessoas são infectadas pelo HTLV em todo o mundo (DE THE & KAZANJI, 1996). As taxas de soroprevalência diferem, de acordo com a região geográfica, com a composição sócio-demográfica da população estudada e os comportamentos de risco individuais. Dados epidemiológicos mostram que a infecção pelo

HTLV-1 tem distribuição mundial (DE THE & KAZANJI, 1996), no entanto, algumas áreas são endêmicas para esta infecção: sudoeste do Japão (YAMAGUCHI, 1994; MUELLER *et al,* 1996), África sub-Saara (GESSAIN e DE THE, 1996), regiões do Caribe (HANCHARD *et al,* 1990), áreas localizadas no Irã e Melanésia (MUELLER, 1991) e Brasil (CATALAN-SOARES *et al,* 2004) (Figura 3).

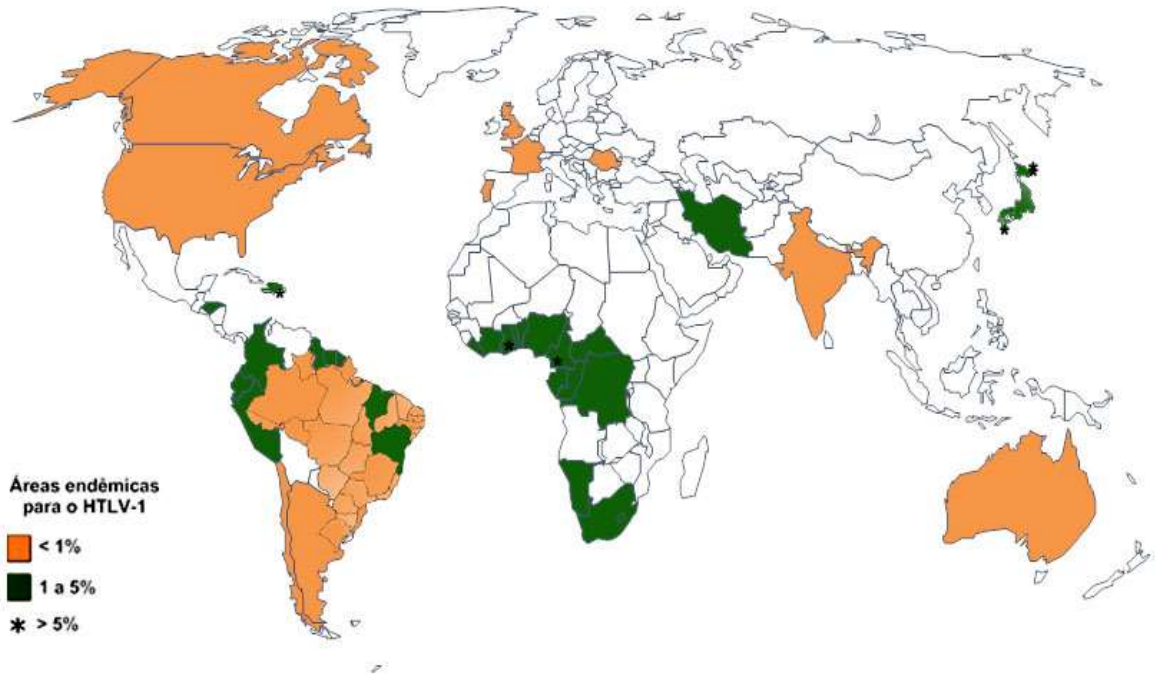


Figura 3. Distribuição do HTLV-1 no mundo. (adaptado de PROIETTI *et al,* 2005).

O Brasil configura-se como uma área endêmica para o HTLV-1, porém com baixo índice de prevalência e com variações nas capitais (CATALAN-SOARES *et al,* 2004) (Figura 4). A maioria destes dados foi obtido a partir de amostras provenientes de bancos de sangue ou amostras de grupos específicos (gestantes, pacientes de clínicas de doenças sexualmente transmissíveis, co-infectados), que não necessariamente representam a população geral (revisado por PROIETTI *et al,* 2005). Entretanto, o único estudo que demonstra dados com base populacional foi o realizado por Dourado e colaboradores (2003) na cidade de Salvador - Bahia, no qual demonstrou- se uma soroprevalência de 1,8 %, estimando-se que existam cerca de 40 a 50 mil indivíduos infectados na cidade.

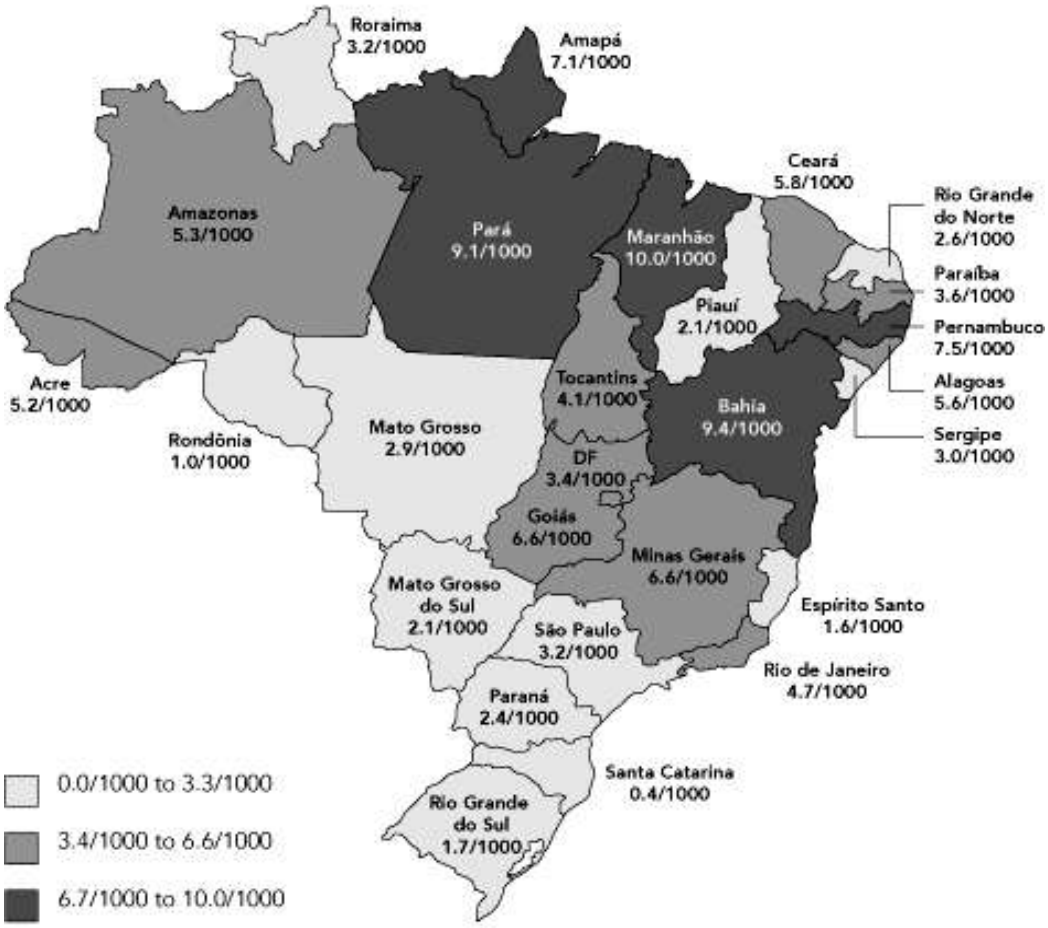


Figura 4. Prevalência de HTLV-1 entre doadores de sangue em capitais de 26 estados brasileiros e no Distrito Federal. Adaptado de CATALAN-SOARES *et al,* 2004.

2.3.2 Epidemiologia molecular do HTLV-1

O HTLV-1 é classificado em vários subtipos virais baseados em diferenças nas sequências do gene *env* e *LTR* de DNA proviral, sendo estratificados em 7 subtipos: “a” ou Cosmopolita (SEIKI *et al,* 1982) que é encontrado em áreas endêmicas em todo o mundo; “b” ou Central Africano; “c” ou da Melanésia (GESSAIN *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” no isolado de pigmeus na República Democrática do Congo (SALEMI *et al,*1998); “f”, de um indivíduo do Gabão (SALEMI *et al,*1998); e “g”, recentemente descrito como um novo subtipo em Camarões, na África Central (WOLFE *et al,* 2005).

O subtipo Cosmopolita (a) é dividido em 5 subgrupos, a depender de sua localização: A –Transcontinental, B – Japonês, C – Oeste Africano, D – Norte Africano e E – Negro do Peru (HAHN *et al,* 1984; GESSAIN *et al,* 1991; MIURA *et al,* 1994 e 1997; GASMI *et al,* 1994; VIDAL *et al,* 1994; VANDAMME, *et al,* 1994; CHEN *et al,* 1995; SALEMI *et al,* 1998; VAN DOOREN *et al,* 1998; WOLFE *et al,* 2005) (Figura 5).

Figura 5. Distribuição mundial dos subtipos e subgrupos do HTLV-1.

2.4 DOENÇAS ASSOCIADAS AO HTLV-1

O HTLV-1 é o agente etiológico da paraparesia espástica tropical/mielopatia associada ao HTLV (TSP/HAM), que é 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). Esta patologia se

desenvolve aproximadamente em 0.3-2% das pessoas infectadas (EDWARDS *et al*, 2011), afetam mais Feminino do que masculino, sendo que a maioria dos indivíduos tem o diagnóstico tardio da doença, cerca da 4ª a 5ª década de vida (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).

Esse vírus 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).

2.5 SISTEMAS DE INFORMAÇÃO E BANCO DE DADOS BIOLÓGICOS

Um sistema de Banco de Dados (BD) é formado basicamente pelos próprios dados, o *software* do SGBD (Sistema de Gerenciamento de Bancos de Dados), o *hardware* do sistema, a mídia de armazenamento e os aplicativos que acessam, recuperam e atualizam componentes dos dados no sistema. Para a criação de um banco de dados específico faz-se necessário a análise de quais informações ele deverá conter, tal qual a sua estrutura, relacionamentos entre as informações, e possíveis restrições, para manter a segurança no armazenamento.

O Banco de Dados Biológicos (BDB) é constituído de tabelas contendo grande quantidade de registros. Por exemplo, o registro associado a uma determinada sequência de nucleotídeo, contém normalmente descrição do nome científico, além de citações na literatura correspondente. Esses BDB utilizam sistemas de banco de dados relacional ou sistemas orientados a objetos e podem ser classificados segundo as informações biológicas armazenadas, como por exemplo, os bancos de dados de sequências de nucleotídeos, que armazenam sequências de nucleotídeos, e as anotações contendo dados de características biológicas sobre as mesmas, esses banco de dados são geralmente associados a um *software* desenvolvido para atualizar, consultar e recuperar informações armazenadas no sistema (BIOINFORMATICS FACTSHEET, 2011).

As anotações consistem na caracterização da sequência genômica, cujo objetivo é a conversão dos dados em informações biologicamente relevantes. (LEMOS 2004; WEISS, 2010). Uma anotação é uma descrição de características em mais alto nível da sequência biológica, ou biossequência. Anotações úteis incluem vários tipos de informações, por exemplo, se um trecho de DNA contém um gene e qual a função dele (LEMOS, 2004).

Os bancos de dados biológicos são classificados em primários, no quais são formados pela deposição direta de sequências de nucleotídeos, aminoácidos ou estruturas protéicas, 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) constituem o *International Nucleotide Sequence Database Collaboration* (INSDC) e os secundários, que são aqueles que derivam dos primários, ou seja, foram formados usando as informações depositadas nos bancos primários (PROSDOCIMI et al., 2002, p.14).

De acordo com Elmasri e Navathe (2005), os bancos de dados biológicos devem ser principalmente:

- Flexíveis ao lidar com tipos de valores e dados, a colocação de restrições deve ser limitada, uma vez que isso pode excluir valores inesperados, sendo que a exclusão desses valores resulta em perda de informação;

- Fáceis em relação à usabilidade, ou seja, as interfaces do banco de dado devem exibir para os usuários informações de maneira que seja aplicável para o problema que eles estejam tentando tratar e reflita a estrutura dos dados de bases;
- Capazes de dar suporte a consultas complexas, pois a definição e a representação destas consultas são extremamente importantes para os estudos biomédicos. Sem conhecimento da estrutura de dados, os usuários comuns não podem construir por conta própria uma consulta complexa através dos dados. Sendo assim, os sistemas devem fornecer ferramentas para que se construam essas consultas.

As redundâncias e inconsistências de dados em BDB são inevitáveis, visto que cada laboratório têm critérios diferentes de qualidade sobre as sequências a serem armazenadas. Portanto, tais bancos de dados apresentam diversos erros, por terem sequências incompletas, contaminadas e com erros vindos do próprio sequenciamento, descrições incompletas e até mesmo errôneas, sobre as sequências submetidas.

3 JUSTIFICATIVA

O presente trabalho propôs a construção de um repositório público para as sequências do HTLV-1 indexadas no *GenBank*, considerando a ausência de uma fonte de dados que disponibilizasse informações mais completas e estruturadas sobre HTLV-1. Informações estas, importantes na avaliação de aspectos clínicos e epidemiológicos, para estudos sobre infecção, patogênese, origem e evolução da infecção pelo vírus. Até o presente momento, existe um banco de dados hospedado no servidor do *BIOAFRICA*, *o RNA Virus Database* (http://newbioafrica.mrc.ac.za/rnavirusdb/virus), contendo apenas os genomas completos do HTLV-1 e uma ferramenta para reconstrução filogenética.

O desenvolvimento do banco de dados para HTLV-1 inspirou-se na iniciativa de outros bancos de dados importantes para as pesquisas biomédicas tais como o de sequências do HIV e HCV de Los Alamos (http://www.hiv.lanl.gov/content/index) e o banco de dados de HIV de Stanford

(http://hivdb.stanford.edu/), que constituem fontes de informações importantes sobre as sequências e epítopos para uso em estudos evolucionários e desenvolvimento de vacinas, dados epidemiológicos, caracterização da epidemiologia viral e de migração populacional.

4 OBJETIVOS

4.1 OBJETIVO GERAL

Construir um repositório de sequências genéticas, dados clínicos/ epidemiológicos e geográficos do HTLV-1, dando subsídio às pesquisas clínicas, relacionadas ao comportamento evolutivo viral, genótipo-fenótipo e desenvolvimento de vacinas.

4.2 OBJETIVOS ESPECÍFICOS

- Realizar levantamento e montagem de banco de dados com todos os isolados do HTLV-1, previamente publicados no banco mundial de sequências (*GenBank);*
- Disponibilizar as sequências (baseado nas informações do banco de dados) organizadas de acordo com a variável selecionada pelo usuário;
- Realizar uma análise descritiva de todas as sequencias coletadas do *GenBank* e organizadas no banco de dados.

5 MATERIAL E MÉTODOS

5.1 REVISÃO BIBLIOGRÁFICA E ESTUDO DE TÓPICOS ESPECIAIS

Realizou-se uma revisão bibliográfica, com foco em repositório de sequências virais, sistemas de gerenciamento de bancos de dados (SGBD) em SQL (*Structured Query Language): MySQL* (http://www.mysql.com), desenho de bases de dados relacionais e linguagem de programação para aplicações de bioinformática, com intuito de estruturar o banco de dados.

5.2 COLETA/MINERAÇÃO E ARMAZENAMENTO DE SEQUÊNCIAS DO HTLV-1

A *priori* as sequências do HTLV-1 e suas respectivas anotações foram coletadas diretamente do *GenBank (http://www.ncbi.nlm.nih.gov/GenBank/)*, um banco público de sequências de nucleotídeos e anotações de apoio bibliográfico e biológico, produzido e mantido pelo *National Center for Biotechnology Information* (NCBI) do *National Institutes of Health* (BENSON, 2010). Estas sequências foram submetidas ao algoritmo BLAST (*Basic Local Alignment Search Tool*) (http://blast.ncbi.nlm.nih.gov/Blast.cgi) com o objetivo de comparar as informações das sequências nucleotídicas obtidas com as contidas na biblioteca ou base de dados da própria ferramenta.

Posteriormente, realizou-se à mineração dos dados, visando garantir a integridade das informações, buscar padrões consistentes e relacionamento sistemático entre as variáveis e por fim, descartar possíveis sequências que não seriam de interesse do presente estudo.

As anotações ausentes nos arquivos obtidos no *GenBank* foram coletadas inicialmente nos respectivos artigos, caso não fossem conseguidas todas as informações, recorreu-se aos autores correspondentes por *email* e/ou telefone.

Os dados foram normalizados e armazenados no SGBD *MySQL* versão 5.5. Para facilitar a adição e a manipulação das informações utilizamos a *IDE*

(*Integrated Development Environment*) *MySQL-Front* versão 5.1 *trial* (http://www.mysqlfront.de/) (Figura 6).

Figura 6. Esquema da coleta/mineração e armazenamento de sequências do HTLV-1.

5.3 DESENHO E IMPLEMENTAÇÃO DO BANCO DE DADOS

O banco de dados foi inicialmente modelado utilizando-se o SGBD *MySQL*, sendo que as informações foram organizadas em uma tabela contendo os seguintes campos:

- **Dados das sequências:**
  - Número de Identificação
  - Região Genômica
  - Status (completa ou parcial)
  - Isolado
  - Subtipo
  - Subgrupo
  - Tamanho da sequência

- **Dados Clínicos e Epidemiológicos**
  - Gênero
  - Idade
  - Etnia
  - Região Geográfica
  - Continente: continente referente à região geográfica;
  - Perfil Clínico
  - Carga Proviral
  - CD4+ (*cluster of differentation 4)*
  - CD8+ (*cluster of differentation 8)*
  - Data de coleta

Além destas informações, também foram incluídos campos para identificação dos autores e origem das publicações:

- Contato
- Artigo
- Autores
- Revista

5.4 DEFINIÇÃO DAS *OPEN READING FRAMES (ORFS)*

Foram coletadas no *GenBank* sequências nucleotídicas no formato *fasta*, arquivo caracterizado por uma ou mais sequências, contendo um cabeçalho com informações indicado pela presença do símbolo de maior (>) na primeira coluna, com a sua região genômica descrita. Essas sequências foram editadas no programa *Genedoc* (NICHOLAS *et al,* 1997) e posteriormente convertidas em aminoácidos; identificou-se o códon de iniciação (ATG) e o códon de parada (TAA, TGA ou TAG) e, em seguida, removeram-se os aminoácidos excedentes, e converteu-se o resultado em nucleotídeos. Sucessivamente, os arquivos finais foram exportados em *fasta* e alinhados com a sequência referência ATK-1, utilizando o programa *Clustal X* (JEANMOUGIN *et al,* 1998) com a localização da posição dos códons de iniciação e os de parada na referência (Figura 7).

Vale ressaltar que os artigos que descreviam as *orfs* não traziam as informações completas.

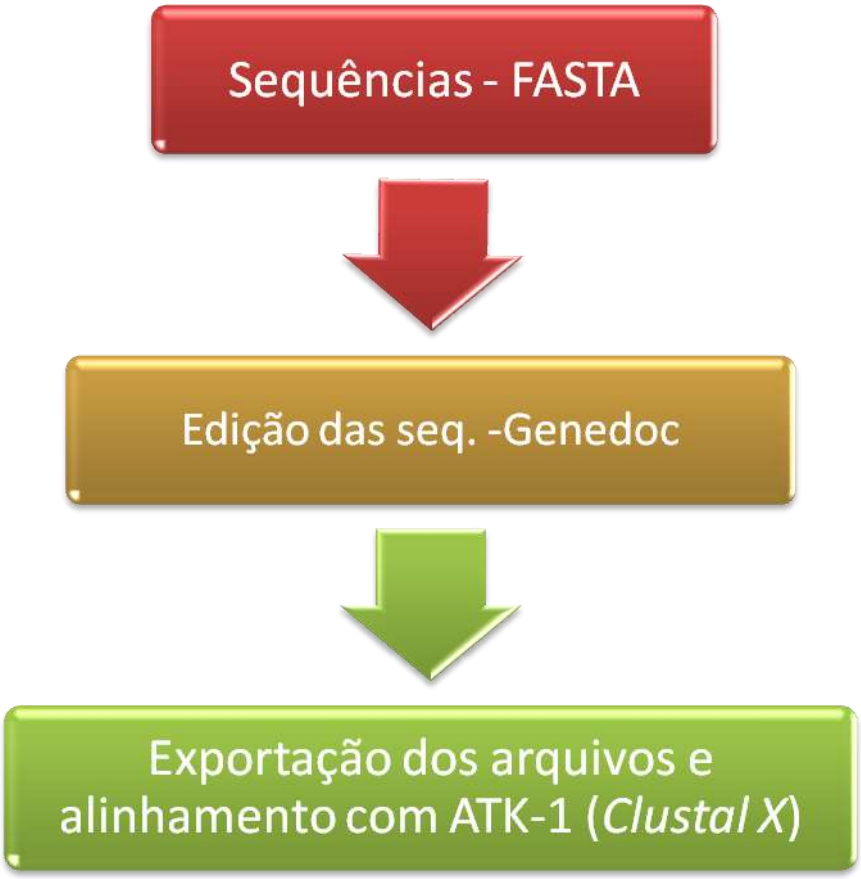


Figura 7. Definição das *Open Reading Frames (Orfs).*

5.5 SUBTIPAGEM DAS SEQUÊNCIAS

As sequências depositadas no banco de dados com a região genômica possuindo *LTR* e/ou *env* que não tinham anotações ou informações nos artigos respectivos sobre o subtipo e/ou subgrupo foram submetidas às ferramentas de bioinformática.

5.5.1 Subtipagem das sequências *LTR*

As sequências com a região genômica *LTR* foram subtipadas com a ferramenta online *LASP HTLV-1 Automated Subtyping Tool (http://www.bioafrica.net/rega-genotype/html/subtypinghtlv.html)*, ferramenta que utiliza métodos filogenéticos para identificar o subtipo de sequências da consulta.

5.5.2 Subtipagem das sequências *env*

O primeiro passo foi à identificação e coleta no *GenBank* de sequências *env* no formato *fasta* com os subtipos específicos:

- Subtipo “a”: AY604882.1, AY604883.1, AY604885
- Subtipo “b”: AY604924.1, AY604928
- Subtipo “c”: L02534.1
- Subtipo “d”: L76414.1
- Subtipo “e”: Y17021.1
- Subtipo “f”: Y17022.1

Até o presente momento, não há no *GenBank* sequência *env* do subtipo “g”.

Posteriormente, foi realizado o alinhamento no programa *Clustal X* (JEANMOUGIN *et al,* 1998) das sequências de interesse presentes no banco de dados, juntamente com todos os subtipos acima citados. As sequências

foram editadas no programa *Genedoc* (NICHOLAS *et al,* 1997) e exportadas no formato *phylip (.phy),* e subsequentemente o arquivo foi adicionado no diretório do pacote para análise filogenética *Phylip (PhyLogeny Inference Package)* versão livre 3.69 (http://evolution.genetics.washington. edu/phylip.html), que permite a análise de vários tipos de dados gerando matrizes de distâncias genéticas diversas e árvores por vários métodos. Utilizou-se o programa de distâncias *DNAdist* e de árvore *Neighbor* para a construção das árvores. Depois de obtido, o arquivo de árvore (*treefile)* foi visualizado no programa *TreeView* versão 1.6.6 (http://taxonomy.zoolo gy.gla.ac.uk /rod/treevi ew.html).

5.6 ESTRUTURA DO *HTLV-1 MOLECULAR EPIDEMIOLOGY DATABASE*

O banco de dados foi desenvolvido utilizando-se o *Webserver Apache* 2.1.6, programa que permite que o computador receba e responda requisições oriundas da rede, associado ao *SGBD Mysql*. As *interfaces* foram desenvolvidas em *html* (*HyperText Markup Language)* e o motor de busca escrito com o *PHP* (*Hypertext Preprocessor*) *Free* versão 5.

A página em *html* foi desenvolvida com campos específicos de busca, permitindo que se faça diversas combinações de dados. A partir do momento que o usuário solicita a busca, a página cria um formulário (*tag form)* contendo os valores (variáveis) selecionados. Este formulário será enviado para um *script* escrito em *php*, responsável em tratar as informações e posteriormente, utilizá-las para executar a busca nos dados armazenados no banco de dados (*Mysql*). Com a posse destes dados, outro script também escrito em *php* irá organizá-los em outra página escrita em *html*, permitindo a visualização e *download* destes dados (Figura 8).

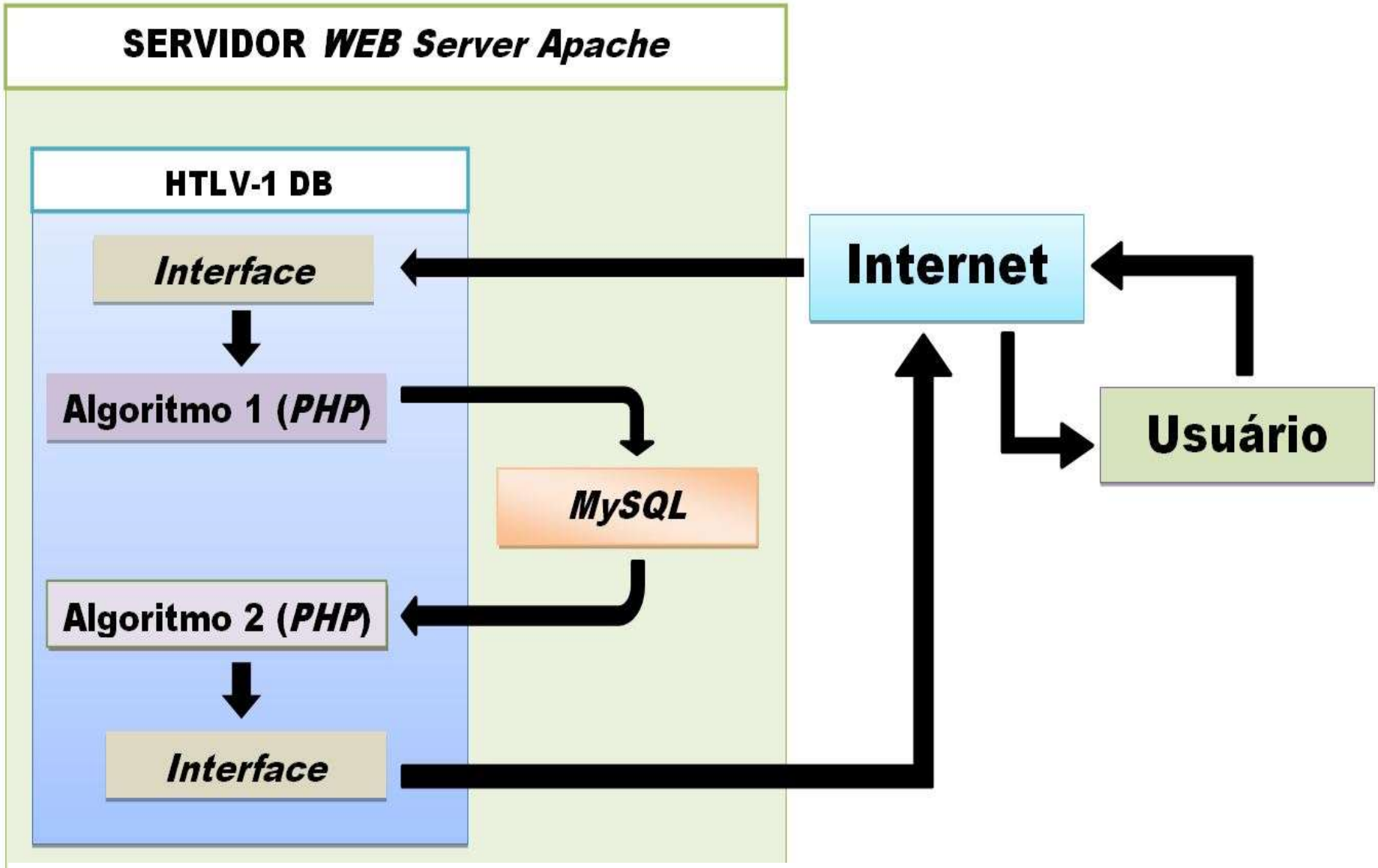


Figura 8. Estrutura do *HTLV-1 Molecular Epidemiology Database.*

6 RESULTADOS

6.1 ANÁLISE DESCRITIVA DOS DADOS

O banco de dados possui 2435 registros de sequências de HTLV-1 do *GenBank* estratificas por regiões genômicas: *env* (30,26%), *pX* (29,69%), *LTR* (34,86%), *gag-pol-env-pX (0,04%), LTR-gag-pol (0,04%), LTR-gag-pol-env-pX (0,04%), pol-env (0,04%), LTR-gag (0,08%), env-pX (0,28%), pX-LTR (1,27%), pol (0,82%), gag (*1,02%)*, gag-pol (*1,02%) *e* genoma completo (0,49%) (Figura 9). Considerando o *status* da região genômica das sequências, 2240 (92%) são parciais e 195 (8%) são completas.

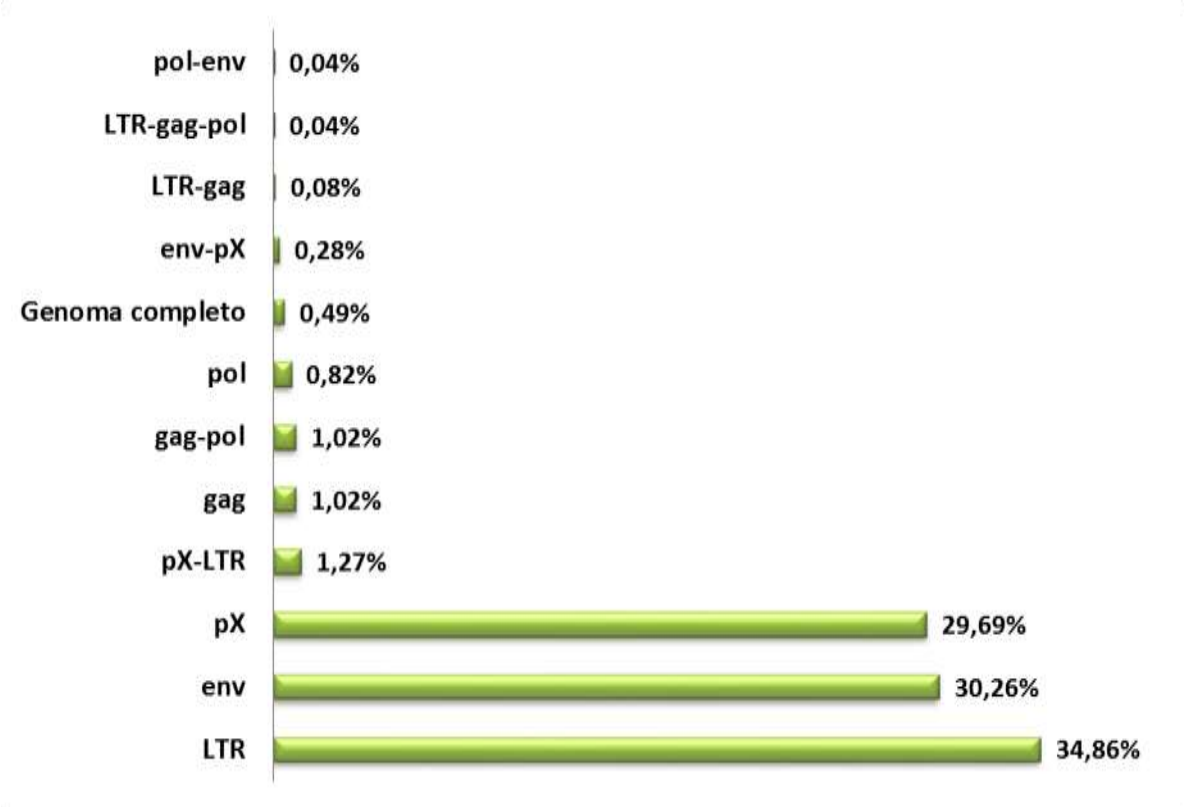


Figura 9. Estratificação das regiões genômicas parciais e completas.

Das 2435 sequências adicionadas ao banco de dados, 1877 (77,08%) representam diferentes isolados, ou seja, as informações foram consideradas uma única vez, mesmo que estes isolados tivessem outros números de acesso, acrescidos de 91 (3,73%) sequências sem o nome do isolado descrito nas

anotações do *GenBank*. Estas sequências sem o nome do isolado foram consideradas a partir da análise da descrição da região genômica e do país de origem, evitando assim, a análise repetida de sequência de um mesmo indivíduo. Ao final, 1968 sequências foram analisadas.

Em relação à distribuição geográfica, das 1968 sequências de isolados, 1822 (92,58%) possuem informações, distribuídas em: Oceania (2,87%), África (12,85%), Europa (3,54%), Ásia (34,51%), sendo que 558/730 (76,43%) das sequências são provenientes do Japão, 1,40% da América do Norte, 3,18% da América Central e 41,61% da América do Sul (Figura 10). Das 756 sequências da América do Sul, 419 (55,42%) são provenientes do Brasil, 154 (20,37%) são da Argentina e as outras 183 (24,20%) sequências são do Chile (1,58%), Bolívia (0,2%), Peru (4,49%), Guiana Francesa (11,9%) e Colômbia (5,15%) (Figura 11).

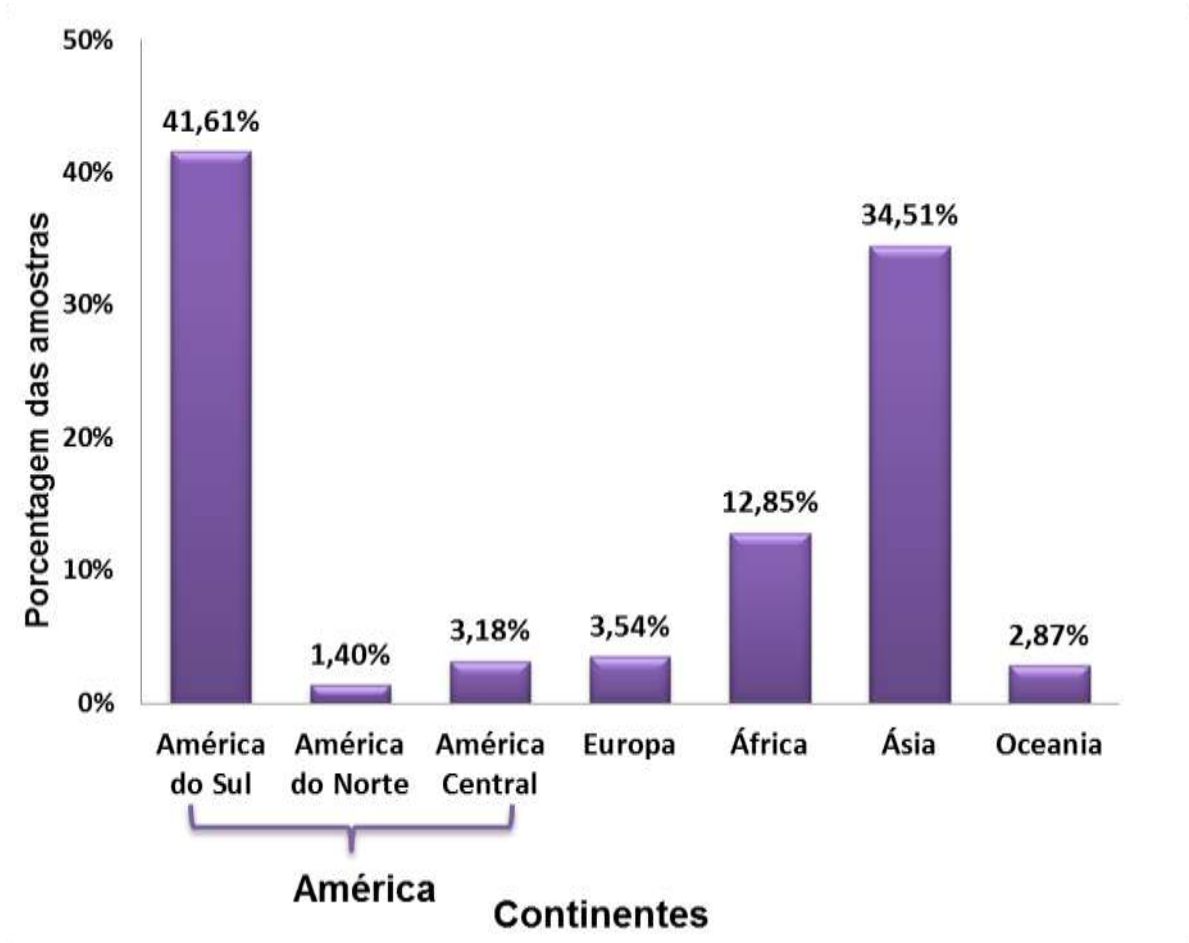


Figura 10. Distribuição geográfica das sequências adicionadas no banco de dados de acordo com os continentes.

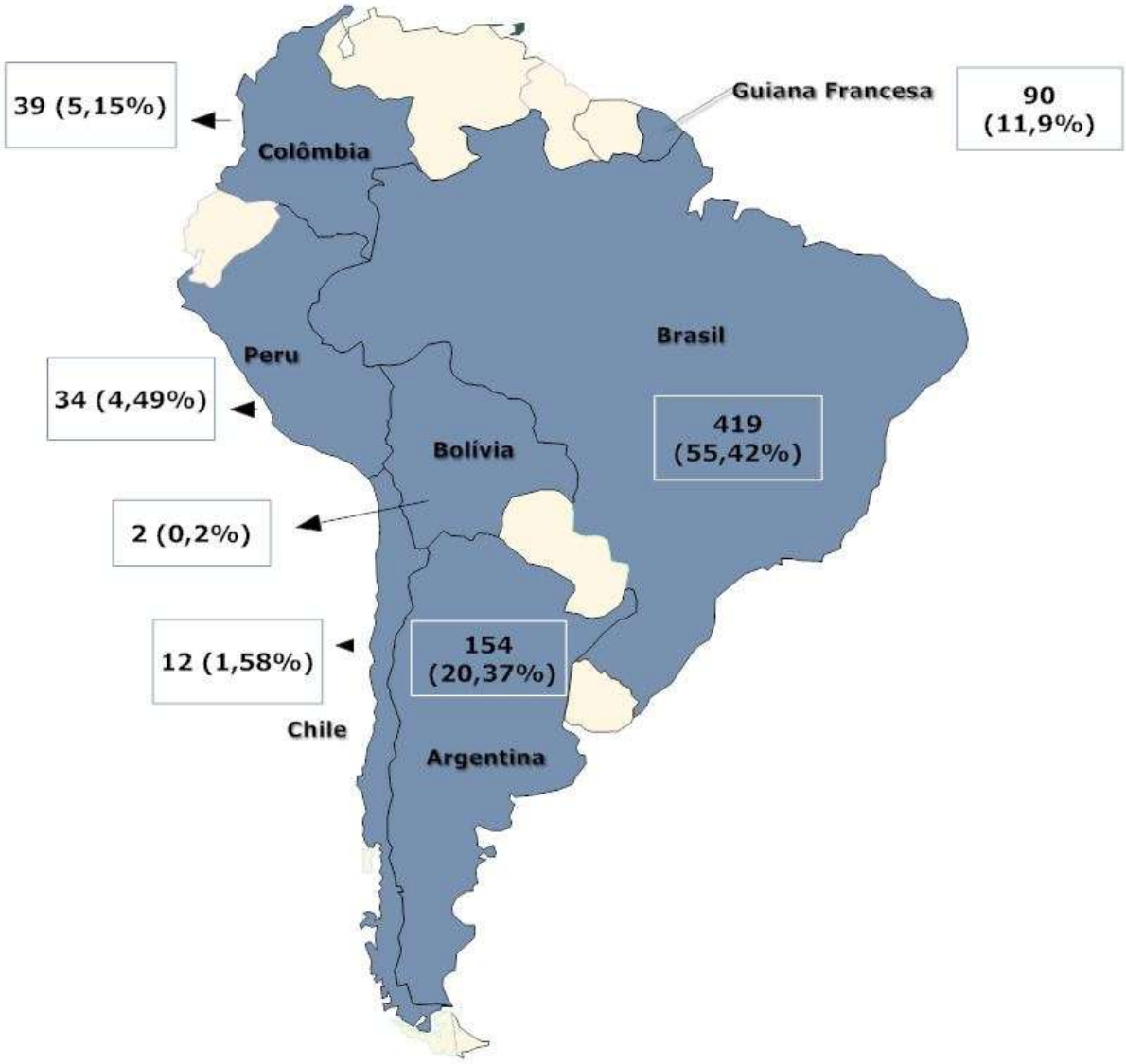


Figura 11. Distribuição geográfica das sequências de isolados da América do Sul adicionadas ao banco de dados.

Das 1968 sequências de isolados, 797 (40,49%) possuem informação sobre o perfil clínico dividido em: TSP/HAM 343/797 (43%), Assintomático 261/797 (32,7%), ATLL 149/797 (18,69%) e outras patologias associadas como dermatite infectiva, Síndrome de *Sjögren*, Síndrome *Sicca*, Síndrome Oculares entre outras 44/797 (5,61%) (Figura 12). Das 343 sequências de isolados com o perfil clínico TSP/HAM, 320 (93,29%) possuem informações sobre a região geográfica, distribuídas na Ásia 218/320 (68,12%), América do Sul 62/320 (19,37%), África 18/320 (5,62%), Europa 8/320 (2,5%) e América Central 10/320 (3,12%), América do Norte 4/320 (1,25%) (Figura 13). Das 149 sequências com o perfil clínico ATLL, 128 (85,9%) das

sequências possuem informação e estão distribuídas geograficamente: Ásia 84/128 (65,62%), América do Sul 12/128 (9,37%), África 12/128 (9,37%), América do Norte 10/128 (7,81%), Europa 7/128 (5,46%) e América Central 3/128 (2,34%) (Figura 14). Enquanto que 260/261 (99,61%) das sequências com o perfil clínico assintomático possuem informação relativa à distribuição geográfica: Ásia 109/260 (42,92%), América do Sul 76/260 (29,2%), África 62/260 (23,84%), América Central 9/260 (3,46%), Europa 2/260 (0,76%) e Oceania 2/260 (0,76%) (Figura 15). Todas as sequências com perfil clínico classificados em outras patologias associadas possuem informação sobre a região geográfica, distribuídas geograficamente em: Ásia 18/44 (40,9%), África 7/44 (15,9%), América Central 12/44 (27,2%), América do Sul 6/44 (13,6%) e Oceania 1/44 (2,27%) (Figura 16).

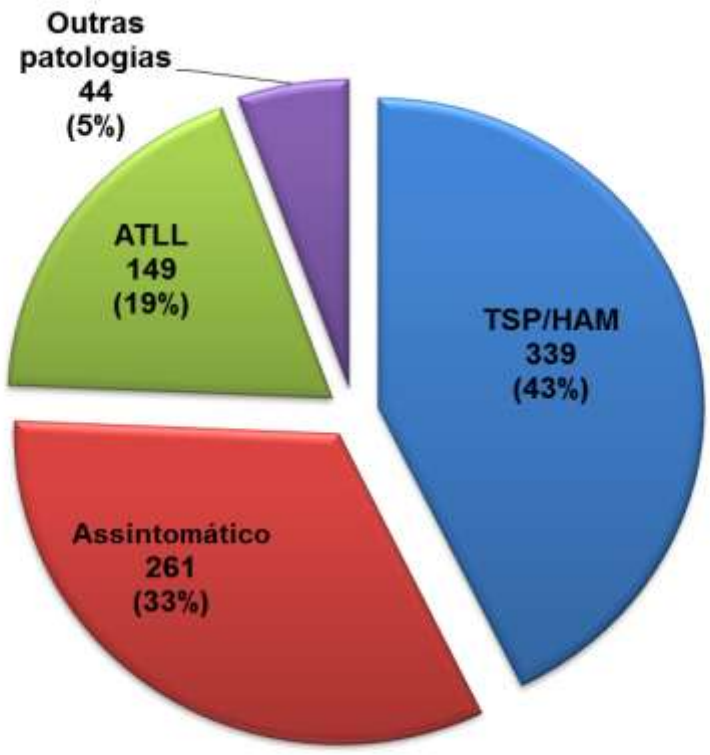


Figura 12. Distribuição em relação ao perfil clínico de 797 sequências de isolados adicionadas no banco de dados.

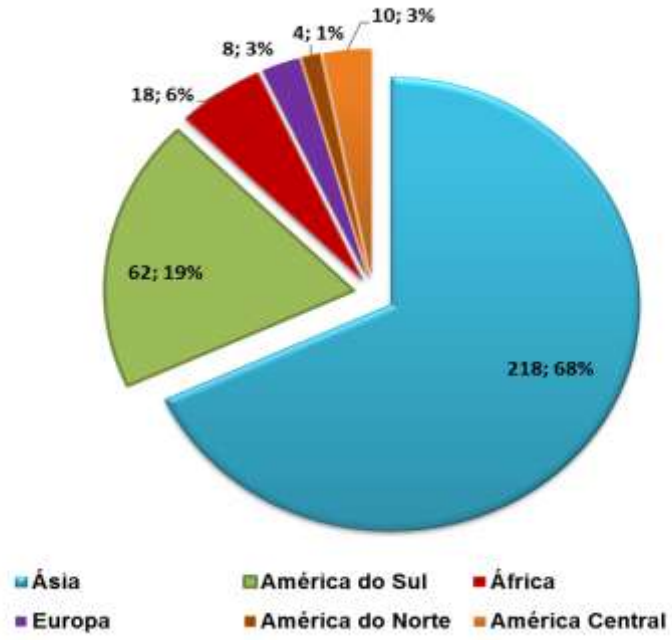


Figura 13. Distribuição geográfica de 320 sequência de isolados com o perfil clínico TSP/HAM (Paraparesia Espática Tropical/ Mielopatia Associada ao HTLV).

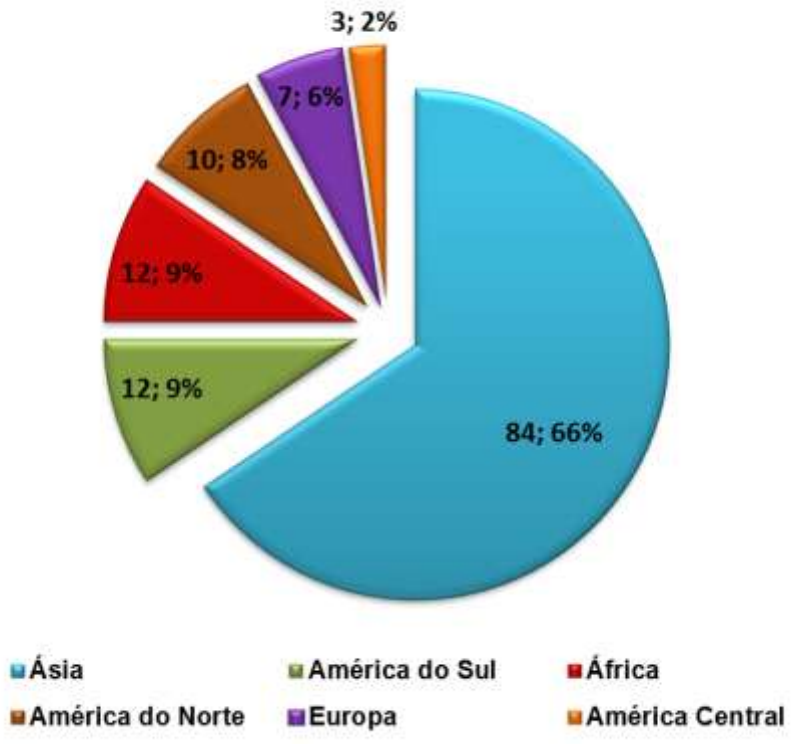


Figura 14. Distribuição geográfica de 128 sequências de isolados com o perfil clínico ATLL (Leucemia/linfoma de células T do adulto).

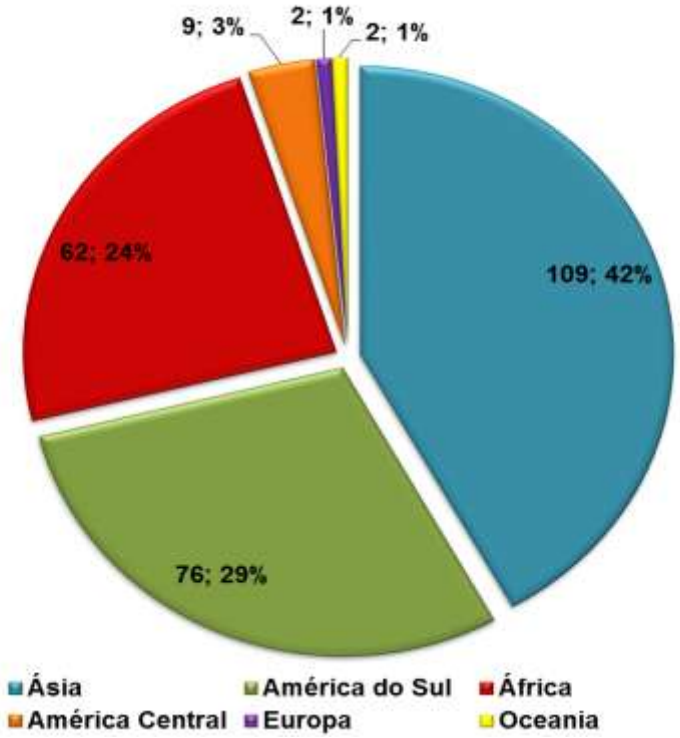


Figura 15. Distribuição geográfica de 260 sequências de isolados com o perfil clínico assintomático.

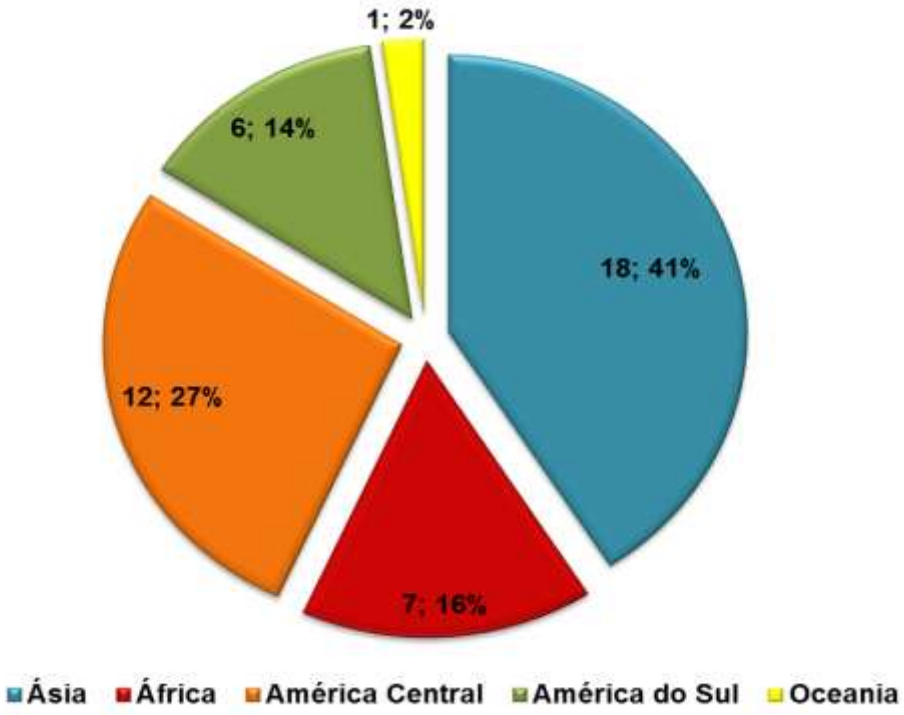


Figura 16. Distribuição geográfica de 44 sequências de isolados com o perfil clínico classificados como outras patologias associadas.

Em relação ao gênero, há informação de 298/1968 (15,14%) dos indivíduos nos quais as sequências dos isolados foram originárias, sendo 63,75% referentes ao gênero feminino. No banco de dados 158/1968 (8,02%) das sequências dos isolados possuem informações concomitantes sobre o gênero e perfil clínico, estratificadas em (Figura 17):

- Assintomático: Feminino (56) e masculino (33);
- TSP/HAM: Feminino (25) e masculino (15);
- ATLL: Feminino (6) e masculino (7);
- Outras patologias associadas: Feminino (8) e masculino (8).

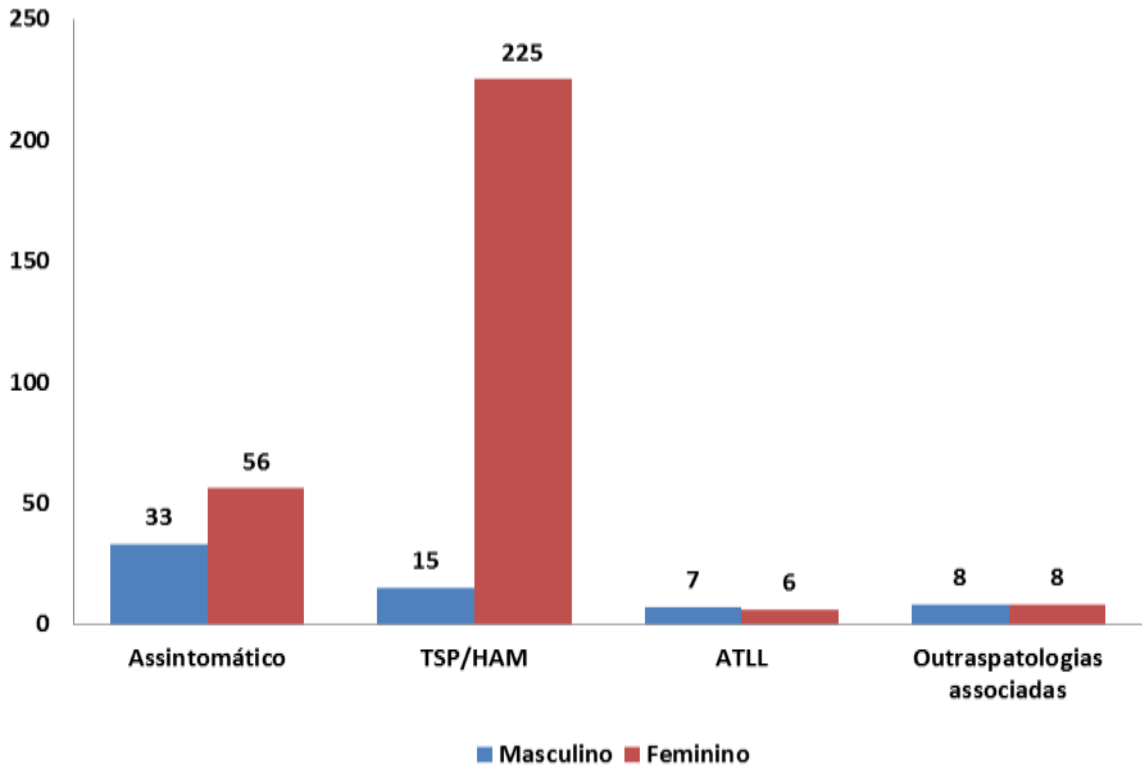


Figura 17. Perfil clínico e gênero de 158 das 1968 sequências de isolados analisadas.

Considerando o gênero e a região geográfica das sequências, 296/1968 (15,04%) possuem as duas informações, distribuídas: Europa 12/296 (4,05%), África 87/296 (29,39%), Ásia 27/296 (9,12%), América do Sul 112/296 (37,83%), América do Norte 6/296 (2,02%), América Central 9/296 (3,04%) e Oceania 43/296 (14,52%) (Figura 18). Em relação aos gêneros específicos, as sequências foram distribuídas geograficamente (Figura 19):

- Europa: Masculino (5) e Feminino (7);
- África: Masculino (33) e Feminino (54);
- Ásia: Masculino (10) e Feminino (17);
- Oceania: Masculino (14) e Feminino (29).
- América:
    - América do Sul: Masculino (44) e Feminino (68);
    - América do Norte: Masculino (0) e Feminino (6);
    - América Central: Masculino (1) e Feminino (8).

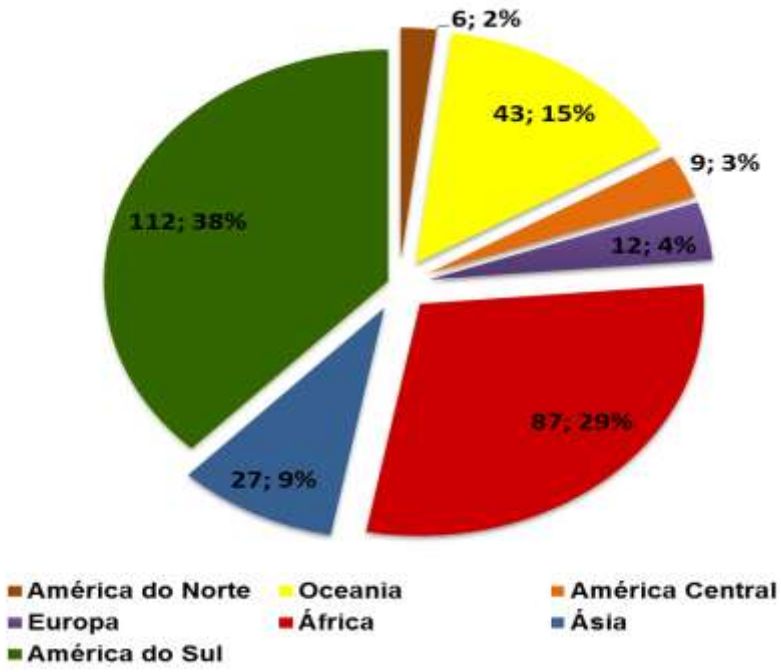


Figura 18. Distribuição geográfica de 296 sequências de isolados do banco de dados com informação em relação ao gênero.

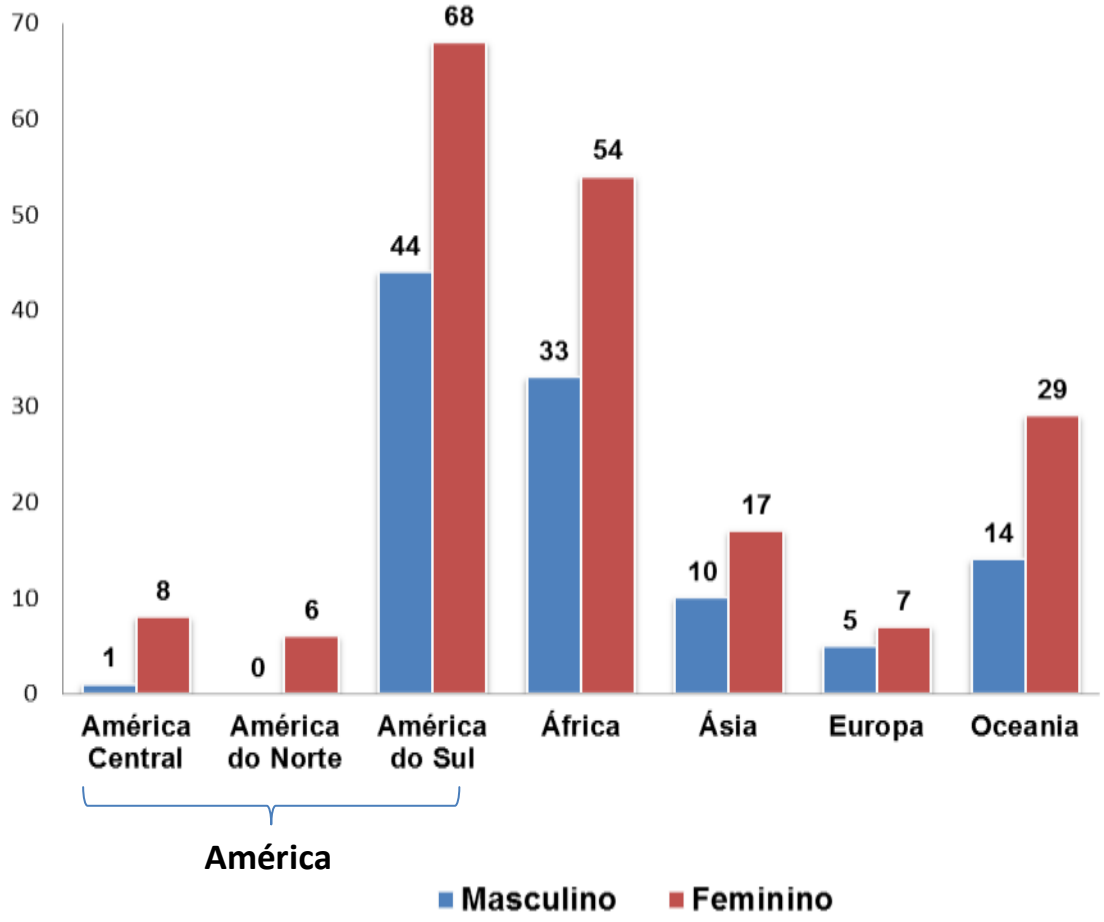


Figura 19. Distribuição geográfica de 296 sequências do banco de dados estratificadas de acordo com o gênero específico.

Levando-se em consideração somente a variável subtipo, 1040 (52,84%) sequências dos isolados têm informações. Considerando concomitantemente as variáveis subtipo e origem geográfica, das 1040, 1008 (96,92%) sequências dos isolados possuem informações, nos quais 892/1008 (88,49%) representam o suptipo "a", distribuído mundialmente, seguidas do subtipo "b" 98/1008 (9,72%), "c" 7/1008 (0,69%), "d" 8/1008 (0,79%), "e" 1/1008 (0,09%), "f" 1/1008 (0,09%) e "g" 1/1008 (0,09%), distribuídas em regiões mais específicas (Figura 20). Em relação à distribuição geográfica dos subgrupos do subtipo "a", 746/892 (83,63%) das sequências dos isolados apresentam informações: subgrupo "A" 661/746 (88,60%), "B" 62/746 (8,31%), "C" 15/746 (2,01%), "D" 4/746 (0,53%) e "E" 4/746 (0,53%) (Figura 21).

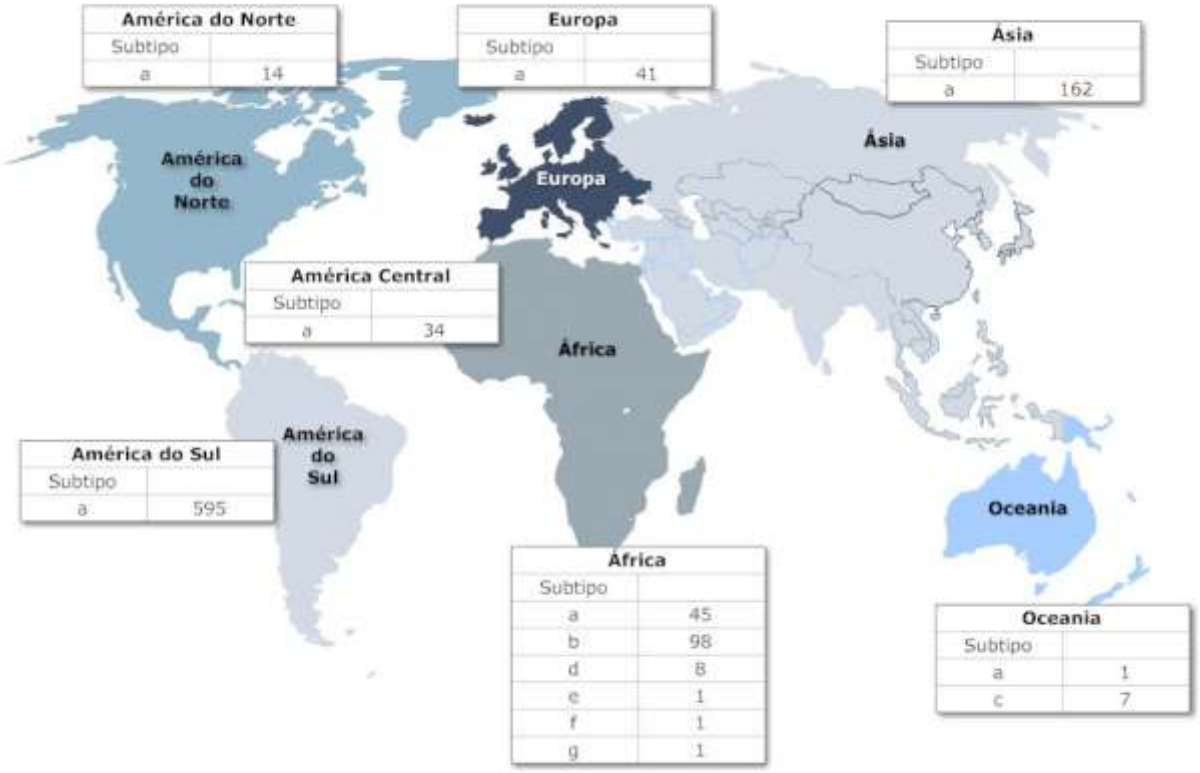


Figura 20. Distribuição geográfica dos subtipos do HTLV-1 de 1008 sequências do banco de dados.

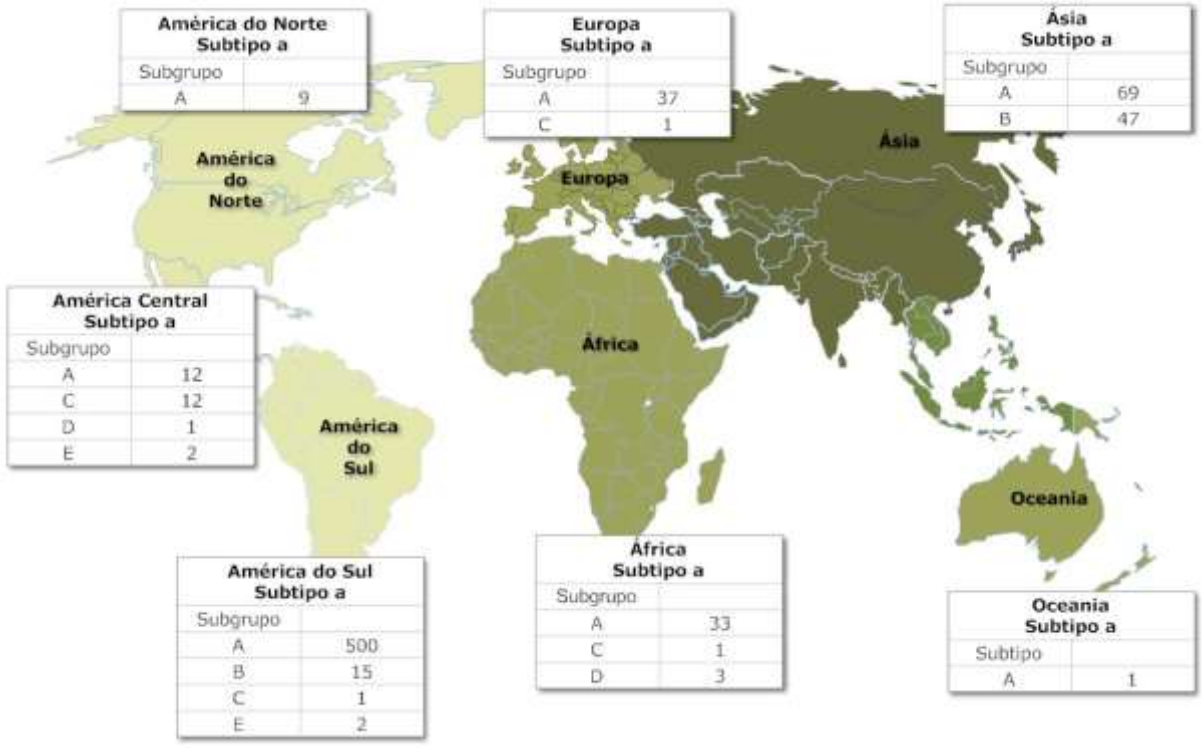


Figura 21. Distribuição geográfica dos subgupos do subtipo “a” do HTLV-1 de 746 sequências do banco de dados.

Considerando a idade, 208 (10,56%) das 1968 sequências dos isolados apresentam informações. As idades variaram de 3 a 100 anos, com uma mediana de 44 anos.O gênero feminino tem a maior representatividade nas informações com 57,7% das sequências dos isolados.

Nas sequências com o perfil clínico TSP/HAM, a mediana de idade foi de 47 anos. Enquanto com o perfil clínico ATLL, a mediana de idade foi de 48,5 anos.

Informações em relação à carga proviral (4/1968; 0,20%), CD4+ (13/1968; 0,66%), CD8+ (9/1968; 0,45%), etnia (62/1968; 3,15%) e data de coleta (87/1968; 4,42%) não foram representativas no banco de dados. Das 1968 sequências, nenhuma apresentou informações de todas as variáveis concomitantemente.

6.2 *HTLV-1 MOLECULAR EPIDEMIOLOGY DATABASE*

O banco de dados desenvolvido disponibiliza informações referentes às sequências indexadas no *GenBank*, no qual o usuário tem a possibilidade de escolher critérios de busca (Figura 23) de acordo com a sua necessidade. Após a escolha, o programa gera uma resposta que é mostrada em outra interface denominada *output*, no qual aparecerão as sequências e informações referentes à busca do usuário. Este poderá baixar as sequências no formato *fasta* e a tabela com as informações no formato *xls* (planilhas do Microsoft Excel) (Figura 24).

O Banco de dados de Epidemiologia Molecular do HTLV- 1 (HTLV-1 *Molecular Epidemiology Dabase*) está hospedado no servidor do Centro de Pesquisa Gonçalo Moniz/ FIOCRUZ–BA com o endereço para acesso: http//www.htlv1db.bahia.fiocruz.br.

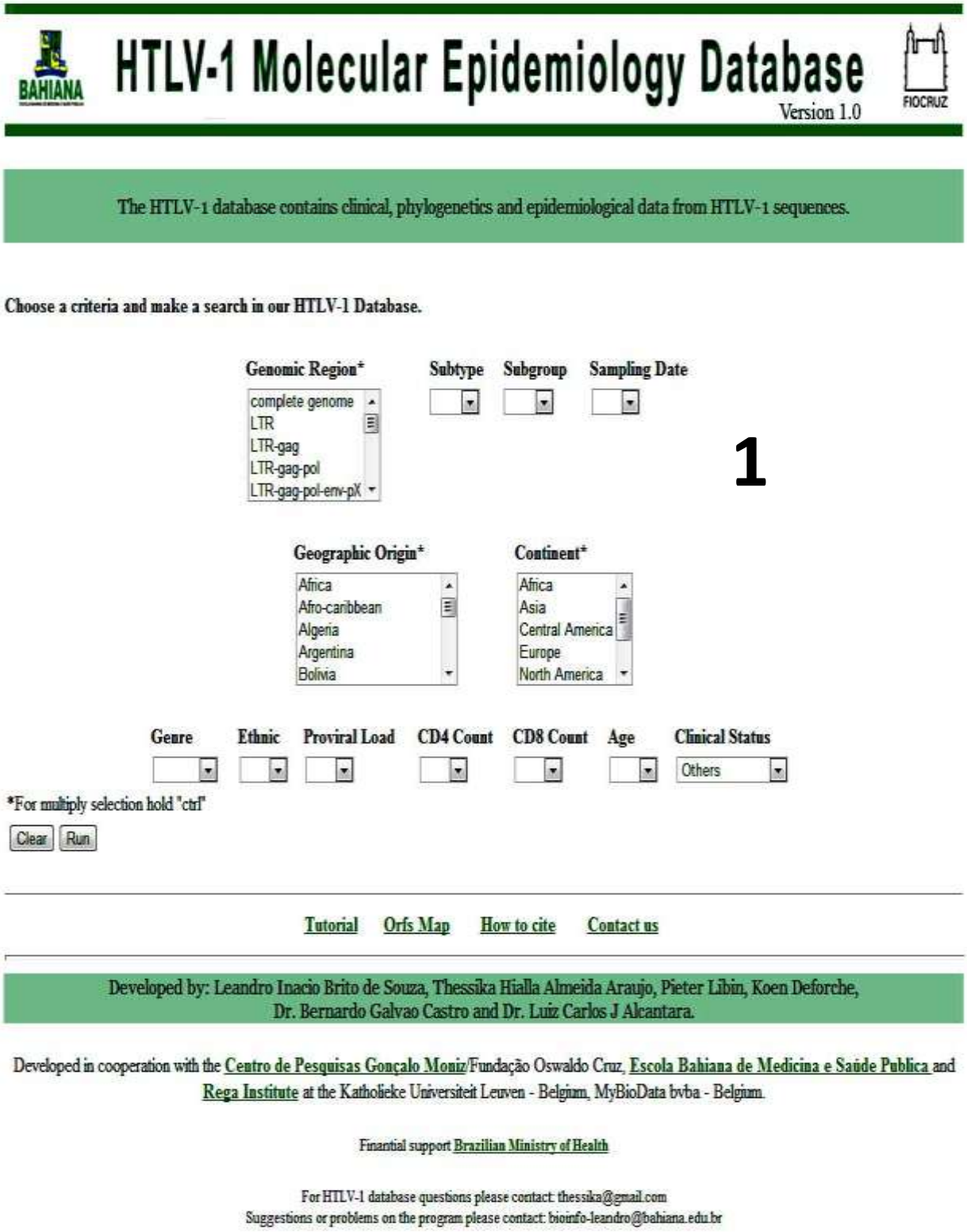
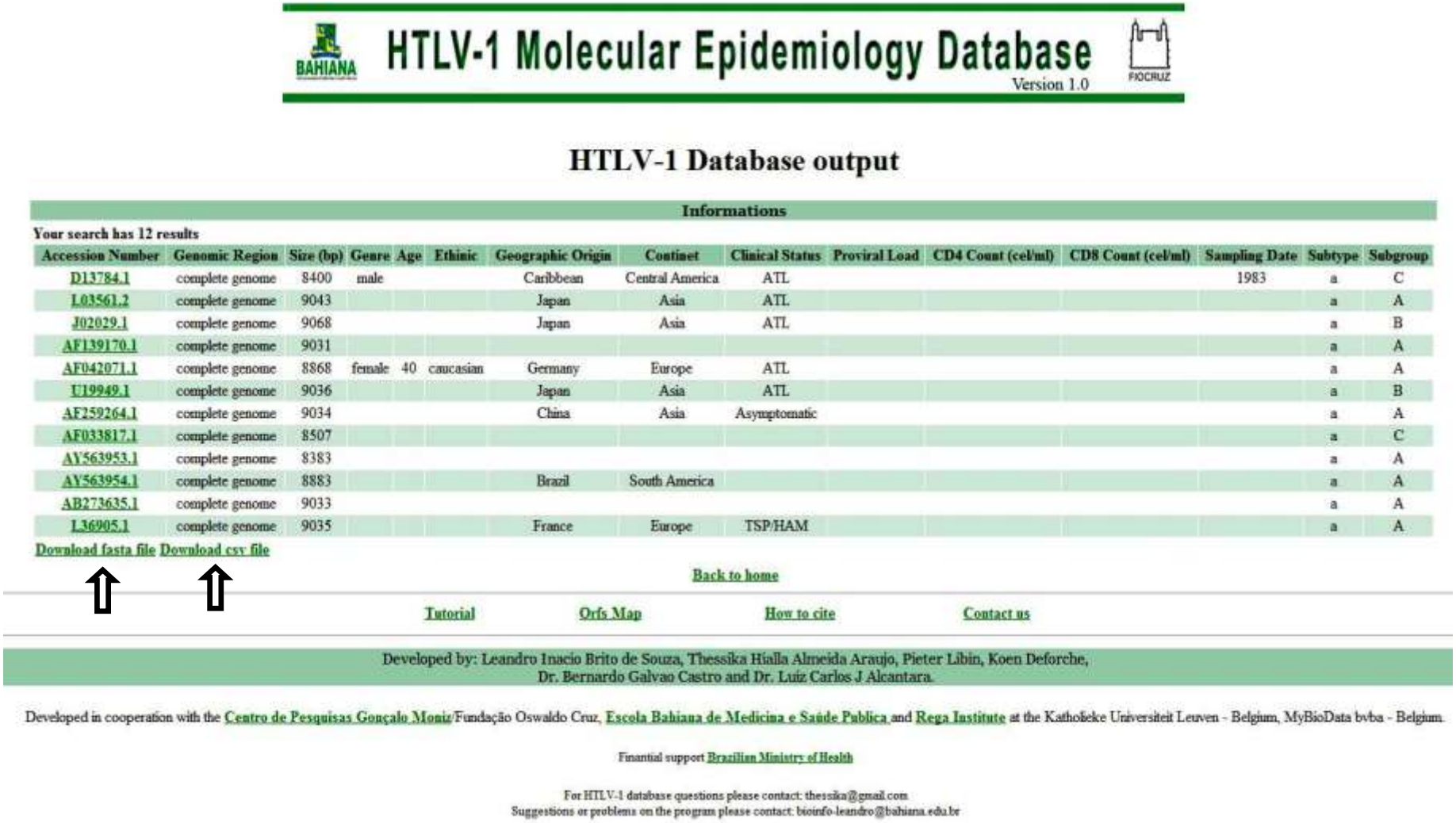


Figura 22. *Interface i*nicial do HTLV-1 *Molecular Epidemiology Database*. 1) Opção para a escolha dos critérios de busca das sequências.



HTLV-1 Molecular Epidemiology Database
Version 1.0

HTLV-1 Database output

Informations

Your search has 12 results

Accession Number	Genomic Region	Size (bp)	Genre	Age	Ethinic	Geographic Origin	Continet	Clinical Status	Proviral Load	CD4 Count (cel/ml)	CD8 Count (cel/ml)	Sampling Date	Subtype	Subgroup
D13784.1	complete genome	8400	male			Caribbean	Central America	ATL				1983	a	C
L03561.2	complete genome	9043				Japan	Asia	ATL					a	A
J02029.1	complete genome	9068				Japan	Asia	ATL					a	B
AF139170.1	complete genome	9031											a	A
AF042071.1	complete genome	8868	female	40	caucasian	Germany	Europe	ATL					a	A
U19949.1	complete genome	9036				Japan	Asia	ATL					a	B
AF259264.1	complete genome	9034				China	Asia	Asymptomatic					a	A
AF033817.1	complete genome	8507											a	C
AY563953.1	complete genome	8383											a	A
AY563954.1	complete genome	8883				Brazil	South America						a	A
AB273635.1	complete genome	9033											a	A
L36905.1	complete genome	9035				France	Europe	TSP/HAM					a	A

Download fasta file Download csv file

Back to home

Tutorial Orfs Map How to cite Contact us

Developed by: Leandro Inacio Brito de Souza, Thessika Hialla Almeida Araujo, Pieter Libin, Koen Deforche, Dr. Bernardo Galvao Castro and Dr. Luiz Carlos J Alcantara.

Developed in cooperation with the Centro de Pesquisas Gonçalo Moniz/Fundação Oswaldo Cruz, Escola Bahiana de Medicina e Saúde Publica and Rega Institute at the Katholieke Universiteit Leuven - Belgium, MyBioData bvba - Belgium.

Financial support Brazilian Ministry of Health

For HTLV-1 database questions please contact: thessika@gmail.com
Suggestions or problems on the program please contact: bioinfo-leandro@bahiana.edu.br

Figura 23. *Interface* de apresentação do resultado (*output)* com as opções de *download* das sequências no formato *fasta* e da tabela com as informações referentes à sequências.

6.3 MAPA E DESCRIÇÃO DAS *ORFS*

Nós construímos o mapa das *orfs* (Figura 25), considerando a sequência referência ATK-1. Este mapa está disponível para os usuários no banco de dados (*HTLV-1 Molecular Epidemiology Database)*, como guia de classificação da(s) região(ões) genômica(s) das sequências. Por exemplo, se a sequência começar no nucleotídeo de número 824 e terminar no número 2052, representa a região *gag* parcial.

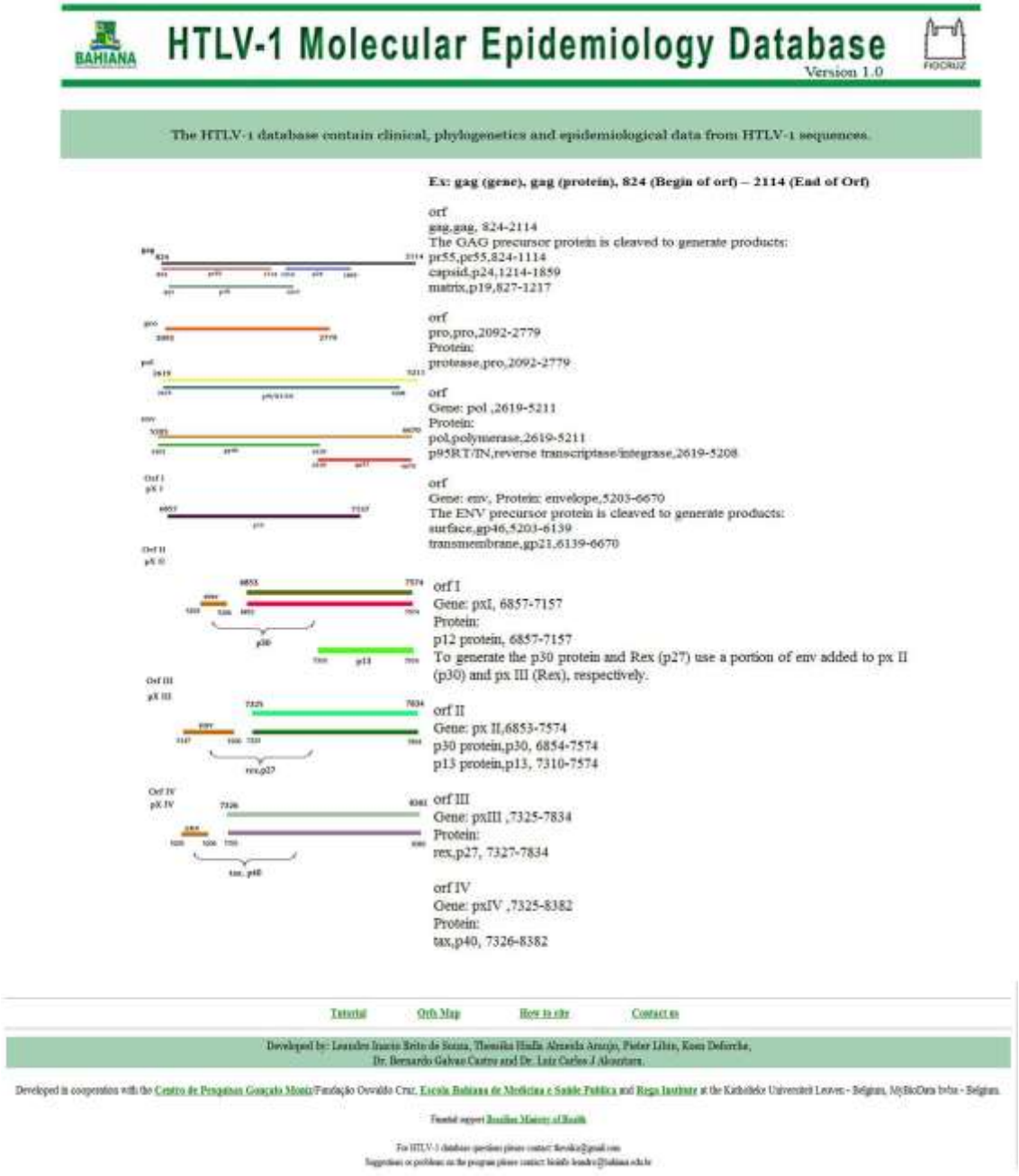


Figura 24. Mapa das *Orfs* do HTLV-1 disponível no *HTLV-1 Molecular Epidemiology Database.*

6.4 QUALIDADE DAS INFORMAÇÕES

Todos os dados coletados sejam as sequências de nucleotídeo do HTLV-1 e suas anotações tiveram como origem o *GenBank*, artigos respectivos ou informações diretamente adquiridas com os autores correspondentes. Estes pesquisadores desenvolvem os seus estudos em diversos centros de pesquisas no mundo inteiro, os quais realizam sequenciamento e anotações sobre as sequências, o que espera que ocorra divergências nas anotações e até mesmo erros no sequenciamento.

Levando-se em consideração que este trabalho propõe oferecer as informações coletadas no *GenBank*, artigos respectivos e diretamente com os autores em um modelo relacional, não são de responsabilidade do nosso estudo os eventuais erros de anotação e sequenciamento.

7 DISCUSSÃO

A mineração das sequências coletadas no *GenBank* foi importante para selecionar somente aquelas de interesse para o banco de dados proposto. Foram excluídos os clones e as sequências de pacientes que tinham na descrição o termo “HTLV-1”, porém a sequência era de outro vírus, por exemplo, nós excluímos aquelas que tinham na descrição o termo “HTLV-1”, entretanto, eram sequências do HIV.

A partir dos dados coletados realizou-se estatística descritiva, considerando as variáveis (região geográfica, idade, gênero, subtipo, subgrupo, entre outras), sendo que os valores foram apresentados em forma de gráfico, colocando em evidência as tendências, as ocorrências ocasionais e os valores mínimos e máximos dos dados. Os dados não foram submetidos a cálculos de significância estatística (nível-p) por serem oriundos de uma base de conveniência, o que não garante a representatividade da população, além de que o tamanho da amostra não irá se manter constante, visto que novas sequências e suas respectivas informações serão adicionadas ao banco de dados, quando estas forem publicadas no *GenBank.*

A disponibilidade no banco de dados de importantes variáveis ausentes no *GenBank*, tais como carga proviral, no qual estudos demonstram a sua relação com a patogênese da TSP/HAM, na predição do desenvolvimento da ATLL e outras doenças como uveíte e artrite reumatóide (AKIZUKI *et al,* 1987; GREETEN *et al,* 1998; NAGAI *et al,* 2001; SAKAI *et al,* 2001; OSAME *et al,* 2002; ONO *et al,* 1998; YAKOVA *et al,* 2005; PROIETTI, 2005), contagem de CD4+ e CD8+ como marcador de risco para doenças associadas ao HTLV-1 (GOON et al, 2003; KUBOTA *et al,* 2000; SANTOS *et al,* 2004), data de coleta, importante pra estimar o relógio molecular (molecular-clock) e para a realização de análises bayesianas e a etnia, para possíveis inferências de susceptibilidade e até mesmo de relações filogenéticas destes grupos étnicos, considerando a estabilidade genômica do HTLV-1 (*SONODA et al*, 2011; MIURA *et al,* 1994).

A utilização de um Sistema de Gerenciamento de Banco de Dados (SBDG) foi fundamental na organização, armazenamento e atualização das sequências e de suas respectivas informações. O banco de dados relacional formado, prover acesso facilitado aos dados, possibilitando que os usuários utilizem uma grande variedade de abordagens no tratamento das informações, uma vez que é possível a escolha em vários pontos ou critérios.

A fácil usabilidade foi priorizada nas *interfaces* criadas, nas quais os usuários tenham um uso simplificado, considerando-se o acesso de usuários mais ou menos experientes, e que a sua utilização seja eficaz, produtiva, além de segura.

Os bancos de dados biológicos são importantes ferramentas para auxiliar pesquisadores na compreensão, avaliação e comparação dos dados gerados a partir das suas pesquisas, entretanto, devido a grande quantidade de conhecimento biológico gerado, tão qual a sua distribuição entre diferentes bases de dados gerais ou especializadas, como por exemplo, o *GenBank,* torna-se difícil assegurar a coerência e fidedignidade destas informações. O banco de dados desenvolvido neste trabalho se apresenta com a proposta de concentrar o máximo de informações relacionadas às sequências do HTLV-1, diminuindo a redundância e preservando a consistência desses dados, sejam estes oriundos do *GenBank* ou dos artigos relacionados às sequências.

Os bancos de dados biológicos como o *HCV Sequence Database* (KUIKEN *et al,* 2005)*, HIV Drugs Resistance Database* - Stanford (http://hivdb.stanford.edu/) e *HIV Databases – Los Alamos (http://www.hiv.lanl.gov/content/index)* constituem-se

como importantes bases de dados para as pesquisas. Visto que, além de disponibilizarem ferramentas para análises específicas, como as de resistência aos antirretrovirais, no caso do HIV, reúnem uma grande quantidade de informações geradas em todo o mundo, o que permite traçar um cenário global da infecção desses vírus, favorecendo o conhecimento clínico e epidemiológico. Além disso, estes dados são eficazes para decisões políticas públicas em saúde e para planejamento de programas de combate à propagação das infecções e de assistências aos pacientes.

A coleta das sequências e suas anotações foram feitas inicialmente no *GenBank*, tendo-se o cuidado em avaliar estas anotações porque segundo Guimarães (2006), um dos problemas mais comuns relatados no processo de anotação estão relacionados com erros/falta das informações nos bancos de dados públicos e a falta de padronização dos vocabulários de anotação. Para completar as informações do banco de dados, recorreu-se aos artigos das respectivas sequências, que na sua grande maioria traziam informações como gênero, perfil clínico, idade, origem geográfica, entre outros, ausentes nas anotações do *GenBank,* ou consultou-se diretamente com os autores correspondentes.

Considerando as dificuldades encontradas na aquisição das informações, ressalva-se a importância do fornecimento e disponibilidade pública dos dados gerados, respeitando a propriedade intelectual, quando estes forem factíveis para o conhecimento científico sobre o HTLV-1.

A construção e disponibilidade de um banco de dados com o máximo possível de informações sobre o HTLV-1 irá favorecer toda a comunidade científica que se interessa e que pesquisa o tema, logo se torna cada vez mais importante o desenvolvimento de bancos de dados biológicos completos ou fazer com que os existentes se tornem completos, com intuito de facilitar o acesso aos dados, e por consequência, favorecer os estudos ligados à infecção, patogênese, origem e evolução do vírus. O gerenciamento e atualização dos dados serão feitos localmente no servidor do Centro de Pesquisa Gonçalo Moniz, com uma periodicidade quinzenal.

8 CONCLUSÕES

- ✓ Através da interface principal, os usuários terão acesso às informações, tais como as sequências, de acordo com os critérios de busca selecionados.

- ✓ A partir das informações inseridas no banco de dados é possível ao usuário realizar uma análise descritiva das sequências oriundas do *GenBank.*

- ✓ Ressalva-se a escassez de informações completas sobre as sequências, ressaltando a importância da disponibilidade do maior número de dados, importantes para as pesquisas sobre HTLV-1.

REFERÊNCIAS

ALCANTARA, LCJ; VAN DOOREN, S; GONÇALVES, MS; KASHIMA, S; COSTA, MC; SANTOS, FL; BITTENCOURT, AL; DOURADO, I; FILHO, AA; COVAS, DT; VANDAMME, AM; GALVÃO-CASTRO, B. Globin haplotypes of human T-cell lymphotropic virus type I-infected individuals in Salvador, Bahia, Brazil, suggest a post-Columbian African origin of this virus. **J Acquir Immune Defic Syndr** 33:536-542, 2003.

ALCANTARA, LC; DE OLIVEIRA, T; GORDON, M; PYBUS, O; MASCARENHAS, RE; SEIXAS, MO; GONÇALVES, M; HLELA, C; CASSOL, S; GALVÃO-CASTRO, B. Tracing the Origin of Brazilian HTLV-1 as determined by Analysis of Host and Viral genes. **AIDS** 20:780-782, 2006.

AKISUKI S., NAKAZATO, O., HIGUCHI Y., *et al*. Necropsy findings in HTLV-I associated myelopathy. **Lancet**. v.1, p.156, 1987.

BASTIAN I., J.; GARDNER, D. WEBB; I. GARDNER. Isolation of a human T-lymphotropic virus type I strain from Australian aboriginals. **Journal of Virology.** 67:843-51, 1993.

BENSON, D. A.; MIZRACHI, I. K.; LIPMAN, D. J.; OSTELL, J.; SAYERS, Eric W. *GenBankGenBank.* **Nucleic Acids Research***,* v. 38, 2010.

BIOINFORMATICS FACTSHEE. Disponível em http://www.ncbi.nlm.nih.gov /About/primer/bioinformatics.html. Acessado em 12 de Abril de 2011.

BORNBERG-BAUER, E. and PATON, N.W. Conceptual data modeling for bioinformatics. **Briefings in Bioinformatics** . v 3. n 2. 166–180, 2002.

BURKE, D.S. Recombination in HIV: An important evolutionary strategy. **Emergent Infectious Diseases, 3:** 253-258, 1997.

CALATTINI S, CHEVALIER SA, DUPREZ R, BASSOT S, FROMENT A, MAHIEUX R, GESSAIN A.Discovery of a new human T-cell lymphotropic virus (HTLV-3) in Central Africa. **Retrovirology**. 9;2:30, 2005.

CATALAN-SOARES, B; CARNEIRO-PROIETTI, A.B.; PROIETTI F.A. Vírus-T linfotrópico humano em familiares de candidatos a doação de sangue soropositivos: disseminação silenciosa. **Ver. Panam Salud Publica/Pan Am J Public Health.** v. 16, n.6, p. 387-394, 2004

CHEN, J., L. ZEKENG, M. YAMASHITA, J. TAKEISHA, T. MIURA, E.IDO, I. MBOUDJEKA, J.M.TSAGUE, M. HAYAMI ; L. KAPTUE. HTLV isolated from a Pygmy in Cameroon is related but distinct from the known Central African type*. AIDS* **Research and Human Retroviruses** .11:1529-1531,1995.

CHEN, J. Y., Carlis, J. V., and Gao, N. **A complex biological database querying method**. In SAC, pages 110–114. ACM., 2005.

CIMINALE V, PAVLAKIS GN, DERSE D, CUNNINGHAM CP, FELBER BK. Complex splicing in the human T-cell leukemia virus (HTLV) family of retroviruses: novel mRNAs and proteins produced by HTLV type 1. **J Virol**. 1992: 66-1737-1745.

DE THE G, KAZANJI M. An HTLV-I/II vaccine: from animal models to clinical trials? **J Acquir Immune Defic Syndr Hum Retrovirol.** v. 13, 1996.

DOURADO, I; ALCANTARA, LCJ; BARRETO, ML; DA GLORIA, TM; GALVAO-CASTRO, B. HTLV-I in the general population of Salvador, Brazil: a city with African ethnic and sociodemographic characteristics. **J Acquir Immune Defic Syndr** 34(5):527-531, 2003.

EDWARDS, D; FENIZIA, C; GOLD, H.; CASTRO-AMARANTE, MF; BUCHMANN, C; PISE-MASISON CA; FRANCHINI, G. HTLV-1 *Orf-I* and *Orf-II*-Encoded Proteins in Viral Infection and Persistence**. Viruses ,** *3*, 1-x manuscripts;, 2011

ELMASRI, Ramez E. & NAVATHE, Shamkant. **Sistemas de banco de dados.** 4. ed. Addison-Wesley, 2005.

FILHO, FC; PROSDOCIMI, F; CERQUEIRA, GC; BINNECK, E; SILVA, AF; REIS, A N; JUNQUEIRA, ACM, SANTOS, ACF; *et al.* **Bioinformática: Manual do Usuário**. Disponível em: <http://www.biotecnologia.com.br/revista/bio29/bioinf.pdf>. Acessado em: 11 de julho de 2010.

FUKUMOTO R, ANDRESEN V, BIALUK T *et al*. In vivo genetic mutations define predominant functions of the human T-cell leukemia/lymphoma virus p12I protein. **Blood**.113:3726-3734, 2009.

GADELHA, S.R. e cols. Ethnic Differences in the Distribution of Interleukin-6 Polymorphisms Among Three Brazilian Ethnic Groups. **Human Biology.** v. 77. n. 4, p. 509-514, 2005.

Gallo RC. Kyoto Workshop on some specific recent advances in human tumor virology. **Cancer Res.**41**:**4738-4739, 1981.

GASMI, M., B. FAROUQI, M. D'INCAN, C. Desgranges. Long terminal repeat sequence analysis of HTLV type I molecular variants identified in four north African patients. ***AIDS Research and Human Retroviruses***. 10:1313-15, 1994.

GESSAIN A, BARIN F, VERNANT JC, GOUT O, MAURS A, CALENDER A, DE THÉ G. Antibodies to human T-lymphotropic virus type I in patients with tropical spastic paraparesis. **Lancet.,** 2: 407-409, 1985.

GESSAIN A, YANAGIHARA R, FRANCINI G, *et al*. Highly divergent molecular variants of human T-lymphotropic virus type from isolated populations in Papua New Guinea and the Solomon Islands. ***Proc NATLL Acad Sci U S A***. 88:7694-7698, 1991.

GESSAIN, A. & DE THE,G. Geographic and molecular epidemiology of primate T lymphotropic retroviruses: HTLV-I, HTLV- II, STLV-I, STLVPP and PTLV-L. **J Acq Immun Def Synd Hum Retrovirol.** v. 13, Suppl 1, S228 – S235, 1996.

GOON PK, IGAKURA T, HANON E, MOSLEY AJ, ASQUITH B, GOULD KG, TAYLOR GP, WEBER JN, BANGHAM CR. High circulating frequencies of tumor

necrosis factor alpha- and interleukin-2-secreting human T-lymphotropic virus type 1 (HTLV-1)-specific CD4+ T cells in patients with HTLV-1-associated neurological disease. **J Virol.** 77(17):9716-22, 2003.

GONÇALVES, DU; GUEDES, AC; PROIETTI, ABFC. Dermatologic lesions in asymptomatic blood donors seropositive for Human T-cell lymphotropic virus type-1**. American Journal of Tropical Medicine and Hygiene.** 68;562-565, 2003.

GRANT W, BIA FJ, CHACKO TM, JEAN-BAPTISTE M, GRIFFITH BP. Comparison of enzyme-linked immunosorbent and indirect immunofluorescence assays for the detection of human T-cell lymphotropic virus type-I antibodies in sera from rural Haiti. **Diagn Microbiol Infect Dis.** Feb;15(2):121-4, 1992.

GREEN P.L; CHEN SY.Human T-cell leukemia virus types 1 and 2. In DM Knipe, PM Howley (eds), **Fields Virology**, 4th ed., Lippincott Williams & Wilkins,Philadelphia, PA, 2001.

GREETEN T.F., SLASKY J.E., KUBOTA R., *et al*. Direct visualization of antigen-specific T cells: HTLV-1 Tax11-19- specific CD8(+) T cells are activated in peripheral blood and accumulate in cerebrospinal fluid from TSP/HAM patients. **Proc. NATLL. Acad. Sci.** USA. v.95,1998.

GUIMARÃES, Ana Carolina Ramos. **Identificação, classificação e anotação de enzimas análogas em tripanossomatídeos.** 106 f. Dissertação (Mestrado em Ciências). Pós-Graduação em Biologia Celular e Molecular. Instituto Oswaldo Cruz, Rio de Janeiro, 2006.

HAHN, B.H., G.M. SHAW, M. POPOVIC, *et al*. Molecular cloning and analysis of a new variant of human T-cell leukemia virus (HTLV-Ib) from an African patient with adult T-cell Leukemia-lymphoma. **International Journal of Cancer**. 34(5):613-618, 1984.

HANCHARD B, GIBBS WN, LOFTERS W, CAMPBELL M, WILLIAMS E, WILLIAMS N, JAFFE E, CRANSTON B, PANCHOOSINGH LD, LAGRENADE L, WILKS R, MURPHY E, BLATTNER W AND MANNS A. **Human Retrovirology: HTLV**. Blattner WA (ed). Raven Press: New York, pp. 173–183, 1990.

JEANMOUGIN, F; THOMPSON, JD; GOY, M; HIGGINS, DG; GIBSON, TJ. Multiple sequence alignment with clustal X. **Trends Biochem Sci** 23, 403-405, 1998.

KALYANARAMAN, V.S., M.G. SARNGADHARAN, M. ROBERT-GUROFF, *et al*. A new subtype of human T-cell leukemia virus (HTLV-1I) associated with a T-cell variant of hairy cell leukemia. **Science**. 218:571-573, 1982.

KAJIYAMA, W., KASHIWAGI, S., IKEMATSU, H., HAYASHI, J., NOMURA, H. & OKOCHI, K. Intrafamilial transmission of adult T-cell leukemia virus. **J Infect Dis** .154**,** 851-857, 1986.

KORALNIK IJ, FULLEN J, FRANCHINI G. The p12I, p13II, and p30II proteins encoded by human T-cell leukemia/lymphotropic virus type I open reading frames I and II are localized in three different cellular compartments. **J Virol**.;67:2360-2366, 1993.

KORALNIK IJ, LEMP JF JR, GALLO RC, FRANCHINI G. In vitro infection of human macrophages by human T-cell leukemia/lymphotropic virus type I (HTLV-I). **AIDS Res Hum Retroviruses**; 8:1845-1849, 1992.

KUBOTA R, NAGAI M, KAWANISHI T, OSAME M, JACOBSON S. Increased HTLV type 1 tax specific CD8+ cells in HTLV type 1-asociated myelopathy/ tropical spastic paraparesis: correlation with HTLV type 1 proviral load**. AIDS Res Hum Retroviruses.**Nov 1;16(16):1705-9, 2000.

KUIKEN C, YUSIM K, BOYKIN L, RICHARDSON R. HCV sequence database: The Los Alamos HCV Sequence Database. **Bioinformatics.** 21(3):379-84, 2005.

LA GRENADE L, MANNS A, FLETCHER V et al. Clinical, pathologic, and immunologic features of human Tlymphotropic virus type I-associated infective dermatites in children. **Arch Dermatol .**134:439-44,1998.

LEHESRAN J; DELAPORTE E; GAUDEBOUT C; Demographic factors associated with HTLV-1 infection in a Gabonese community.**Int J Epidemiol**. 23:812-17, 1994.
LEMOS, M. *Workflow para Bioinformática*. **PhD thesis**, Departamento de Informática da PUC-Rio,2004.

LIFSCHITZ, S.Algumas pesquisas em banco de dados e bioinformática. In: Workshop de Biologia Computacional. **Anais do XXVI Congresso da SBC**. Campo Grande, MS, 2006.

LOUREIRO P *et al.* Carga proviral do HTLV-1 em doadores de sangue, pacientes leucemia/linfoma T do adulto e paraparesia espática tropical/mielopatia associada ao HTLV-1 em Pernambuco. In: Congresso Brasileiro de Hematologia e Hemoterapia, 30.,São Paulo, **Anais: Sociedade Brasileira de Hematologia e Hemoterapia.p362, 2007.**

MAHIEUX R, IBRAHIM F, MAUCLERE P, HERVE V, MICHEL P, TEKAIA F, CHAPPEY C, GARIN B, VAN DER RYST E, GUILLEMAIN B, LEDRU E, DELAPORTE E, DE THE G, GESSAIN A. 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. **J Virol.**71(2):1317-33, 1997.

MARTINS, ML; STANCIOLI, EFB. Vírus linfotrópico de células T humanas Tipos 1 e 2 (HTLV – 1/2) – patogênese da infecção pelo HTLV. **Cadernos Hemominas. HTLV.** Volume XIII, 4ª edição, Belo Horizonte, p.21-45, 2006.

MCCALLUM, R.M., PATEL, D.D., MOORE, J.O. & HAYNES, B.F. Arthritis syndromes associated with human T cell lymphotropic virus type I infection. **Medical Clinics of North America.** v**.**81:261-276, 1997.

MIURA, T; FUKUNAGA, T; IGARASHI, T; YAMASHITA, M; IDO, E; FUNAHASHI, S; ISHIDAC, T; WASHIO, K; UEDAC, S; HASHIMOTO, K; YOSHIDA, M; OSAME, M; SINGHAL, BS; ZANINOVIC, V; CARTIER, L; SONODA, S; TAJIMA, K; INA, Y; GOJOBORI, T; HAYAMI, M. Phylogenetic subtypes of human T-lymphotropic virus type I and their relations to the anthropological background. **Proc NATLL Acad Sci USA** 91:1124-1127, 1994.

MIURA, T., M. YAMASHITA, V. ZANINOVIC, L. CARTIER, J. TAKEHISA, T. IGARASHI, E. IDO, T. FUJIYOSHI, S. SONODA, K.TAJIMA; M. HAYAMI. Molecular phylogeny of human T-cell leukemia virus type I and II of Amerindians in Colombia and Chile. **Journal of Molecular Evolution**. 44:S76-S82, 1997.

MOCHIZUKI M, WATANABE T, YAMAGUCHI K, *et al*. Uveitis associated with human T lymphotropic virus type I: seroepidemiologic, clinical, and virologic studies. **J. Infect. Dis**. 166:943–944, 1992.

MOROFUJI-HIRATA M, KAJIYAMA W, NAKASHIMA K, NOGUCHI A, HAYASHI J, KASHIWAGI S. Prevalence of antibody to human T-cell lymphotropic virus type I in Okinawa, Japan, after an interval of 9 years. **Am J Epidemiol.** Jan 1;137(1):43-8, 1993.

MALONEY, E.M., CLEGHORN, F.R., MORGAN, O.S., RODGERS-JOHNSON, P., CRANSTON, B., JACK, N., BLATTNER, W.A.,BARTHOLOMEW, C. & MANNS, A. Incidence or HTLV-I associated myelopathy/tropical spastic paraparesis)HAM/ TSP) in Jamaica and Trinidad. **J Acquir Immune Defic Syndr Hum Retrovirol**.17**,** 167-70, 1998.

MUELLER, N. The epidemiology of HTLV-1 infection.**Cancer Causes Control**. 2:37-52, 1991.

MUELLER, N.;OKAYAMA, A.; STUVER, S; TACHIBANA, N. Findings from the Miyazaki Cohort Study. **J. Acq Immun Def Synd Hum Retrovirol** 13 Suppl 1, S2-S7, 1996.

MURPHY, E.L., FIGUEROA, J.P., GIBS, W.N., HOLDING-COBHAM, M., CRANSTON, B., MALLEY, K. & BODNER, A.J. Human T lymphotropic virus type I (HTLV-I) seroprevalence in Jamaica.I. **Demographic determinants Am J Epid** 1991;133**,** 1114-1124,1991.

NAGAI M, KUBOTA R, GRETEN TF, SCHNECK JP, LEIST TP & JACOBSON S. Increased activated human T-cell lymphotropic virus type I (HTLV-1) tax 11-19-specific memory and effector CD8+ cells in patients with HTLV- 1-associated myelopathy / tropical spastic paraparesis: correlation with HTLV-1 provirus load. **J. Infect Dis.** 183: 197-205, 2001.

NICHOLAS, KB; NICHOLAS, HBJ, DEERFIELD, DW. GeneDoc: analysis and visualization of genetic variation. **Embnew News** 4, 14, 1997.

NISHIOKA, K. HTLV-1 arthropathy and Sjögren syndrome. Journal of Acquired Immune Deficiency Syndromes and Human. **Retrovirology .**13(S1):57-62, 1996.

ONO A et al. Provirus load in patients with human T-cell leukemia virus type 1 uveitis correlates with precedent Gravees´ disease and disease activities. **Jpn. J. Cancer Res.**89: 608-614,1998.

ORLAND JR, ENGSTROM J, FRIDEY J, SACHER RA, SMITH JW, NASS C, GARRATTY G, NEWMAN B, SMITH D, WANG B,LOUGHLIN K, MURPHY E. Prevalence and clinical features of HTLV neurologic disease in the HTLV Outcomes Study. **Neurology** .61:1588-1594, 2003.

OSAME M, USUKU K, IZUMO S, IJICHI N, AMITANI H, IGATA A, MATSUMOTO M, TARA M. HTLV-I associated myelopathy, a new clinical entity. **Lancet** .1: 1031-1032, 1986.

OSAME M. Pathological mechanisms of human T-cell lymphotropic virus type I-associated myelopathy (TSP/HAM). **J Neurovirol**. ;8(5):359-64, 2002.

POIESZ, BERNARD J.; RUSCETTI, FRANCIS W.; GAZDAR, ADI F. *et al*. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. **Proc NATLL Acad Sci USA**.77(12):7415–7419, 1980.

PORTO AF, OLIVEIRA FILHO J, NEVA FA, ORGE G, ALCANTARA L, GAM A, CARVALHO EM. Influence of human T-cell lymphocytotropic virus type 1 infection on serologic and skin tests for strongyloidiasis. **Am J Trop Med Hyg**.65(5), 610-3, 2001.

PORTO AF, SANTOS SB, ALCANTARA L, GUERREIRO JB, PASSOS J, GONZALEZ T, NEVA F, GONZALEZ D, HO JL, CARVALHO EM. HTLV-1 modifies the clinical and immunological response to schistosomiasis. **Clin Exp Immunol**.137(2), 424-9, 2004.

PROIETTI, FA;CARNEIRO-PROIETTI ABF; CATALAN-SOARES, BC; MURPHY, EL. Global epidemiology of HTLV-1 infection and associated disease. **Oncogene.** 24:6058-6068, 2005.

PROSDOCIMI, F. et al. Bioinformática: Manual do Usuário. **Biotecnologia Ciência & Desenvolvimento.** n. 29, 2002.

QUEIROZ, Alexandre. **Apostila de Introdução à Bioinformática.** Rio Grande do Norte: UFRN, 2002.

SABOURI, A.H. e cols. Differences in viral and host genetic risk factors for development of human T-cell lymphotropic virus type 1 (HTLV-1)-associated myelopathy/tropical spastic paraparesis between Iranian and Japanese HTLV-1-infected individuals. **J Gen Virol**. v.8, p.509-513, 2005.

SALEMI M, VAN DOOREN S, AUDENAERT E, *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**. 246:277-287, 1998.

SAKAI JA, NAGAI M, BRENNAN MB, MORA CA, JACOBSON S. In vitro spontaneous lymphoproliferation in patients with human T-cell lymphotropic virus type I-associated neurologic disease: predominant expansion of CD8+ T-cells. **Blood.** 98: 1506-1511, 2001.

SANDERS RC, WAI'IN PM, ALEXANDER SS, LEVIN AG, BLATTNER WA, ALPERS MP. The prevalence of antibodies to human T-lymphotropic virus type I in different population groups in Papua New Guinea. **Arch Virol.** 130(3-4):327-34, 1993.

SANTOS SB, PORTO AF, MUNIZ AL, DE JESUS AR, MAGALHÃES E, MELO A, DUTRA WO, GOLLOB KJ, CARVALHO EM. Exacerbated inflammatory cellular immune response characteristics of TSP/HAM is observed in a large proportion of HTLV-I asymptomatic carriers. **BMC Infect Dis**. Mar 2;4:7, 2004.

SEIBEL, L. F. B. and Lifschitz, S. **An overview of genomic databases research** issues**.** In *SBBD*, page 10. UFRGS, 2002.

SEIKI, M., HATTORI, S. & YOSHIDA, M. Human adult T-cell leukemia virus: molecular cloning of the provirus DNA and the unique terminal structure. **Proc NATLL Acad Sci** USA; 79, 6899-6902, 1982.

SEIKI, M., S. HATTORI, Y. HIRAYAMA, M. YOSHIDA**.** Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. **Proc. NATLL. Acad. Sci. USA.**;80:3618–3622,1983.

SHIMOTOHNO, K., TAKAHASHI, Y., SHIMIZU, N., GOJOBORI, T., GOLDE, D.W., CHEN, I.S.Y. Complete nucleotide sequence of an infectious clone of human T-cell leukemia virus type II: an open reading frame for the protease gene. **Proceedings of the National Academy of Sciences of the USA, 82:** 3101-3105, 1985.

SONODA S, LI HC, TAJIMA K. Ethnoepidemiology of HTLV-1 related diseases: ethnic determinants of HTLV-1 susceptibility and its worldwide dispersal. **Cancer Sci.**102(2):295-30, 2011.

TAKAYANAGUI, OM; CASTRO-COSTA, CM. Vírus linfotrópico de células T humanas Tipos 1 e 2 (HTLV – 1/2) – mielopatia associada ao HTLV-1 / paraparesia

espástica tropical (TSP/HAM). **Cadernos Hemominas. HTLV.** Volume XIII, 4ª edição, Belo Horizonte, p.115-139, 2006.

TATA, S., Patel, J. M., Friedman, J. S., and Swaroop, A. Declarative querying for biological sequences. In *ICDE*, page 87. **IEEE Computer Society** ,2006.

TSUJIMOTO, A., T. TERUCHI, J. IMAMURA, *et al*. Nucleotide sequence analysis of a provirus derived from HTLV-1 associated myelopathy (HAM). ***Mol. Biol. Med***.;5:29–42,1988.

UCHIYAMA, T.;YODOI J.;SAGAWA K.; TAKATSUKI, K.; UCHINO, H. Adult T-cell leukemia:clinical and hematologic features of 16 cases. **Bloob..** v.50, p.481-92, 1977.

VALLINOTO AC, MUTO NA, PONTES GS, MACHADO LF, AZEVEDO VN, DOS SANTOS SE, RIBEIRO-DOS-SANTOS AK, ISHAK MO, ISHAK R. Serological and molecular evidence of HTLV-I infection among Japanese immigrants living in the Amazon region of Brazil. **Jpn J Infect Dis**.57(4):156-9, 2004.

VAN DOOREN, S, GOTUZZO E, SALEMI M, *et al*. Evidence for a Post-Colombian introduction of human T-cell lymphotropic virus in Latin America. **J Gen Virol**; 79:2695-2708, 1998.

VANDAMME, A.M., H.-F. LIU, P. GOUBAU AND J. DESMYTER. Primate T-lymphotropic virus type I *LTR* sequence variation and its phylogenetic analysis: compatibility with an African origin of PTLV-I. **Virology.**202:212-223, 1994.

VANDAMME,A.M., SALEMI,M., VAN BRUSSEL,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. **J Virol** .72(5):4327-4340, 1998.

VIDAL, A.U., A. GESSAIN, M. YOSHIDA, *et al*. Phylogenetic classification of human T cell leukaemia/lymphoma virus type I genotypes in five major molecular and geographical subtypes. **Journal of General Virology.** 75:3655-66, 1994.

VINE, A.M. e cols. Polygenic control of human T lymphotropic virus type 1 (HTLV-1) provirus load and the risk of HTLV-1-associated myelopathy/tropical spastic paraparesis. **J Infect. Dis.,** v.186, n.7, p. 932-939, 2002.

WEISS, Vinicius Almir. ***Estratégias de Finalização da Montagem do Genoma da Bactéria Diazotrófica Endofítica Herbaspirillum seropedicae SmR1*.** Curitiba, 2010. 72 f. Dissertação (mestrado em Ciências - Bioquímica). Departamento de Bioquímica. Universidade Federal do Paraná

WOLFE ND, HENEINE W, CARR JK, ET AL. Emergence of unique primate T-lymphotropic viruses among central African bushmeat hunters. ***Proc NATLL Acad Sci USA.***102 (22):7994-9, 2005.

Yakova M, Lezin A, Dantin F, Lagathu G, Olindo S, Jean-Baptiste G, Arfis, Cesaire R. Increased proviral load in HTLV-1-infected patients with rheumatoid arthritis or connective tissue disease. **Retrovirology**. 2 (1): 4, 2005.

YAMAGUCHI, K. Human T-lymphotropic virus type I in Japan. **Lancet.** v. 343, 213-216, 1994.

YAMAGUCHI, K. & WATANABE, T. HTLV-I and adult T-cell leukemia in Japan. **Int J Hematol.** 76 s2**,** 240-5, 2002.

YOSHIDA M, MIYOSHI I, HINUMA Y. Isolation and characterization of retrovirus from cell lines of human adult T-cell leukemia and its implication in the disease**. Proc NATLL. Acad Sci U S A.** 79(6):2031-5,1982.