UNIVERSIDADE DO ESTADO DA BAHIA (UNEB)
VAGNER DE SOUZA FONSECA

Estudo da correlação entre frequência de códons em genomas virais e a abundância de espécies de RNA transportador cognato na célula hospedeira humana

Salvador, Bahia, Brasil.
2013

VAGNER DE SOUZA FONSECA

Estudo da correlação entre frequência de códons em genomas virais e a abundância de espécies de RNA transportador cognato na célula hospedeira humana

Trabalho de Conclusão de Curso apresentado ao Colegiado do Curso de Bacharelado em Sistemas de Informação da Universidade do Estado da Bahia como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.
Orientador: Diego Gervásio Frías Suarez.

Salvador, Bahia, Brasil.
2013

VAGNER DE SOUZA FONSECA

Estudo da correlação entre frequência de códons em genomas virais e a abundância de espécies de RNA transportador cognato na célula hospedeira humana

Trabalho de Conclusão de Curso apresentado ao Colegiado do Curso de Bacharelado em Sistemas de Informação da Universidade do Estado da Bahia como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.
Orientador: Diego Gervásio Frías Suarez.

Trabalho aprovado em 06 de dezembro de 2013:

_____

Diego Gervásio Frías Suarez
Doutorado em Modelagem Computacional
Universidade do Estado da Bahia (UNEB)

_____

Joana Paixão Monteiro Cunha
Doutorado em Biotecnologia em Saúde e Medicina Investigativa
Universidade Federal da Bahia (UFBA)

_____

Maria Inês Valderrama Restovic
Mestrado em Engenharia Elétrica
Universidade do Estado da Bahia (UNEB)

Salvador, Bahia, Brasil.
2013

Agradecimentos

Agradeço a Deus, pelo dom da vida e encontrar em suas palavras sabias conforto para superar as minhas dificuldades.

Em segundo lugar, agradeço minha mãe Neide Fonseca, a minha esposa Simara Teixeira e meu pai adotivo Leandro Coelho por estarem sempre ao meu lado em todos os momentos de minha vida me incentivando e confiando em meu potencial. E todo amor incondicional demonstrado a mim.

Aos meus irmãos Verusca, Tiago e Mariana, aos meus irmãos de coração Aécio e Saulo e aos meus pais de coração Antônio e Teresa, por compartilhar com eles toda a minha trajetória acadêmica e serem compreensivos a todos os momentos.

Agradeço aos grandes mestres e doutores que ensinaram na graduação, não somente ciência, pois ao longo desses anos tive o prazer de assistir a certas aulas que, com certeza, serão inesquecíveis.

Aos meus colegas de curso, por compartilharem os seus conhecimentos ao longo da jornada do curso, e, mostrando-se sempre disposto a ajudar um a outro. Agradeço, ainda, Ana América por sempre interceder pelos alunos do curso, junto aos professores, sempre estando ao lado do curso com dedicação, orientado e dando ótimos conselhos sempre que necessário.

Ao Dr. Luiz Alcântara e toda sua equipe que esteve sempre disposta a me ajudar com aulas de biologia, explicando todo o funcionamento dos retrovírus e proteínas para entendimento melhor do projeto.

Em especial, agradeço ao professor Diego Frias pela orientação fantástica, ao logo desta pesquisa, e me dar à oportunidade de realizar uma pesquisa interdisciplinar.

Agradeço também aos meus amigos, com quem pude compartilhar conhecimentos e confidências ao longo desses anos.

Resumo

Contextualização: A bioinformática é uma ciência multidisciplinar, que procura armazenar e relacionar os dados biológicos, utilizando o auxilio de métodos computacionais e matemáticos, subsidiando no reconhecimento de padrões que seriam difíceis sem a ajuda da tecnologia da informação. Pesquisas desenvolvidas nesta área utilizam, em geral, um volume considerável de informações que devem ser analisadas em conjunto para proporcionar inferências precisas sobre possíveis hipóteses e possibilitar assim a construção de novas teses. O armazenamento de informações em bancos de dados e a utilização de linguagens computacionais permite aos pesquisadores de todo o mundo compartilhar informações possibilitando aos mesmos estudar estruturas tridimensionais de moléculas, simular o metabolismo de células ao até mesmo desvendar a função biológica de determinada sequência de DNA. O estudo do genoma de patógenos, plantas e animais é uma linha de pesquisa em constante crescimento. Em particular o estudo do genoma de vírus e da evolução viral ocupa a atenção de muitos pesquisadores no mundo todo. Desde a descoberta do vírus da AIDS (Human Immunodeficiency Virus – HIV) nos anos 80 acentuou-se, ainda mais, o uso de sistemas computacionais para auxiliar no processamento de genomas virais, impulsionando a fusão da biologia com a informática. **Problema Abordado:** O objetivo deste trabalho foi o desenvolvimento de uma ferramenta de bioinformática para investigar a composição dos genes retrovirais no nível de códons e estabelecer sua compatibilidade com a maquinaria celular dos hospedeiros para síntese de proteínas. Os resultados deste estudo darão suporte teórico para o estudo de alvos de terapias antirretrovirais idealizadas pelos autores, baseadas na inibição seletiva de espécies de RNA transportador. **Contribuição:** Neste trabalho foi desenvolvida uma ferramenta WEB de bioinformática que permite estudar a frequência de códons nos genes do HIV e do vírus linfotrópico da célula humana (HTLV), assim como a frequência de genes de RNA de transporte no genoma do hospedeiro humano. **Relevância:** A ferramenta permitirá, graças a um procedimento implementado para atualização automática, realizar os estudos considerando todas as sequências virais existentes até o momento do estudo nos bancos de dados públicos. Desta forma os resultados terão maior representatividade e significância estatística, na medida que novas sequências sejam armazenadas nos bancos de dados.

Palavras-chaves: utilização de códons, RNA transportador, HIV, HTLV, terapia, banco de dados.

Abstract

Background: Bioinformatics is a multidisciplinary science, which aims to store and correlate biological data, using the aid of computational and mathematical methods. Bioinformatics supports the recognition of patterns that would be difficult without the help of information technology. Research conducted in this area use, in general, a considerable volume of information that must be analyzed together to provide accurate inferences about possible hypotheses and thus enable the construction of new theories. The storage of information in databases and the use of computer languages allows researchers around the world to share information, enabling them to study, for example, three-dimensional structures of molecules or the metabolism of cells, as well as to find the biological function of a particular DNA or RNA sequence. The study of the genome of pathogens, plants and animals is a line of research in constant growth. In particular the study of viruses and viral genome evolution occupies the attention of many researchers worldwide. Since the discovery of the virus causing AIDS (Human Immunodeficiency Virus-HIV) in the eighties, the use of computer systems to assist in the processing of viral genomes was accentuated, propelling the fusion of biology with computer science. **Addressed Problem**: The aim of this work was the development of a bioinformatics tool to investigate the composition of retroviral genes at the level of codons and establishing their fitness with the host cell machinery, used for protein synthesis. The results of this study will provide theoretical support for studying targets of antiretroviral therapies hypothesized by the authors, which is based on selective inhibition of transfer RNA species. **Contribution**: In this study we developed a bioinformatics web tool that allows to study the frequency of codons in the genes of HIV and Human T cell Lymphotropic Virus (HTLV), as well as the frequency of transfer RNA genes in the human host genome. **Relevance**: The tool will allow, thanks to a procedure implemented for automatic updating, performing studies considering all existing viral sequences in public databases until the moment of the study. Thus the results will be more representative of the virus population as new sequences are stored in such databases.

Keys word: codon usage, tRNA, HIV, HTLV, therapy, database.

Lista de ilustrações

Lista de tabelas

Lista de Abreviaturas e Siglas.

3’	Região carboxi-terminal do ácido nucléico
5’	Região amino-terminal do ácido nucléico
AIDS	Síndrome de Imunodeficiência Adquirida (*Acquired Immunodeficiency Syndrome*)
ATL	Leucemia/linfoma de células T do adulto (*Adult T cell Leukemia*)
BD	Banco de Dados
BDB	Banco de Dados Biológicos
BDE	Banco de Dados Específico
BDP	Banco de Dados Públicos
CATIRT	Códons Alvos de Terapia por Inibição de RNA transportador
CCAT	Códons Candidatos para Alvos Terapêuticos
CD4+	Grupamento de Diferenciação 4 *(Cluster of Differentation 4)*
CD25+	Grupamento de Diferenciação 25 *(Cluster of Differentation 25)*
CDPT	Códons Desfavorecidos no Processo de Tradução
CDS	Sequências Codificantes *(Coding Sequences)*
CRUD	*Create, Read, Update, Delete*
CUB	*Codon Usage Bias*
CUT	Frequências de Códons
DDBJ	*DNA Data Bank of Japan*
DNA	Ácido Desoxirribonucleico *(Deoxyribonucleic Acid)*
EBI	*European Bioinformatics Institute*
EMBL	*European Molecular Biology Laboratory*
gp41	Glicoproteína 41
gp120	Glicoproteína 120
HIV	Vírus da Imunodeficiência Humana *(Human Imunodeficiency Vírus)*
HTLV	Vírus Linfotrópico da Célula Humana *(Human T cell Lymphotropic vírus)*
HTML	*HyperText Markup Language*
INSDC	*International Nucleotide Sequence Database Collaboration*
KDD	*Knowledge Discovery in Database*
LCP	Leucemia de Célula Pilosas
mRNA	Ácido Ribonucleico Mensageiro *(Messenger Ribonucleic Acid)*

NCBI	*National Center for Biotechnology Information*
NIH	*National Institutes of Health*
ORF	Fase de Leitura aberta *(Open Reading Frame)*
PHP	*Hypertext Preprocessor (*origenalmente: *Personal Home Page)*
RNA	Ácido Ribonucleico (*Ribonucleic Acid*)
SCAM	Sistema Configurável Automático de Mineração
SGBD	Sistema de Gerenciamento de Bancos de Dados
SQL	*Structured Query Language*
STLV	*Simian T-cell leukemia virus*
TRIT	*Inhibition Therapy*
tRNA	Ácido Ribonucleico Transportador *(Transfer Ribonucleic Acid*)
UFM	*Universal Feature Method*
XML	*Extensible Markup Language*

Sumário

1. Introdução

A bioinformática é uma ciência multidisciplinar, que aborda o armazenamento e o relacionamento dos dados genômicos, transcriptômicos[1] ou de expressão e proteômicos[2], com o auxílio de técnicas e métodos computacionais e matemáticos. Em particular, faz um uso intensivo de técnicas de reconhecimento de padrões em sequências e conjuntos de sequências que seriam difíceis de detectar e reconhecer sem a ajuda da tecnologia da informação. Neste aspecto pesquisas bioinformáticas processam, em geral, um volume considerável de informações que deve ser analisado em conjunto para proporcionar inferências precisas sobre possíveis hipóteses e possibilitar assim a construção de novas teses sobre o papel ou funcionamento biológico de determinada proteína ou organismo.

Tomando por exemplo as pesquisas realizadas sobre o genoma humano, de outras espécies e dos patógenos[3], acentua-se, ainda mais, a necessidade do uso de sistemas computacionais para auxiliar no processamento deste volume de informação. Tal fato tem impulsionado a fusão da biologia com a informática.

O armazenamento de informações em um banco de dados e a utilização de linguagens computacionais permitem que pesquisadores de todo o mundo compartilhem informações possibilitando aos mesmos estudarem estruturas tridimensionais de moléculas, simular o metabolismo de células e até mesmo desvendar a função biológica dos genes.

1.1. Biologia Molecular

A biologia molecular explora a vida em escalas moleculares estudando a genética, o Ácido Desoxirribonucleico (DNA), a produção de proteínas e o material genético em geral. A biologia molecular estuda os padrões moleculares baseados em aprofundamento genético e bioquímico, a principal diferença entre genética e a bioquímica é a forma de trabalho, enquanto a genética e a bioquímica trabalham com análise macro ou microscópica de tecidos ou células a biológica molecular trabalha em

[1] Estudo de todos os RNAs que existem em uma célula, tecido ou órgão.
[2] Estuda o conjunto de proteínas e suas isoformas contidas em uma célula, tecido ou órgão.
[3] Microrganismo causador de doença.

um nível submicroscópico. Os padrões estudados definem as e funções do material genético e seus produtos de expressão (proteínas), analisa a interação entre diversos sistemas celulares entre eles a interação entre o DNA, o Ácido Ribonucleico (RNA) e a síntese proteica, compreendendo assim os processos de replicação, transcrição, tradução do material genético e a regulação desses processos.

A replicação das células é possível, pois todas as células vivas da Terra armazenam suas informações hereditárias na forma de moléculas de DNA de fita dupla em longas cadeias poliméricas pareadas não ramificadas, formadas por nucleotídeos, cada nucleotídeo consiste em duas partes: um açúcar (desoxirribose), com um grupo de fosfato ligado a ele e uma base que pode ser de Adenina (A), Guanina (G), Citosina (C) ou Timina (T) (Figura 1) (ALBERTS *et al*, 2010).

O processo de replicação do DNA envolve a participação de diversas enzimas, entre elas, as polimerases[4], que atuam no processo da síntese da nova molécula de DNA. Essa nova formação precisa que desoxirribonucleotídeos livres, sejam posicionados sobre um molde (cadeia de polinucleotídica), e sejam unidos entre si, formando, assim, uma nova cadeia complementar à cadeia mãe.

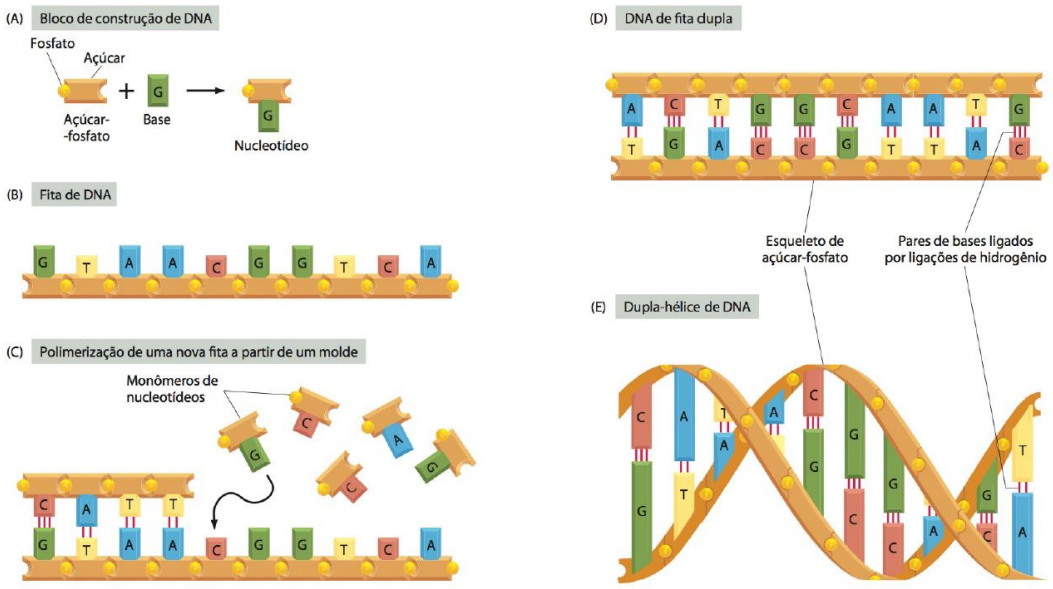

Figura 1 - Blocos de construção do DNA
Fonte: ALBERTS *et al*, 2010

[4] Enzima que catalisa a formação de um novo DNA e RNA a partir de uma molécula existente de DNA ou RNA.

Como o DNA não direciona a síntese proteica diretamente, devido a suas informações estarem localizadas quase que totalmente no núcleo das células, e a síntese proteica ocorre no citoplasma das células, ele utiliza uma molécula intermediaria semelhante a um fio que leva a informação do núcleo para os ribossomos, denominada de RNA. Com isso quando a célula necessita de uma proteína especifica a sequência de nucleotídeos da região apropriada de uma molécula de DNA é inicialmente copiada sob a forma de RNA por um meio denominado de transcrição. São essas cópias que são usadas diretamente como moldes para direcionar a síntese da proteína em um processo chamado tradução (Figura 2). Esse processo é tão fundamental para a célula que ele é denominado o dogma central da biologia molecular, onde todas as células existentes realizam esse processo.

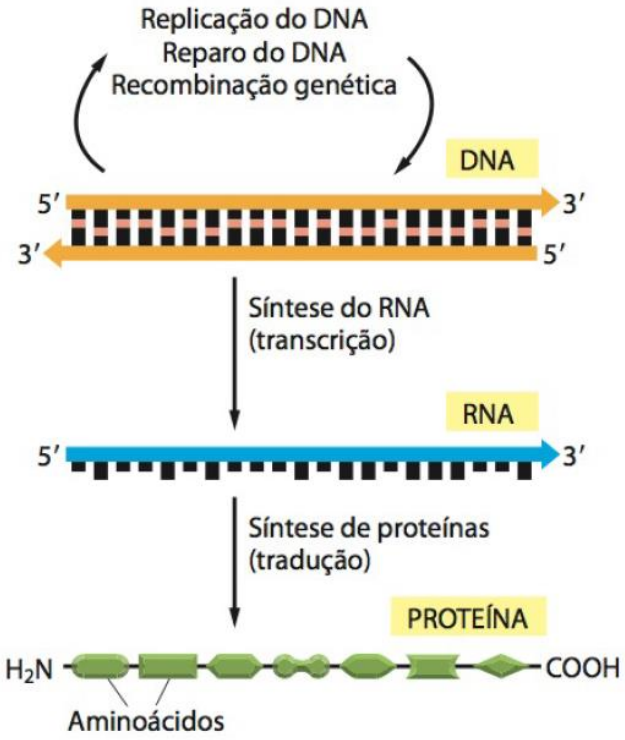

Figura 2 - Processo de replicação, transcrição e tradução.
Fonte: ALBERTS *et al*, 2010

O RNA é formado por um açúcar (ribose) ligeiramente diferente do açúcar (desoxirribose) do DNA e há, também, uma diferença em uma das bases do nucleotídeo a Uracila (U) substituindo a Timina (T). Durante a transcrição os nucleotídeos de RNA são alinhados e selecionados para a polimerização a partir de uma fita-molde de DNA, processo parecido com a replicação do DNA, como consequência tem-se uma molécula

de polímero cuja sequência de nucleotídeos representa fielmente uma parte da informação genética da célula. Como a informação genética do DNA é fixa e inviolável os transcritos de RNA são produzidos em massa e são descartáveis (Figura 3), estes transcritos funcionam como intermediários na transferência da informação genética atuando principalmente como RNA mensageiro (mRNA) para guiar a síntese de proteínas de acordo com as instruções genéticas armazenadas no DNA.

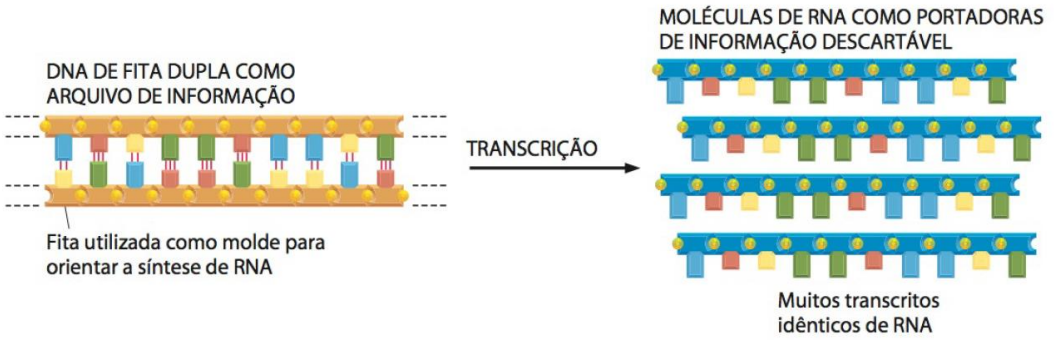

Figura 3 - Processo de transcrição do DNA para o RNA
Fonte: ALBERTS *et al*, 2010

As proteínas são cadeias poliméricas longas não ramificadas formadas por sequências de blocos construtores monoméricos (aminoácidos) retirados de um repertório padrão semelhantes em todas as células. As proteínas carregam informações em forma de sequências lineares de símbolos e retirando a água das células, as proteínas constituem a maior parte da massa de uma célula. Enquanto no DNA e RNA existem 4 monômeros nas proteínas existem 20 tipos diferentes, denominados de aminoácidos.

A tradução da informação genética contida em uma sequência de mRNA é lida em códons (grupos de três nucleotídeos) por vez, cada trinca de nucleotídeo especifica codifica para um único aminoácido na proteína correspondente. O código é lido por uma classe especial de moléculas pequenas de RNA, os RNAs transportadores (tRNAs). Cada tipo de tRNA liga-se a uma extremidade de um aminoácido especifico, apresentando em sua outra extremidade uma sequência especifica de um anticódon que o habilita a reconhecer pelo pareamento de bases, um códon ou um grupo de códons no mRNA (Figura 4).

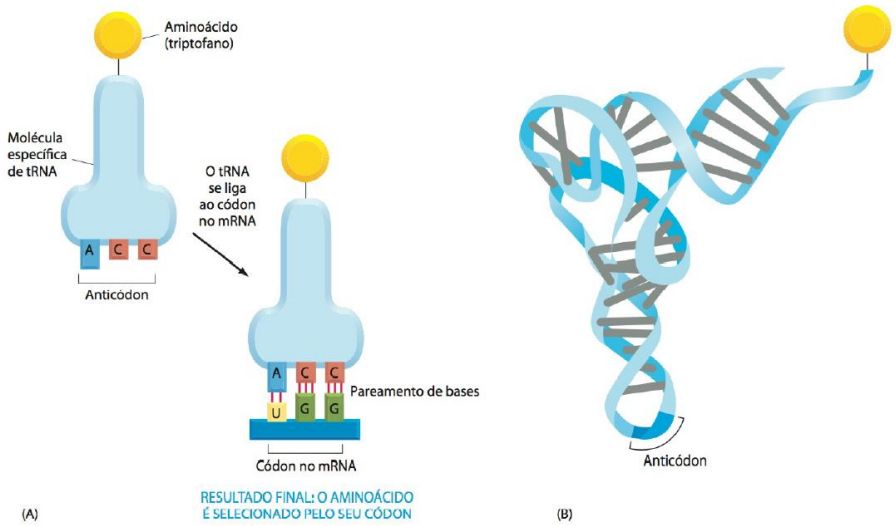

Figura 4 - Processo de tradução de RNA para aminoácidos
Fonte: ALBERTS *et al*, 2010

Para a síntese proteica, uma sucessão de moléculas de tRNA carregadas com os seus aminoácidos apropriados deve unir-se a uma molécula de mRNA e, por meio do pareamento os anticódons dos tRNAs emparelham-se com cada um dos seus códons sucessivos. Os aminoácidos devem, então, ser ligados uns aos outros para alongar a cadeia de proteínas crescente, e os tRNAs, atenuados de suas cargas, devem ser liberados. Todo este conjunto de processos é realizados por uma gigantesca máquina multimolecular, o ribossomo, formado por duas cadeias principais de RNA, chamadas de RNAs ribossomos (rRNAs) (Figura 5), junto a mais de 50 proteínas diferentes. Essa molécula evolutivamente antiga agarra-se à porção terminal de uma molécula de mRNA e se desloca ao longo dela, capturando moléculas de rRNA carregadas para formar uma nova cadeia de proteínas pela ligação dos aminoácidos que elas transportam.

A tradução geralmente começa com o códon AUG e termina com um dos códons UAA, UGA e UAG. Terminado, assim, o processo de formação das proteínas. Todo este processo permite que a informação genética contida no núcleo chegue ao citoplasma para produzir as proteínas necessárias para um organismo mesmo ambos estando separados pela membrana nuclear.

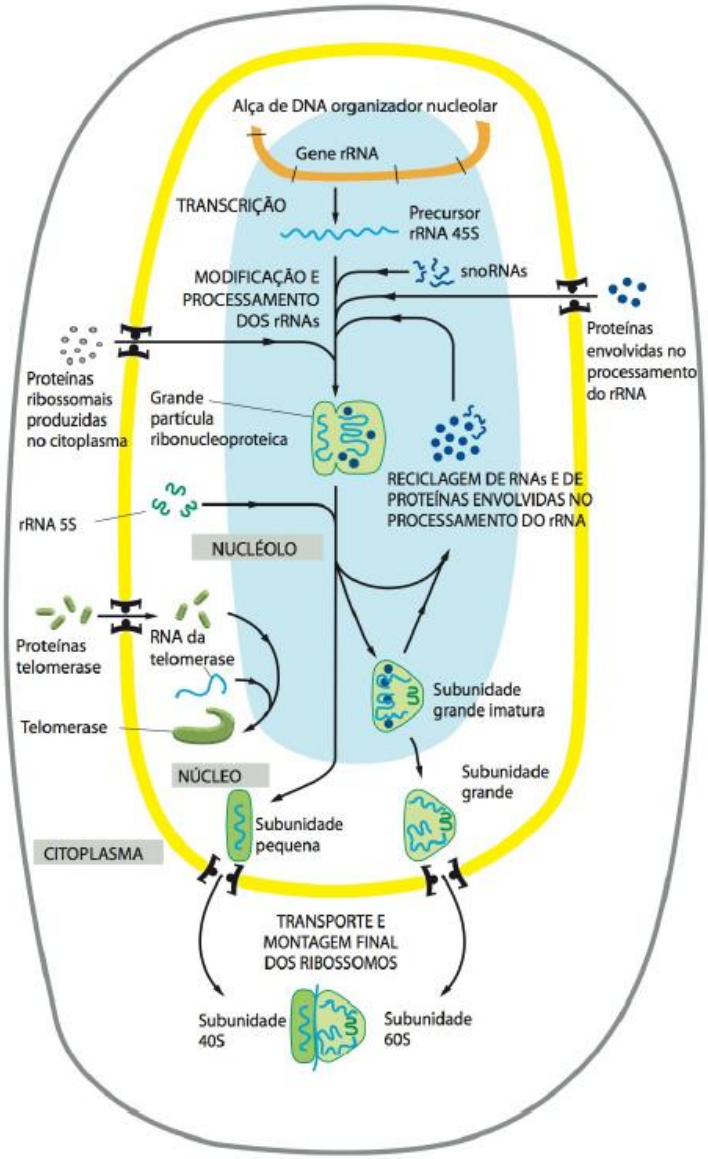

Figura 5 - Função do nucléolo na síntese do ribossomo e de outras ribonucleoproteínas
Fonte: ALBERTS et al, 2010

1.2. Vírus e genoma viral

Entre os patógenos, os vírus causam muitas doenças humanas comuns como gripe varicela, resfriados, diarreia, sarampo, rubéola e caxumba. Algumas enfermidades viróticas como a raiva, febre hemorrágica, encefalites, poliomielites, febre amarela e a Síndrome de Imunodeficiência Adquirida ou *Acquired Immunodeficiency Syndrome* (AIDS) são mortais. Estima-se que existam entre 1000 e 1500 espécies de vírus, das quais aproximadamente 250 têm sido identificadas como patógenos humanos, até o momento. Na Tabela 1 mostramos uma lista resumida dos vírus que mais infectam o hospedeiro humano.

Tabela 1 - Lista resumida dos vírus que mais infectam o hospedeiro humano
Adaptada de xxxxxxxxxx

TIPO	VIRUS	DOENÇA
Adenovirus	----------	Resfriado comum
Bunyavirus	Hantaan La Crosse Sem nome	Insuficiência renal Encefalites Síndrome pulmonar
Caliivirus	Norwalk	Gastroenterites (diarreia e vómitos)
Coronavirus	Corona	Resfriado comum
Filovirus	Ebola Marburg	Febre hemorrágica Febre hemorrágica
Flavivirus	Hepatite C (no A, no B). Febre Amarela	Hepatitis Hepatite, hemorragia
Hepadnavirus	Hepatite B (VHB)	Hepatites, cancer de fígado.
Herpes vírus	Citomegalovirus Vírus Epstein-Barr (VEB) Herpes simples tipo 1 Herpes simples tipo 2 Vírus herpes humano 8 (VHH8) Varicela-zoster	Defeitos de nascimento Mononucleose, câncer nasofaringe. Herpes labial Lesões genitais Sarcoma de Kaposi Varicela, herpes zoster.
Ortomixovírus	Influenza tipos A e B	Gripe
Papovavirus	Vírus do Papiloma Humano (VPH)	Verrugas, câncer de colo do útero.
Picornavírus	Coxsakievirus Echovirus Hepatite A Poliovirus Rinovírus	Miocardite (infecção do musculo cardíaco) Meningite Hepatite infecciosa Poliomielite Resfriado comum
Paraminovirus	Sarampo Parotidite infecciosa Parainfluenza	Sarampo Parotidite infecciosa Resfriado comum, infeção do ouvido.
Parvovírus	B19	Eritema infeccioso, anemia.

À diferença de outros patógenos como as bactérias e fungos, o vírus não possui mecanismo próprio para síntese de suas proteínas. Levando-se em conta que as proteínas constituem a estrutura dos organismos, entre múltiplas outras funções, pelo fato de não ser capaz de sintetizar seu proteoma (conjunto de todas as proteínas codificadas no genoma de um organismo) os vírus não são considerados seres vivos. Para suprir esta carência os vírus infectam células do organismo hospedeiro e utilizam a maquinária de tradução de RNA em proteínas dessas células.

Por exemplo, para poder replicar-se o vírus precisa sintetizar diversas glicoproteínas que formam sua estrutura externa ou capsídeo[5], além de outras proteínas do seu núcleo onde guarda o código genético da sua espécie. De acordo com isto, a taxa de tradução de proteínas viróticas está diretamente relacionada com a taxa de replicação do vírus,

[5] Material envoltório dos vírus, um invólucro protetor constituído de proteínas, que protege e facilita sua proliferação, e além de proteger o genoma (DNA ou RNA, mas nunca os dois como nos demais seres).

impactando na virulência (capacidade de infectar outros indivíduos) do patógeno a na própria evolução das doenças causadas por eles.

Pesquisas recentes (FRIAS, *et al.*, 2010) com uma família de vírus altamente patogênica, chamada de retrovírus, à qual pertence o HIV causador da AIDS, sugerem que deve existir uma correlação entre as frequências de códons nos genes viróticos e a abundância de espécies de RNA transportador na célula hospedeira, e que, a depender de quão compatível seja essa correlação, a taxa de replicação viral será maior ou menor.

1.3. Objetivos da pesquisa

Visando aprofundar na solução dessa hipótese, este trabalho tem como objetivo desenvolver uma ferramenta de bioinformática para subsídio ao estudo de uma nova abordagem de terapia antiviral baseada em interferência seletiva de espécies de RNA transportador. Para subsidio ao objetivo serão realizados os seguintes procedimentos:

- Projetar e desenvolver um Banco de Dados específico (BDE) para a ferramenta de bioinformática.
- Desenvolver um robô para realizar a atualização automática, mas controlada do BDE a partir de diversos Bancos de Dados Públicos (BDP).
- Desenvolver um portal WEB para disponibilizar resultados baseados em estudos de correlações entre: (a) as frequências de códons nos genomas dos patógenos e dos hospedeiros e (b) as frequências das espécies cognatas (semelhantes) do RNA transportador no hospedeiro, calculadas segundo modelo parametrizado de compartilhamento do RNA transportador.

A ferramenta WEB permitirá que os pesquisadores obtenham conhecimento inovador que poderia ser utilizado para: (a) inibir a replicação viral nas células infectadas, mediante algum procedimento (fármaco) que diminuísse a abundância das espécies de RNA transportador (que transportam os aminoácidos codificados por esses códons); (b) criar cepas de vírus com baixas taxas de replicação substituindo os códons essenciais por códons sinônimos classificados como códons desfavorecidos no processo de tradução (CDPT), gerando cepas que podem ser utilizadas para a fabricação de vacinas.

No portal WEB os pesquisadores poderão acessar os resultados dos estudos realizados e interagir com o sistema para caracterizar genomas virais submetidos segundo o indicador de viabilidade terapêutica desenvolvida.

Como o volume de dados armazenados nos bancos de dados públicos relacionados com o genoma humano, de outras espécies e dos patógenos, é muito grande e cresce exponencialmente a cada dia, o uso de sistemas computacionais para processar esse volume de informação e realizar estudos específicos é inevitável e de fato tem impulsionado o desenvolvimento vertiginoso da bioinformática, para gerenciamento e refinamento dessas informações.

A realização de estudos com novos focos terapêuticos são essenciais no combate aos retrovírus dado o continuado fracasso da comunidade científica internacional nas tentativas de desenvolver uma vacina eficaz contra *Human Imunodeficiency Vírus* (HIV) e *Human T cell Lymphotropic vírus* (HTLV).

2. Referencial Teórico

Neste capítulo serão apresentados alguns conceitos que embasam o desenvolvimento desta pesquisa. Na sessão 2.1 será apresentada uma contextualização sobre a classificação dos retrovírus. Na sessão 2.2 aborda-se o HIV, apresentando sua classificação, os mecanismos de infecção e replicação no hospedeiro. Na sessão 2.3 mostra-se o HTLV apresentando sua classificação, os mecanismos de infecção e replicação no hospedeiro.

Na sessão 2.4, são apresentados os conceitos de códons que forma aminoácidos, RNA mensageiro e transportador, abordando a correlação entre a frequência de códons e RNA transportador. Na sessão 2.5 são explicados como realizar o cálculo da frequência de códons em genes e genomas de sequências completas de DNA. Por fim aborda-se a mineração das informações para a modelagem do Banco de Dados Biológico, apresentando os principais Bancos de Dados Públicos, de onde serão coletadas as sequências codificantes que darão subsidio a esta pesquisa.

2.1. Classificação dos retrovírus

Retroviridae é que caracteriza os retrovírus humanos e são classificados em: alpharetrovírus (espécies de vírus da leucemia aviária), betaretrovírus (espécies de vírus do tumor mamário dos ratos), gammaretrovírus (espécies de vírus da leucemia de camundongos e vírus da leucemia felina), deltaretrovírus (espécies de vírus da leucemia bovina e os vírus HLTV, que afetam humanos), epsilonretrovírus (espécies de vírus do sarcoma de dérmica Walleye, e vírus da hiperplasia epidérmica Walleye 1 e 2, infectam geralmente peixes), lentivírus (espécies vírus da imunodeficiência humana (AIDS) e da imunodeficiência de felinos e símios) e spumavírus (espécies de vírus exógenos que têm morfologia específica com picos de superfície de destaque). Os deltaretrovírus pertencem a um grupo especial de vírus que quando infectam a células têm a capacidade de alterar o ciclo celular induzindo o desenvolvimento de tumores, e é composto pelos HTLV-1, HTLV-2, HTLV-3 e HTLV-4. Os lentivírus são vírus com longo período de incubação associados a doenças neurológicas e imunossupressoras, e é composto pelos HIV-1, HIV-2 e HIV-3 (CUELLAR, 1998). Sendo assim, HIV é um lentivírus, da família dos retrovírus que apresentam um genoma de RNA contido dentro de um capsídeo e um envelope lipídico (CHINEN; SHEARER, 2002).

2.2. HIV

A AIDS é causada pelo HIV (RAAPHORST *et al*., 2002). O HIV, (Figura 6), foi identificado e descoberto no inicio da década de 80, pelos pesquisadores Luc Montagnier, Françoise Barré-Sinoussi e Odete Ferreira do Instituto Pasteur na França (BARRE-SINOUSSI *et al.*, 1983). As infecções, oportunistas ou malignas, do HIV (REPETTO *et al*., 1996) vêm sido descritas como um processo continuo da perda das células responsáveis pela regulação imunológica do organismo humanos, conhecidas como células do linfócito T regulatório (Treg) (CD4+,CD25+). Atualmente os subtipos HIV-1 e HIV-2, infectam entre 10 e 20 milhões de pessoas em diferentes regiões do mundo[6].

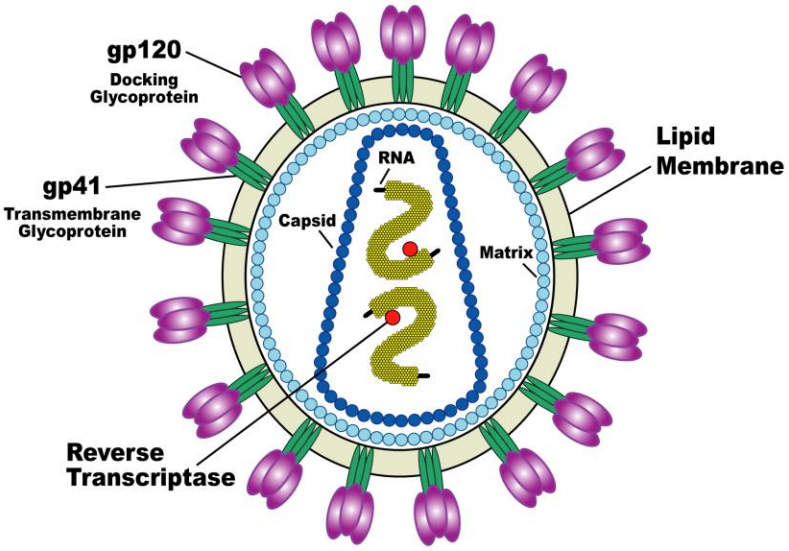

Figura 6 - Estrutura do HIV.
Fonte: http://www.stanford.edu/group/virus/retro/2005gongishmail/hiv1.jpg

O HIV é dividido em três tipos HIV-1, HIV-2 e HIV-3, o HIV-1 possui uma variabilidade genética de até 30% e são classificados em quatro grupos: M, O, N e P. O grupo M possui nove subtipos puros (não recombinantes), identificados pelas letras A, B, C, D, F, G, H, J e K; os subtipos A é dividido em A1 e A2 e os subtipos F é dividido F1 e F2, e há também as formas recombinantes circulantes, conhecidas por CRF. Na CRF há uma mutação de dois ou mais subtipos puros do grupo M (Tabela 2). Segundo Pereira (2010), o subtipo C é responsável por aproximadamente 56% das infecções em

[6] Dados do Departamento de DST, Aids e Hepatites Virais do Ministério da Saúde, disponível no endereço eletrônico http://www.aids.gov.br/pagina/coinfeccoes.

todo mundo. O HIV-2 possui sete subtipos puros (não recombinante), identificados pelas letras A, B, C, D, E, F e G; e até o momento existe uma recombinante CRF chamada de HIV2-CRF01_AB. A epidemiologia molecular do HIV é complexa e com o surgimento de novas variantes diferentes no mundo, apresentando um número maior de genes em relação aos outros vírus (PEREIRA, 2010). Além dos lentivírus apresentarem em seu genoma os genes estruturais *gag*, *pol* e *env* comuns em todos os retrovírus, eles ainda apresentam genes acessórios e regulatórios. O processo da replicação do vírus HIV é um tópico em constante estudo. A entrada nas células ocorre por meio de um complexo de duas glicoproteínas virais denominada de gp120 e gp41, situadas no envelope viral. A gp120 possui uma fita complexa com alta analogia para células CD4+. A gp41 é o elemento que realiza a fusão do envelope viral com a membrana plasmática da célula, permitindo que o genoma e as proteínas dos vírus entrem para o citoplasma (JANEWAY; TRAVERS, 1997).

Tabela 2 - Formas Recombinantes Circulantes do HIV-1
Fonte: adaptada http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html

Nome	Subtipos	Autor	Nome	Subtipos	Autor
CRF01_AE	A, E	JK Carr	CRF30_0206	CRF02, CRF06	M. Peeters
CRF02_AG	A, G	JK Carr	CRF31_BC	B, C	M. Soares
CRF03_AB	A, B	K. Liitsola	CRF32_06A1	CRF06, A1	M. Adojaan
CRF04_cpx	A, G, H, K, L	D. Paraskevis	CRF33_01B	CRF01, B	KP Ng & KK Tee
CRF05_DF	D, F	T. Laukkanen	CRF34_01B	CRF01, B	FE McCutchan
CRF06_cpx	A, G, J, K	RB Oelrichs	CRF35_AD	A, D	FE McCutchan
CRF07_BC	B, C	R. Wagner	CRF36_cpx	CRF01, CRF02, A, G	R. Powell
CRF08_BC	B, C	FE McCutchan	CRF37_cpx	CRF01, CRF02, A, G, L	R. Powell
CRF09_cpx	A, G, L	FE McCutchan	CRF38_BF	B, F1	C. Lopez-Galindez
CRF10_CD	C, D	IN Koulinska	CRF39_BF	B, F1	MG Morgado
CRF11_cpx	A, E, G, J, L	M. Peeters	CRF40_BF	B, F1	MG Morgado
CRF12_BF	B, F1	JK Carr	CRF41_CD	C, D	S. Tovanabutra
CRF13_cpx	CRF01, A, G, J, L	K. Wilbe	CRF42_BF	B, F1	JC. Schmit
CRF14_BG	B, L	R. Najera	CRF43_02G	CRF02, G	C. Brennan
CRF15_01B	CRF01, B	FE McCutchan	CRF44_BF	B, F1	M. Thomson
CRF16_A2D	A2, D	U. Visawapoka	CRF45_cpx	A, K, L	M. Peeters
CRF17_BF	B, F1	JK Carr	CRF46_BF	B, F1	SS Sanabani
CRF18_cpx	A1, F, G, H, K, L	M. Thomson	CRF47_BF	B, F1	M. Thomson
CRF19_cpx	A1, D, L	M. Thomson	CRF48_01B	CRF01, B	Y. Takebe
CRF20_BG	B, L	M. Thomson	CRF49_cpx	A1, C, J, K, L	T. de Silva & M.

					Cotten
CRF21_A2D	A2, D	FE McCutchan	CRF50_A1D	A1, D	G. Foster
CRF22_01A1	CRF01, A1	JK Carr	CRF51_01B	CRF01, B	OT Ng
CRF23_BG	B, L	M. Thomson	CRF52_01B	CRF01, B	J. Li
CRF24_BG	B, L	M. Thomson	CRF53_01B	CRF01, B	KK T
CRF25_cpx	A, G, L	JK Carr	CRF54_01B	CRF01, B	KK T
CRF26_AU	A, L	M. Peeters	CRF55_01B	CRF01, B	X. Han
CRF27_cpx	A, E, G, H, J, K, L	M. Peeters	CRF57_BC	B, C	L. Li
CRF28_BF	B, F1	R. Diaz	CRF59_01B	CRF01, B	X. Han
CRF29_BF	B, F1	R. Diaz	CRF61_BC	B, C	X. Li

A entrada do vírus na célula do hospedeiro depende também da presença de um co-receptor na membrada celular (SLEASMAN; GOODENOW, 2003). A interação da gp120 com o marcador CD4+ promove a ligação da gp120 com o co-receptor, esse acontecimento aciona gp41 que promove a unificação do envelope viral com a membrana celular (SHERMAN; GREENE, 2002). Depois dessa união o RNA viral será transcrito em DNA com o auxílio de uma proteína chamada transcriptase reversa. A transcriptase reversa dos lentivírus erra frequentemente nesse processo pelo que o DNA viral incorporado no genoma da célula hospedeira difere do RNA original. Isto causa uma elevada variabilidade genética do genoma viral (SLEASMAN; GOODENOW, 2003). Após a integração do DNA viral com o genoma das células hospedeiras ele pode persistir nestas células sem a produção de partículas virais completas (*virion*) capazes de sair célula hospedeira.

2.3. HTLV

O HTLV é um deltavírus da família dos retrovírus que apresenta um genoma de RNA em duas fitas simples com polaridade positiva (BURKE, 1997), sendo composto por HTLV-1, HTLV-2, HTLV-3 e HTLV-4. O HTLV-1 (Figura 7) foi o primeiro retrovírus a ser descrito e isolado na década de 80 nos Estados Unidos (POIESZ *et al.*, 1980). Cerca de três anos antes, Uchiyama e seus colaboradores (1977), encontrou em pacientes no Japão um tipo especifico de leucemia com propriedades morfológica celular especial, denominada de Leucemia de células T do Adulto (ATLL). 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 *et al.*, 1981). Kalyanaraman e seus colegas no inicio da década de 80

encontrou e isolou em um paciente uma forma incomum da Leucemia de células T pilosas, o HTLV-2. Os tipos do HTLV-3 e HTLV-4 foram encontrados recentemente em pacientes na África Central (CALATTINI *et al*., 2005).

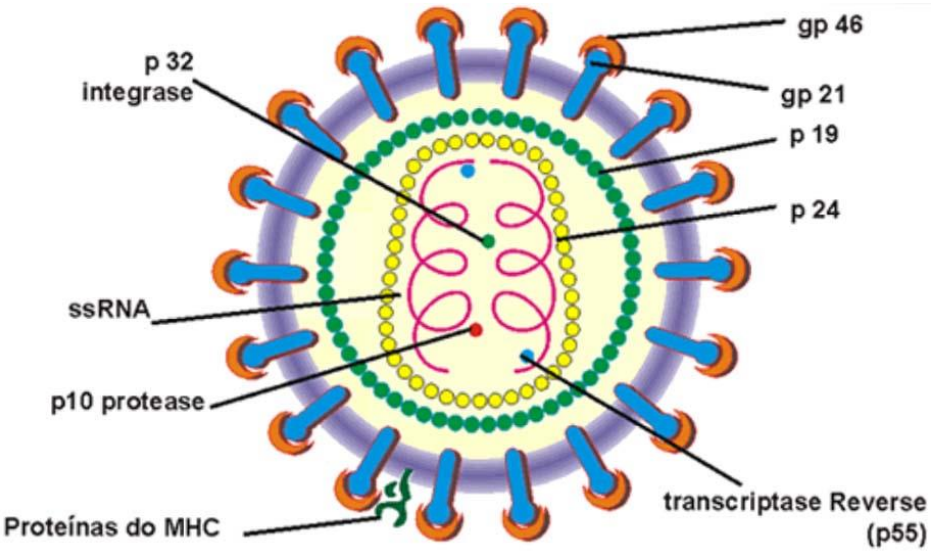

Figura 7 - Estrutura do HTLV-1.
Fonte: http://www.htlv.com.br

O HTLV-1 é classificado em 7 subtipos estratificados em “a”, “b”, “c”, “d”, “e”, “f” e “g”. O subtipo “a” é encontrado em quase todas as regiões geográficas do mundo, o “b” é comum na África Central, o “c” é comum na Melanésia, o “d” e “e” foi isolado em pigmeus africanos, o subtipo “f” foi encontrado no Gabão e o “g” foi recentemente em pacientes em Camarões (WOLFE *et al.*, 2005). O subtipo “a” esta dividido em 5 subgrupos estratificados em A (transcontinental), B (japonês), C (oeste africano), D (norte africano) e E (isolado em negros do Peru).

O HTLV-1 está associado a doenças graves neurológicas crônico-degenerativas que atinge o sistema nervoso central causando um aumento do distúrbio frequente em lesões congênitas dos membros inferiores (OSAME *et al.*, 1986). Esta patologia afeta mais mulheres do que homens, sendo que a maioria dos indivíduos infectados possui um o diagnostico tardio (OSAME *et al.*, 1986). Ainda não são conclusivas as pesquisas em relação à determinação da doença em pacientes, acreditando-se em uma possível influência do modo de transmissão do HTLV-1. É importante salientar, que grande maioria dos indivíduos infectados pelo HTLV-1 não desenvolve doenças, permanecendo assintomáticos (QUEIROZ, 2000).

A estrutura do genoma do HTLV-1 e do HTLV-2 englobam cerca de 70% da sequência de nucleotídeos possuindo, assim, uma paridade entre eles, pois ambos têm os genes virais essenciais *gag*, *pol* e *env* (FEUER; GREEN, 2005). Apesar do HTLV-2 ser considerado menos patogênico, comparado ao HTLV-1, ele associando-se a doenças neurológicas semelhantes a algumas variantes da Leucemia de Célula Pilosas (LCP) (FEUER; GREEN, 2005). A LCP é uma doença do sistema linfático, chamada também de linfoproliferativa, esta é uma doença crônica de acontecimento raro nos pacientes portadores da LCP, ela apresenta uma perda de peso, um aumento do volume do baço e uma diminuição global de elementos celulares do sangue (GONSALEZ *et al*., 1998). A confirmação da LCP ocorre através do exame da medula óssea ou da analise do sangue do paciente, assim é notada a presença de células do sistema imune que defendem o organismo contra invasão de agentes estranhos (GONSALEZ et al., 1998).

Os tipos 1 e 2 do HTLV são inteiramente relacionados às famílias dos vírus STLV-1 e STLV-2 (*Simian T Cell Leukemia Virus*), respectivamente, recomendando a pressuposição da transmissão do HTLV-1 por intermédio do relação entre humanos e/ou primatas não humanos (CALATTINI *et al.*, 2005). Após testes realizados em moradores rurais da África Central, mais precisamente em Camarões, foram encontrados dois novos os tipos de HTLV, os tipos 3 e 4. (CALATTINI *et al.*, 2005). Switzer e seus colaboradores (2006) notaram que o HTLV-3 é idêntico geneticamente ao STLV-3.

2.4. A correlação entre a frequência de códons e RNA transportador

O dogma central da biologia molecular estabelece que os genes que são expressos em uma dada célula de qualquer organismo, são inicialmente transcritos do DNA no núcleo da célula por um complexo de proteínas e enzimas liderado pela transcriptase, produzindo moléculas lineares (fitas) chamadas de mRNAs. Os mRNAs são formados por sequências de três nucleotídeos, chamadas códons (que codificam aminoácidos) ou que indicam o ponto de início (start códon) ou fim da tradução (stop códon) da cadeia do RNA mensageiro. As bases nitrogenadas presentes nos nucleotídeos são Adenina (A), Guanina (G), Citosina (C) e Timina (T). As duas primeiras são purinas e as duas últimas são pirimidinas. A diferença entre purinas e pirimidinas é sua estrutura, pois nas purinas o anel de carbono de seis membros é ligado a um anel de cinco membros e nas

pirimidinas possui apenas uma conformação do anel de seis membros. Na transcrição de um gene expresso contido na fita de DNA para a fita de RNA, a base Timina é substituída pela Uracila (U). Como existem 4 bases nitrogenadas no mRNA é possível formar 64 combinações diferentes de três bases. Dessas 64 combinações, 3 são utilizadas para indicar o fim da mensagem (UAA, UGA e UAG), restando 61 códons que codificam aminoácidos. Os mRNAs transcritos migram até o citoplasma celular passando através da membrana nuclear, para serem traduzidos nos ribossomos. A tradução implica na produção (síntese) de uma cadeia polipeptídica segundo a sequência de códons no mRNA, a qual após modificações pós-tradução, se converte em uma proteína. O código genético estabelece uma relação entre os 61 códons codificantes e os 20 aminoácidos que compõem as proteínas em todos os seres vivos no planeta (SPENCER *et al*, 2012). Dois (triptofano - W e metionina - M) dos 20 aminoácidos são codificados por um único códon, 1 aminoácido (isoleucina - I) é codificado por três códons, 9 aminoácidos são codificados por 2 códons, 5 aminoácidos são codificados por 4 códons e 3 aminoácidos por 6 códons, como indicado na Tabela 3.

Tabela 3 - Tradução dos Códons em aminoácidos
Fonte:

c	codon	aa	c	codon	aa	c	codon	aa	c	codon	aa
1	AAA	K	17	GAA	E	33	CAA	Q	49	UAA	*
2	AAG		18	GAG		34	CAG		50	UAG	*
3	AAC	N	19	GAC	D	35	CAC	H	51	UAC	Y
4	AAU		20	GAU		36	CAU		52	UAU	
5	AGA	R	21	GGA	G	37	CGA	R	53	UGA	*
6	AGG		22	GGG		38	CGG		54	UGG	W
7	AGC	S	23	GGC		39	CGC		55	UGC	C
8	AGU		24	GGU		40	CGU		56	UGU	
9	ACA	T	25	GCA	A	41	CCA	P	57	UCA	S
10	ACG		26	GCG		42	CCG		58	UCG	
11	ACC		27	GCC		43	CCC		59	UCC	
12	ACU		28	GCU		44	CCU		60	UCU	
13	AUA	I	29	GUA	V	45	CUA	L	61	UUA	L
14	AUG	M	30	GUG		46	CUG		62	UUG	
15	AUC	I	31	GUC		47	CUC		63	UUC	F
16	AUU		32	GUU		48	CUU		64	UUU	

Os aminoácidos que formam as proteínas são levados até o ribossomo por moléculas de tRNA (Figura 8), moléculas altamente especializadas para esta função, que são

transcritas no núcleo e que também migram para o citoplasma que é seu lugar de trabalho. Todo tRNA possui uma região chamada anticódon que lhe permite se aderir a um sítio específico do ribossomo (chamado de sítio P) no qual o ribossomo "expõe" um único códon da molécula de mRNA que está sendo traduzida, sempre que: (a) esteja carregando um aminoácido, o que em biologia molecular se diz estar aminoacilado e (b) seu anticódon seja "compatível" com o códon exposto pelo ribossomo. A compatibilidade requer que, pelo menos, as duas primeiras bases do códon sejam complementares, em sentido inverso, às duas últimas bases do anticódon do RNA transportador. Consideremos por exemplo o códon AGU exposto no ribossomo, isto é, que tem A na primeira posição (A1), G na segunda (G2) e U na terceira (U3). Neste caso o anticódon compatível seria XCU (ou X1, C2 e U3), de forma que a primeira base do códon, A1, se liga à base complementar U3 na terceira posição do anticódon, a segunda base do códon, G2, à base complementar C2 na segunda posição do anticódon, e a terceira base do códon U3 se liga à base genérica X1 na primeira posição do anticódon. Quando a primeira base do anticódon é complementar à terceira base do códon, neste exemplo seria X = A, se diz que acontece um pareamento clássico ou de Watson-Crick[7] entre o códon e o anticódon. Contudo, podem acontecer pareamentos não clássicos ou de tipo oscilante (*wobble*) quando X = G, ou X = C. Estes pareamentos tipo *wobble* são os que permitem na célula humana a tradução de 16 dos 32 códons no código genético terminados em pirimidinas (C3, U3) que não possuem RNA transportador no genoma humano. Se supõe que estes 16 códons são traduzidos pelos RNAs de transporte dos outros 16 códons terminados em pirimidinas, através do pareamento C3 <> A1 e U3 <> G1 (FRIAS et. al, 2013). O pareamento *wobble* acontece devido a modificações pós-transcricionais nos tRNAs (SPENCER *et al*, 2012).

[7] James Dewey Watson e Francis Harry Compton Crick vencedores do Premio Nobel de Fisiologia/ Medicina de 1962 por ter publicado o modelo de dupla hélice para a estrutura da molécula de DNA.

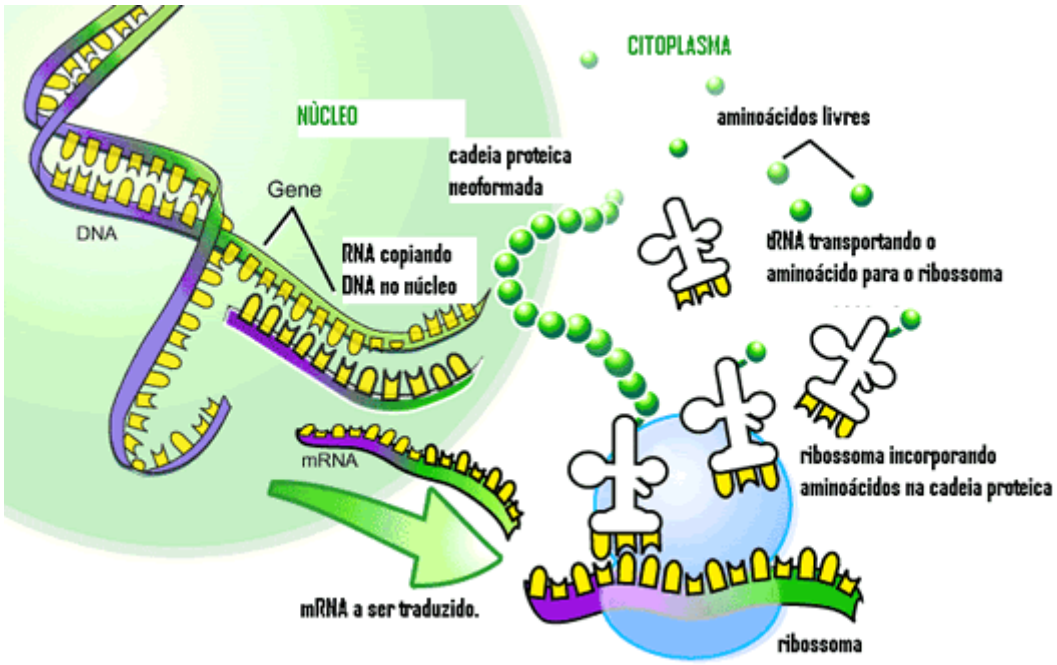

Figura 8 - Processo da sintese de proteínas
Fonte: http://www.sobiologia.com.br/conteudos/figuras/Citologia2/DNA2.gif

Na maioria dos organismos, há uma preferência dos genes por certos códons, chamados de códons ótimos, parecendo haver uma correlação entre a frequência desses códons no mRNA e a abundância de tRNA no citoplasma celular. Em teoria, quanto mais abundante uma espécie de tRNA, mais rapidamente serão traduzidos os códons compatíveis (cognatos) dado que haverá um menor tempo de espera pelo tRNA aminoacilado cognato ao códon exposto no ribossomo para ser traduzido (FRIAS et al, 2013). Outros resultados mostraram que o duplo mecanismo de pareamento do códon com o anticódon permite um movimento mais rápido do ribossomo ao longo da molécula de mRNA, possibilitando assim, uma maior taxa e maior precisão na tradução de códons (SPENCER *et al*, 2012). Por outro lado a presença de códons com baixa frequência, também conhecidos como códons raros, parece estar associada a uma necessidade de pausa na síntese da cadeia polipeptídica, para permitir a conformação espacial correta da proteína que está sendo sintetizada nos ribossomos (YADAVA e OCKENHOUSE, 2003).

Sendo assim, um requisito importante para o sucesso de uma vacina de proteína recombinante é a habilidade de produção de proteínas biologicamente ativas que podem ser facilmente utilizadas para produção em massa. A produção em massa da vacina depende do microambiente utilizado pela célula hospedeira, em particular da correlação

entre a abundância de tRNA e a frequência de códons na molécula de mRNA que codifica a proteína sendo produzida. (YADAVA e OCKENHOUSE, 2003).

Wong e seus colaboradores relatam que dois modelos têm sido propostos para explicar o uso dos códons no vírus da influenza A (vírus da gripe). Estes são os modelos de tradução seletiva e o modelo de mutação. No modelo de tradução seletiva existe uma coadaptação do uso do códon e a abundância do RNA transportador para aperfeiçoar a eficiência na tradução. No modelo de mutação existem restrições da composição genética que influenciam a probabilidade de fixação da mutação, esse fenômeno vem sendo encontrado em diversas espécies (WONG et al, 2012).

Nos estudos de Wong (2012), utilizando a análise de correspondência nos valores do uso de códons sinônimos relativos verificou-se que o vírus H1N1 (vírus da gripe) da pandemia de 1918, continha genes com padrões de utilização de códons virais de mamíferos, isso indica que a introdução deste vírus para os seres humanos não foi por meio único e exclusivo da transferência do vírus da gripe aviária. Esse estudo permitiu observar que muitos genes virais tiveram mudanças na frequência de códons ao longo do tempo, após o primeiro isolamento viral. Estas alterações reduziram o teor global de GC e do uso de G na terceira posição do códon no genoma viral, o que se acredita seja uma estratégia para otimizar a tradução dos genes virais (WONG *et al*, 2012).

2.5. Cálculo da frequência de códons em genes e genomas

Para calcular a frequência dos 61 códons codificantes nos genes é preciso o sequenciamento completo do DNA nos cromossomos e a posterior busca nessas sequencias pelos segmentos que são genes, isto é, que codificam proteínas ou moléculas de RNA funcionais (RNA ribossomal, RNA transportador, etc). É importante notar que a molécula de DNA é formada por duas fitas complementares, mas com orientações opostas, enroladas em forma de espiral. A complementaridade implica que pares de nucleotídeos A-T ou G-C aparecem em cada posição da sequência, um deles em uma das fitas e o outro na fita complementar. Devido a isso, para economizar trabalho, dinheiro e espaço em memória, no sequenciamento automatizado se armazena apenas a sequência de uma das duas fitas de DNA. A orientação da fita é determinada pelo

sentido em que o complexo de transcrição a percorre sendo por tanto o sentido de leitura do terminal 5’ ao terminal 3’, como mostrado no seguinte exemplo:

5’ –> AGTTCGACTGGACT –> 3’
|||| || | |||||||
3’ <– TCAAGCTGACCTGA <– 5’

Dada uma sequência de DNA podem se extrair 6 sequências de códons (tripletos) diferentes, 3 em cada fita, que originam sequências de aminoácidos diferentes. Cada uma das sequências de códons está associada a um quadro de leitura (*reading frame*). Usando o exemplo anterior teríamos:

Na fita superior:

Frame+1 ***AGT-TCG-ACT-GGA***-CT.
Frame+2 ..A-***GTT-CGA-CTG-GAC***-T..
Frame+3 .AG-***TTC-GAC-TGG-ACT***-…

Na fita inferior:

Frame-1 ***AGT-CCA-GTC-GAA***-CT.
Frame-2 ..A-***GTC-CAG-TCG-AAC***-T..
Frame-3 .AG-***TCC-AGT-CGA-ACT***-...

Os genes se encontram completamente em uma das duas fitas, sem sobreposição, isto é, existe uma sequência não codificante entre dois genes consecutivos chamada de região intergênica. Contudo um gene pode estar composto por várias sequências codificantes (CDSs) chamadas de éxons[8] que estão separadas por sequências não codificantes chamadas de íntrons[9]. Por isso é, extremamente importante identificar a fita e o quadro de leitura onde se encontram as CDS, assim como, onde começam e terminam as mesmas.

Para identificar as CDS, após o sequenciamento do genoma das espécies, são utilizadas técnicas de reconhecimento de padrões em sequências de DNA. As mais utilizadas

[8] Regiões codificantes do RNA mensageiro.
[9] Regiões não-codificantes do RNA mensageiro.

inicialmente foram baseadas em Cadeias de Markov[10] e redes neurais (BORODOVSKY e MCININCH, 1993). Posteriormente, vários modelos probabilísticos de múltiplos estágios chamados Modelos Oculto de Markov (*Hidden Markov Models*), foram propostos, os quais fazem uma integração de toda a informação disponível sobre a estrutura dos genes (KROGH *et al*, 1994; BALDI e BRUNAK, 2001). A taxa de sucesso desse modelo depende da representatividade do conjunto de treinamento utilizado para calibrar o modelo probabilístico (ABU-HANNA, 1999). O grande diferencial dos HMMs é o uso do conceito de auto-formação, que integra informações estruturais genéricas permitindo estender a aplicabilidade do modelo em busca de genes no genoma com pouca ou nenhuma informação prévia. Nos estudos genômicos e proteômicos, é importante determinar corretamente todas as CDS que fazem parte de um mesmo gene. Isto não é uma tarefa fácil, pois um CDS pode estar mais perto de um gene vizinho do que da CDS mais próximo do próprio gene. Contudo, quando o foco é o uso de códons, o que é mais relevante é a correção da fita, do quadro de leitura, da posição e tamanho das CDS, pois erros na estimação destes dados podem causar grandes erros na contagem dos diferentes tipos de códons o que impacta no cálculo das frequências relativas destes.

Para minimizar estes erros se realiza primeiro uma varredura dos seis quadros de leitura em uma sequência de DNA de entrada buscando por regiões abertas de leitura (*open reading frames* ou ORFs). Uma ORF é um fragmento de uma sequência de DNA que contem um número finito de tripletos de nucleotídeos começando com um códon de início de transcrição (normalmente ATG) e terminando em um códon de finalização da tradução (TAA, TAG ou TGA). Embora não todas as ORFs são codificantes, todas as CDS estão contidas em uma ORF.

Carels e colaboradores (2009) desenvolveram um método estatístico eficaz para classificar as ORFs como codificantes ou não. Tratando-se de um método baseado em padrões estatísticos da frequência de códons, sua eficácia (acurácia) depende do comprimento (dado em bases nitrogenadas) das ORFs analisadas. O método em questão classificou corretamente 95% das ORFs contendo, pelo menos, entre 150 e 200 bases em P. *falciparum* e C. *reinhardtii*, 300 bases em D. *melanogaster* e 350 bases em

[10] Cadeia de Markov é uma sequencia de variáveis aleatórias, chamada assim em homenagem ao matemático russo Andrei Andreyevich Markov.

H. *sapiens* (CARELS et al, 2009). O método proposto por Carels *et al* permitiu uma melhor sensibilidade na classificação de éxons *vs* íntrons em ORFs de 50 a 150 bases em relação a outros métodos, confirmando ser independente do uso de códons e da espécie, não precisando de uma etapa de treinamento. Por isso este método pode ser útil na extração de ORFs codificantes para montar conjuntos de treinamento para os métodos de predição de genes *ab-initio*[11] que apreendem padrões, no estudo de genomas para os quais há pouca informação prévia disponível (CARELS *et al*, 2009). No âmbito do nosso projeto, este método será utilizado para o controle de qualidade das CDS que foram mineradas dos bancos de dados públicos, antes de serem utilizadas para o cálculo do uso de códons, nos genes e nos genomas dos organismos e patógenos estudados neste trabalho.

2.6. A mineração de informações para a modelagem dos Bancos de Dados Biológicos (BDB).

Um banco de dados (BD) é formado por coleções de informações que se relacionam entre si para criar um sentido. Um Sistema de Gerenciamento de Bancos de Dados (SGBD), é um conjunto de *softwares* com objetivo de gerenciar o acesso, a manipulação e organização das informações, disponibilizando uma interface para o usuário manipular os dados. Para criação de um banco de dados especifico (BDE) será necessária a análise das informações que deverão compô-lo, para a elaboração e execução do seu modelamento.

Os bancos de dados biológicos (BDB) são compostos por tabelas que relacionam-se entre si, armazenando uma grande quantidade de registros. As informações contidas neste banco de dados são registros de uma determinada sequência de nucleotídeo, essa sequência normalmente possui uma descrição do nome cientifico, com as citações na leitura correspondente da sequência. Os BDB possuem a mesma modelagem dos bancos de dados relacionais ou orientados a objetos, apenas o que os caracterizam como biológico são as informações contidas nele. Esses BDB geralmente são associados a um *software* de interface desenvolvido para realização das quatros operações básicas conhecida por CRUD's (*create, read, update, delete,* i.e *insert, select, update* e *delete,* respectivamente) (BIOINFORMATICS FACTSHEET, 2011).

[11] Palavra em latim com significado desde princípio

Uma meta-informação é constituída por características da uma sequência genômica, onde seu objetivo é a tradução dos dados em informações biologicamente importantes. (LEMOS 2004; WEISS, 2010). Essas meta-informações são uma descrição de características em mais alto nível da biossequência. Meta-informações úteis contêm vários tipos de informações, como exemplos, um trecho de DNA que contém um gene e a sua função (LEMOS, 2004).

Há dois tipos de classificação para BDB, primários e secundários. Os primários são constituídos pela colocação direta de sequências de nucleotídeos, aminoácidos ou estruturas proteicas, sem qualquer processamento ou análise prévia dessas informações. Como exemplo desses BDB podemos citar o banco de dados públicos do *GenBank* pertencente ao *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) sob domínio do *International Nucleotide Sequence Database Collaboration* (INSDC). Os secundários são aqueles que derivam dos bancos de dados primários, ou seja, são constituídos utilizando informações específicas que são coletadas dos bancos de dados públicos primários (PROSDOCIMI *et al*., 2002, p.14).

Segundo Elmasri e Navathe (2005), os BDB precisam 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 base;
- 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.

Devido a que os bancos de dados públicos primários não realizam nenhum tipo de processamento ou análise prévia, as redundâncias e/ou inconsistências das informações são irremissíveis, pois os laboratórios que alimentam esses bancos possuem critérios particulares sobre a qualidade das biossequências a serem publicadas. Com isso, alguns dados armazenados apresentam erros, por possuírem sequências incompletas, corrompidas, e com erros devidos a falhas vindas do próprio sequenciamento, e mesmo assim elas são submetidas a estes bancos de dados.

Por este motivo o projeto implantou o método *Universal Feature Method* (UFM) (CARELS e FRIAS, 2009 e 2013) para controle de qualidade das sequências de vírus humanos mineradas pelo robô desenvolvido neste projeto.

2.7. Descoberta de conhecimento e *data mining*

Com a rápida evolução dos recursos computacionais ocorridas nos últimos anos permitiu que fossem geradas, também, grandes volumes de dados armazenados. Estima-se que a quantidade de informação no mundo dobra a cada 20 meses e que o tamanho e a quantidade armazenada nos bancos de dados crescem em uma velocidade maior ainda (DILLY, 1999). O crescimento exponencial desse volume de dados tem gerado uma urgente necessidade de novas técnicas e ferramentas capazes de transformar, de forma inteligente e automática, a grande massa de dados em informações valiosas. As informações extraídas dos bancos de dados são de grande valia para a tomada de decisões, essas informações na verdade estão implícitas e/ou escondidas sob uma montanha de dados, e não podem ser facilmente identificadas utilizando-se sistemas convencionais de gerenciamento de bancos de dados. A partir dessa necessidade surgiu a mineração de dados, denominada de *data mining*.

A *data mining* é uma parte do processo conhecido como Descoberta de Conhecimento de Bases de Dados ou *Knowledge Discovery in Database (KDD)*. Este conceito surgiu nos anos 80 para dar vazão ao grande volume de dados que se expandiam exponencialmente, sendo necessário para automatizar a exploração, reconhecimento padrões na modelagem das informações. O KDD é uma tecnologia que surgiu da interseção das áreas da estatística clássica, inteligência artificial e aprendizado de máquina. Segundo Addrians e Zantinge (1996) o KDD permite uma extração não trivial

de conhecimento previamente desconhecido e potencialmente útil de um banco de dados. Esse conceito é ressaltado por Fayyad e seus colaboradores (1996b) afirmando que a mineração de dados é um processo não trivial de identificações de padrões, desconhecidos, potencialmente úteis e no final das contas, compreensíveis em dados.

Considerando uma hierarquia de complexidade, se algum significado em especial é atribuído a um dado qualquer, esse dado se transforma em uma informação ou fato. Para Sade (1996) se uma norma ou regra é elaborada, a interpretação do confronto entre o fato e a regra constitui em um conhecimento. O processo KDD é constituído de várias etapas, (Figura 9), que são executadas de forma interativa e iterativa. As etapas são interativas porque envolvem a cooperação da pessoa responsável pela análise de dados, cujo conhecimento sobre o domínio orientará a execução do processo (BRACHMAN e ANAND, 1996). Por sua vez, a iteração deve-se ao fato de que, com frequência, esse processo não é executado de forma sequencial, mas envolvem repetidas seleções de parâmetros e conjunto de dados, aplicações das técnicas de *data mining* e posterior análise dos resultados obtidos, a fim de refinar os conhecimentos extraídos.

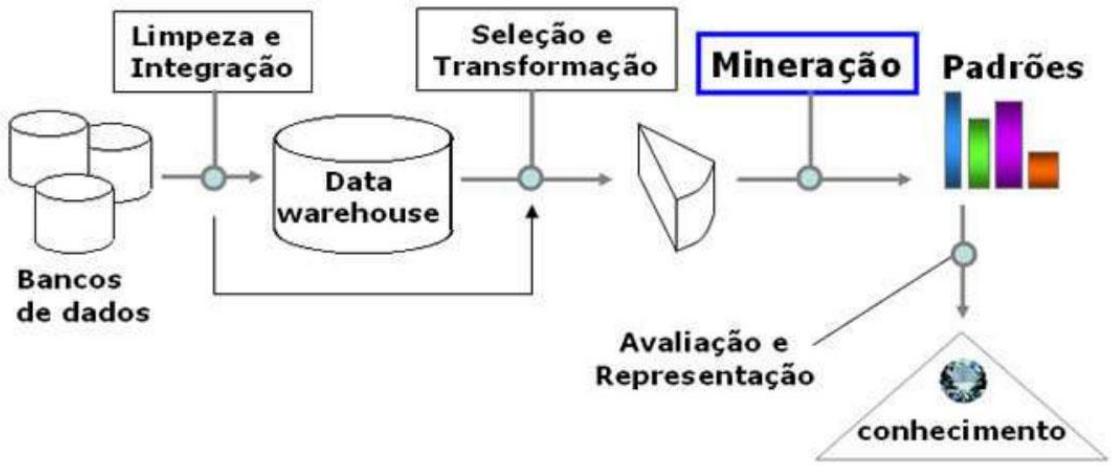

Figura 9 - As etapas do processo de KDD.
Fonte: http://www.lsi.ufu.br/documentos/publicacoes/ano/2004/JAI-cap5.pdf

Esse processo tem início com o entendimento do domínio da aplicação e dos objetivos a serem atingidos. Em seguida, é realizado um agrupamento organizado da massa de dados alvo da descoberta. Como em toda análise quantitativa, a qualidade dos dados é essencial para a obtenção de resultados confiáveis. Portanto, dados limpos e compreensíveis são requisitos básicos para o sucesso da mineração de dados (DINIZ e

LOUZADA-NETO, 2000). A limpeza dos dados, identificada na literatura como *Data Cleaning* é realizada por meio de um pré-processamento, visando assegurar a qualidade dos dados selecionados. Segundo Mannila (1996), essa etapa pode tomar até 80% do tempo necessário para todo o processo, devido às dificuldades de integração de bases de dados heterogêneas.

Os dados pré-processados devem passar por outra transformação, que os armazena adequadamente, visando facilitar o uso das técnicas de *Data Mining*. Nessa fase, o uso de *Data Warehouses* expande-se consideravelmente, já que, nessas estruturas, as informações estão alocadas da maneira mais eficiente. O *Data Warehouse* como um depósito central de dados, extraído de dados operacionais, em que a informação é orientada a assuntos, não volátil e de natureza histórica (ADDRIANS e ZANTINGE, 1996). Devido a essas características o *Data Warehouses* tendem a se tornar grandes repositórios de dados extremamente organizados, facilitando a aplicação do *Data Mining*.

Prosseguindo no processo KDD, chega-se especificamente à fase de *Data Mining*. O objetivo principal desse passo é a aplicação de técnicas de mineração nos dados pré-processados, o que envolve ajuste de modelos e/ou determinação de características nos dados. Em outras palavras, exige o uso de métodos inteligentes para a extração de padrões ou conhecimentos dos dados.

É importante destacar que cada técnica de *Data Mining* utilizada para conduzir as operações de mineração de dados adapta-se melhor a alguns problemas do que a outros, o que impossibilita a existência de um método de *Data Mining* universalmente melhor. Para cada problema particular, tem-se uma técnica particular.

Portanto, o sucesso de uma tarefa de *Data Mining* está diretamente ligado à experiência e à intuição do analista. A etapa final do processo de mineração consiste no pós-processamento, que engloba a interpretação dos padrões descobertos e a possibilidade de retorno a qualquer um dos passos anteriores. Assim, a informação extraída é analisada (ou interpretada) em relação ao objetivo proposto, sendo identificadas e apresentadas as melhores informações. Dessa forma, o propósito do resultado não consiste somente em visualizar, gráfica ou logicamente, o rendimento da *data mining*,

mas, também, em filtrar a informação que será apresentada, eliminando possíveis ruídos, ou seja, padrões redundantes ou irrelevantes que podem surgir no processo.

A mineração de dados, pode ser entendido como o processo de extração de informações, sem conhecimento prévio, de um grande banco de dados e seu uso para tomada de decisões. É uma metodologia aplicada em diversas áreas que usam o conhecimento, como empresas, indústrias e instituições de pesquisa. *Data Mining* define o processo automatizado de captura e análise de grandes conjuntos de dados para extrair um significado, sendo usado tanto para descrever características do passado como para predizer tendências para o futuro.

Para encontrar respostas ou extrair conhecimento interessante, existem diversos métodos de *Data Mining* disponíveis na literatura. Mas, para que a descoberta de conhecimentos seja relevante, é importante estabelecer metas bem definidas. Essas metas são alcançadas por meio dos seguintes métodos de mineração de dados: classificação, modelos de relacionamento entre variáveis, análise de agrupamento, sumarização, modelo de dependência, regras de associação e análise de séries temporais (FAYYAD *et al*, 1996a). É importante ressaltar que a maioria desses métodos é baseada em técnicas das áreas de aprendizado de máquina, reconhecimento de padrões e estatística. Essas técnicas vão desde as tradicionais da estatística multivariada, como análise de agrupamentos e regressões, até modelos mais atuais de aprendizagem, como redes neurais, lógica difusa e algoritmos genéticos.

Os métodos tradicionais de *Data Mining* são:

- *Classificação*: associa ou classifica um item a uma ou várias classes categóricas pré-definidas. Uma técnica estatística apropriada para classificação é a análise discriminante. Os objetivos dessa técnica envolvem a descrição gráfica ou algébrica das características diferenciais das observações de várias populações, além da classificação das observações em uma ou mais classes predeterminadas. A ideia é derivar uma regra que possa ser usada para classificar, de forma otimizada, uma nova observação a uma classe já rotulada. Para Mattar (1998), a análise discriminante permite que dois ou mais grupos possam ser comparados, com o objetivo de determinar se diferem uns dos outros e, também, a natureza da diferença, de forma que, com base em um conjunto de variáveis

independentes, seja possível classificar indivíduos ou objetos em duas ou mais categorias mutuamente exclusivas.

- *Modelos de Relacionamento entre Variáveis*: associa um item a uma ou mais variáveis de predição de valores reais, consideradas variáveis independentes ou exploratórias. Técnicas estatísticas como regressão linear simples, múltipla e modelos lineares por transformação são utilizadas para verificar o relacionamento funcional que, eventualmente, possa existir entre duas variáveis quantitativas, ou seja, constatar se há uma relação funcional entre X e Y. Observa-se que o método dos mínimos quadrados ordinários, atribuído a Carl Friedrich Gauss, tem propriedades estatísticas relevantes e apropriadas, que tornaram tal procedimento um dos mais poderosos e populares métodos de análise de regressão conforme (GUJARATI 2000).
- *Análise de Agrupamento* (*Cluster*): associa um item a uma ou várias classes categóricas ou *cluster*s, em que as classes são determinadas pelos dados, diversamente da classificação em que as classes são pré-definidas.
- Os *cluster*s são definidos por meio do agrupamento de dados baseados em medidas de similaridade ou modelos probabilísticos. A análise de *cluster* ou agrupamento é uma técnica que visa detectar a existência de diferentes grupos dentro de um determinado conjunto de dados e, em caso de sua existência, determinar quais são eles. Nesse tipo de análise, o procedimento inicia com o cálculo das distâncias entre os objetos estudados dentro do espaço multiplano constituído por eixos de todas as medidas realizadas (variáveis), sendo, a seguir, os objetos agrupados conforme a proximidade entre eles (PEREIRA, 1999). Na sequência, efetuam-se os agrupamentos por proximidade geométrica, o que permite o reconhecimento dos passos de agrupamento para a correta identificação de grupos dentro do universo dos objetos estudados.
- *Sumarização*: determina uma descrição compacta para um dado subconjunto. As medidas de posição e variabilidade são exemplos simples de sumarização. Funções mais sofisticadas envolvem técnicas de visualização e a determinação de relações funcionais entre variáveis. As funções de sumarização são frequentemente usadas na análise exploratória de dados com geração automatizada de relatórios, sendo responsáveis pela descrição compacta de um conjunto de dados. A sumarização é utilizada, principalmente, no pré-

processamento dos dados, quando valores inválidos são determinados por meio do cálculo de medidas estatísticas, como mínimo, máximo, média, moda, mediana e desvio padrão amostral, no caso de variáveis quantitativas, e, no caso de variáveis categóricas, por meio da distribuição de frequência dos valores. Técnicas de sumarização mais sofisticadas são chamadas de visualização, que são de extrema importância e imprescindíveis para se obter um entendimento, muitas vezes intuitivo, do conjunto de dados. Exemplos de técnicas de visualização de dados incluem diagramas baseados em proporções, diagramas de dispersão, histogramas e *box plots*, entre outros. Autores como Levine *et al.* (2000) e Martins (2001), abordam com grande detalhamento esses procedimentos metodológicos.

- *Modelo de Dependência*: descreve dependências significativas entre variáveis. modelos de dependência existem em dois níveis: estruturado e quantitativo. O nível estruturado especifica, geralmente em forma de gráfico, quais variáveis são localmente dependentes. O nível quantitativo especifica o grau de dependência, usando alguma escala numérica. Análises de dependência são aquelas que têm por objetivo o estudo da dependência de uma ou mais variáveis em relação a outras, sendo procedimentos metodológicos para tanto a análise discriminante, a de medidas repetidas, a de correlação canônica, a de regressão multivariada e a de variância multivariada (PADOVANI, 1995).
- *Regras de Associação*: determinam relações entre campos de um banco de dados. A ideia é a derivação de correlações multivariadas que permitam subsidiar as tomadas de decisão. A busca de associação entre variáveis é, frequentemente, um dos propósitos das pesquisas empíricas. A possível existência de relação entre variáveis orienta análises, conclusões e evidenciação de achados da investigação. Uma regra de associação é definida como *se X então Y*, ou $X \Rightarrow Y$, onde X e Y são conjuntos de itens e $X \cap Y = \emptyset$. Diz-se que X é o antecedente da regra, enquanto Y é o seu consequente. Medidas estatísticas como correlação e testes de hipóteses apropriados revelam a frequência de uma regra no universo dos dados minerados. Vários métodos para medir associação são discutidos por Mattar (1998), de natureza paramétrica e não paramétrica, considerando a escala de mensuração das variáveis.
- *Análise de Séries Temporais*: determina características sequenciais, como dados com dependência no tempo. Seu objetivo é modelar o estado do processo

> extraindo e registrando desvios e tendências no tempo. Correlações entre dois instantes de tempo, ou seja, as observações de interesse, são obtidas em instantes sucessivos de tempo, por exemplo, a cada hora, durante 24 horas, ou são registradas por algum equipamento de forma contínua, como um traçado eletrocardiográfico. As séries são compostas por quatro padrões: tendência, variações cíclicas, variações sazonais e variações irregulares. Há vários modelos estatísticos que podem ser aplicados a essas situações, desde os de regressão linear (simples e múltiplos), os lineares por transformação e regressões assintóticas, além de modelos com defasagem, como os auto regressivos (AR) e outros deles derivados. Uma interessante noção introdutória ao estudo de séries temporais é desenvolvida por Morettin e Toloi (1987).

Diante da descrição sumária de metodologias estatísticas aplicáveis ao procedimento de mineração de dados, registra-se que, embora Hand (1998) afirme que o termo *Data Mining* possa trazer uma conotação simplista para os estatísticos, (FAYYAD *et al.* 1996a) mostraram a relevância da estatística para o processo de extração de conhecimentos, ao afirmar que essa ciência provê uma linguagem e uma estrutura para quantificar a incerteza resultante quando se tenta deduzir padrões de uma amostra a partir de uma população.

A estatística preocupa-se com a análise primária dos dados, no sentido de que eles são coletados por uma razão particular ou por um conjunto de questões particulares *a priori* (HAND 1998). *Data Mining*, por outro lado, preocupa-se também com a análise secundária dos dados, num sentido mais amplo e mais indutivo do que uma abordagem hipotético-dedutiva, frequentemente considerada como o paradigma para o progresso da ciência moderna. Assim, *Data Mining* pode ser visto como o descendente direto da estatística, já que são técnicas metodológicas complementares.

3. Materiais e Métodos

O projeto inicia-se com o entendimento de tópicos especiais nas seguintes temáticas:

- Estrutura e mecanismos de acesso automatizado a bancos públicos para *download* de:
    - sequências codificantes (CDS) de organismos hospedeiros e vírus.
    - frequências de códons (CUT) de organismos hospedeiros e vírus.
    - frequências genômicas de genes de espécies de RNA transportador (tRNA) de organismos hospedeiros.
- Técnicas estatísticas de reconhecimento de padrões para detecção e extração de sequências codificantes no quadro de leitura (CDS) de sequências de DNA genômico ou RNA.
- Papel e funcionamento do RNA transportador no processo de **tradução** de RNA (síntese de proteínas). Pareamento códon-anticódon no ribossomo e pareamento não clássico (*wobble*).
- Mecanismos de transcrição e de tradução. Perfis de expressão nos níveis de tradução e de transcrição. Bancos de dados de expressão.

3.1. Escolha da Plataforma Computacional e a Mineração de Dados

Após uma revisão bibliográfica, com foco em repositório de sequências virais, sistemas de gerenciamento de bancos de dados (SGBD) baseado na linguagem SQL (*Structured Query Language*), optou-se utilizar o banco de dados MySQL na sua versão 5.1, que utilizam de bases de dados relacionais e como linguagem de programação a utilização do PHP (*Hypertext Preprocessor*) na versão 5.4 para criação da aplicação de bioinformática, utilizando o Servidor HTTP Apache na versão 2.2, para interpretar os códigos PHP da ferramenta *Web*.

Desde o ponto de vista formal, a pesquisa foi realizada pelo método quantitativo. Essa abordagem de pesquisa visa enfatiza o desenvolvimento da investigação dentro dos métodos estabelecidos e técnicas específicas. Baseando no teste da hipótese a fim de expressar os fenômenos existentes, a partir dos dados coletados. A pesquisa quantitativa é apropriada quando há a possibilidade de medidas quantificáveis de variáveis e inferências a partir de amostras dos dados, medidas numéricas ou busca padrões que poderão ser encontrados nos bancos de dados públicos.

De fato, o procedimento proposto é uma forma de extração baseadas nos procedimentos dos processos do KDD. A mineração do conhecimento utilizando técnicas do KDD é essencial a esta pesquisa, pois nela consiste aplicações de técnicas próprias inteligentes, podendo detectar e colher informações mais profundas "escondidas" extraindo, assim, dos bancos de dados públicos informações relacionadas com a interação dos patógenos virais com seus hospedeiros.

As principais tecnologias utilizadas no KDD são as de *interfaces*, distribuição de banco de dados, organização de dados (*data warehousing*), redes neurais e sistemas especialistas. Utilizam-se estas tecnologias devido a cada tarefa ser dependente da intenção do usuário e também a mesma ser responsável por uma informação específica que se refere a um algoritmo específico para tratar cada tipo de informação coletada.

A ferramenta de bioinformática, elaborada nesse projeto, existem duas finalidades distintas, a primeiro funcionalidade é a coleta, controle de qualidade e povoamento do BDE, esta funcionalidade será explicada melhor nas sessões 3.2 e 3.3, e a segundo é a página *web* para interação com o usuário. A página *web* inicial da ferramenta foi construída na linguagem *HyperText Markup Language* (HTML) contendo um formulários com campos específicos para realização das busca, permitindo que o usuário faça diversas combinações de dados diferentes. A partir do momento que o ele solicita uma busca, a pagina inicial envia uma requisição para um *script* escrito em PHP, responsável em tratar as informações e posteriormente, utilizá-las para executar a busca nos dados armazenados no BDE. Com a posse dos resultados obtidos por meio da comunicação realizada com o BDE, outro *script* também escrito em PHP irá organizá-los em outra página escrita em HTML, permitindo a visualização e *download* destes resultados obtidos, por meio das combinações escolhidas (Figura 10).

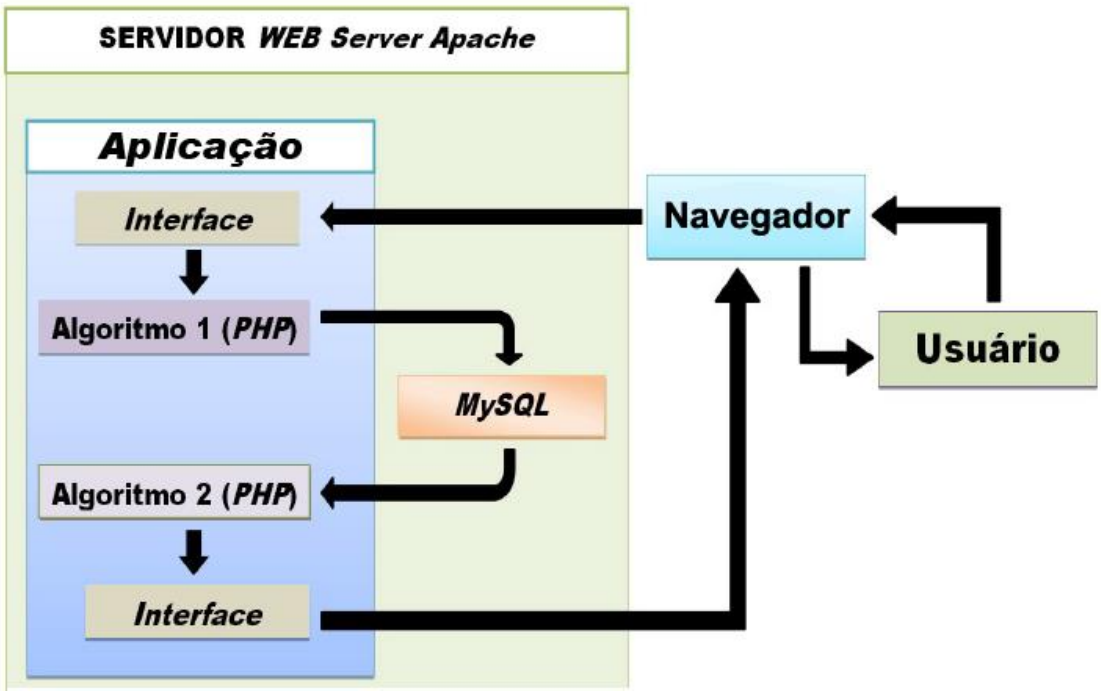

Figura 10 - Arquitetura da ferramenta WEB do BDE

3.2. Coleta, Mineração, Controle de Qualidade e Armazenamento de Sequências Codificantes de Retrovírus.

As sequências codificantes do HIV e HTLV e suas respectivas meta-informações foram coletadas de um BDP chamado *GenBank*, no formato *Extensible Markup Language* (XML), contendo nesse arquivo as sequências de nucleotídeos e as meta-informações que tem apoio bibliográfico e biológico, mantido pelo *National Center for Biotechnology Information* (NCBI) do *National Institutes of Health* (BENSON, 2010).

O primeiro passo foi entender a estrutura do *Genbank* e os mecanismos e ferramentas para minerar sequências contidas nele. Em particular estudamos três *frameworks* para construir um Sistema Configurável Automático de Mineração (SCAM) para realização da mineração, controle de qualidade e armazenamento dos dados genômicas (CDS filtradas e validadas *in-silico*).

O primeiro *framework* estudado foi o BioPerl (LEHVASLAIHO, 2007), escrito na linguagem de script PERL, juntamente com o *software* ACNUC (GOUY *et al*, 1985, GOUY *et al*, 2008). Segundo framework estudado foi o BioJava escrito na linguagem Java, e o terceiro *framework* estudado foi o BioPHP, escrito na linguagem PHP. O

framework BioPerl é o primeiro *framework* de bioinformática construído pela comunidade acadêmica a fim de dar subsídios aos pesquisadores, e, atualmente ele é o mais difundido atualmente, os demais *frameworks* existentes tais como: BioJava e BioPHP surgiram posteriormente a ele, e possuem funcionalidades iguais ao BioPerl, devido a realização de tradução de seus códigos fontes feito na linguagem PERL para as linguagens de cada *framework* em questão.

Após o entendimento da estrutura do *GenBank*, utilizou-se o *framework* BioPHP, apensar do mesmo não ser atualizado desde do ano de 2003, porém seus códigos fontes foram fundamentais para implementação sem complicação das funcionalidades necessárias para a construção SCAM. O robô realizou a mineração, controle de qualidade e povoamento das sequências codificantes coletadas do BDP, como ilustra a estrutura do *GenBank* na Figura 11.

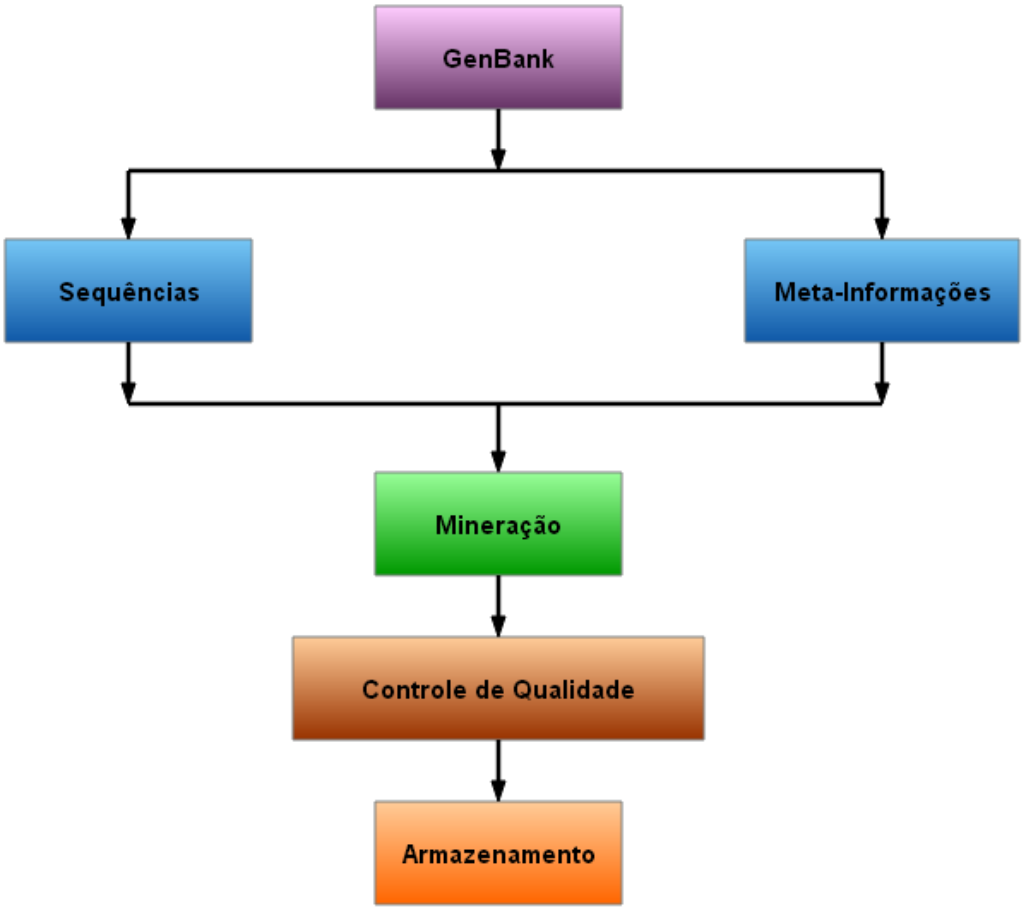

Figura 11 - Fluxo da coleta, mineração e armazenamento das sequências de códons do HIV e HTLV.

Para realização da atualização automática, porém controlada do BDE das sequências coletadas do *GenBank*, o algoritmo do SCAM, primeiramente, verifica todos os arquivos das sequências coletadas de um determinado tipo de vírus, que deseja atualizar, esses arquivos são atribuídas há uma variável indexada (*array*). Posteriormente são consultados no *GenBank* todos os números de identificações das sequências do referido tipo de vírus, que deseja atualizar, essas numerações são denominada pelo *GenBank* como "GI", o resultado dessa consulta vem em um arquivo com a extensão XML, nesse arquivo contém as numerações de identificação das sequências, do referido vírus, são atribuídas há uma nova variável do mesmo tipo da anterior. Após o preenchimento das duas variáveis elas são comparadas uma com a outra e as divergências encontradas, em sua comparação, são atribuídas há uma nova variável do mesmo tipo das anteriores criadas. Em seguida ao processo de comparação, essa nova variável é lida e todas os dados contidas nela serão utilizadas para a realização de nova consulta no *GenBank*, essa nova consulta é pelo numero de identificação da sequência (GI). O resultado dessa nova consulta vem em um novo arquivo no formato XML, a partir desse momento, o arquivo é baixado, atribuído um nome a esse arquivo e na seguida é colocado em uma determinada pasta junto com os outros, nesse arquivo contém as metas-informações e a sequência de nucleotídeos, essas informações são necessárias para povoamento do BDE, essa procedimento permite que o dado coletado não seja consultado novamente e a informação não seja duplicada no BDE. Essa funcionalidade é repetida para cada tipo de vírus.

As sequências depositadas no banco de dados com a região genômica que não tinham anotações ou informações no arquivo *XML*, fornecido pelo BDP sobre o subtipo e/ou subgrupo foram submetidas às ferramentas de bioinformática.

As informações faltantes da região genômica das sequências dos HTLV foram subtipadas utilizando a ferramenta *online LASP HTLV-1 Automated Subtyping Tool* (http://www.bioafrica.net/rega-genotype/html/subtypinghtlv.html), e do HIV foram subtipadas utilizando a ferramenta *online REGA HIV-1 & 2 Automated Subtyping Tool* (http://www.bioafrica.net/rega-genotype/html/subtypinghiv.html), essas ferramentas utilizam métodos filogenéticos para identificar o subtipo de sequências consultadas, e foram ambas projetadas pelo instituto BioAfrica (http://www.bioafrica.net/). Para a submissão das sequências nos *softwares*, foi necessário à construção de um algoritmo

com a capacidade transformar as sequências codificantes presentes no BDE no formato *fasta* aceito pelas ferramentas do BioAfrica. O arquivo *fasta* (LIPMAN e PEARSON, 1985) é caracterizado por uma ou mais sequências, contendo um cabeçalho com informações indicado pela presença do símbolo de maior (>) na primeira coluna e a organização das sequências com no máximo 60 nucleotídeos por linha, para facilitar o alinhamento das sequências.

3.3. Coleta, Mineração e Armazenamento de Frequências de Códons e Frequência de espécies de RNA transportador.

O próximo passo foi compreender a estrutura do *Codon Usage Database* fornecido pelo *Kazusa DNA Research Institute* (http://www.kazusa.or.jp/codon/) e os mecanismos e ferramentas para minerar a frequência de códons contida nele. Após o entendimento da sua estrutura foi implementado uma nova funcionalidade (*script*) ao SCAM para entender esse novo BDP, assim consegui realizar a coletar da frequência de utilização de códons do hospedeiro e vírus, e armazenamento das mesmas. A Figura 12 ilustra o fluxo da realização de tal procedimento.

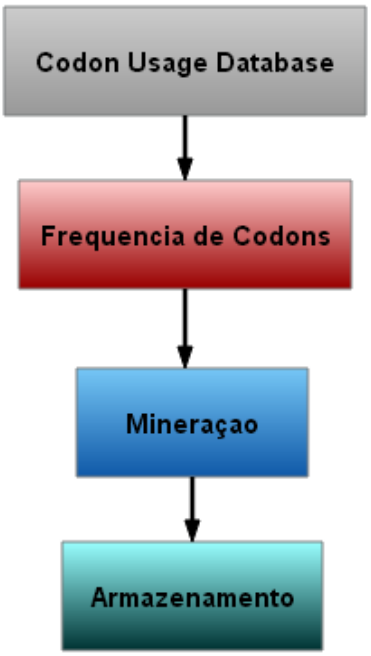

Figura 12 - Fluxo da coleta, mineração e armazenamento das frequências de códons do Hospedeiro.

Por último tivemos que entender a estrutura do *Genomic tRNA Database* para minerar, coletar e armazenar as frequências genômicas de genes de espécies de tRNA de

organismos hospedeiros, obtidas no mesmo. Foi realizada uma última adaptação no robô do SCAM adicionando um novo *script* para comtemplar a mineração e armazenamento do número total de frequências de espécies de RNA transportador nos genomas de hospedeiros eucariotos. A Figura 13, ilustra o fluxo de procedimentos para minerar e armazenar as informações necessárias.

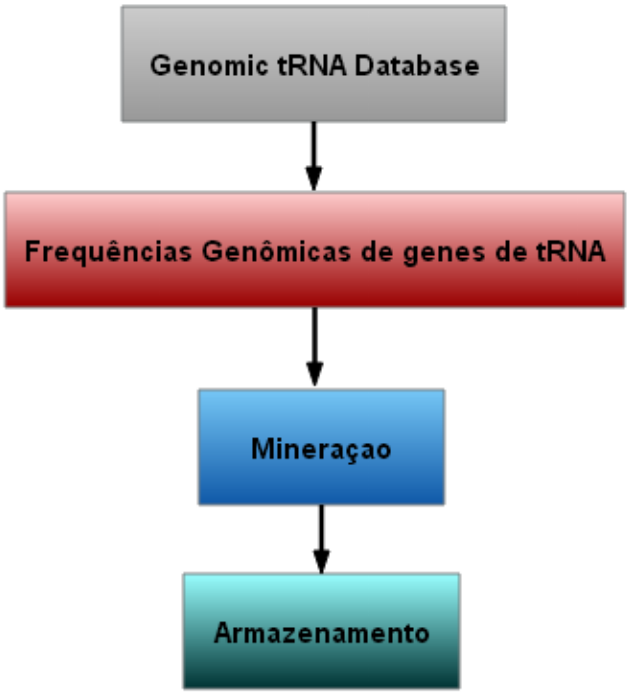

Figura 13 - Fluxo da coleta, mineração e armazenamento das frequências genômicas de genes de espécies de tRNA de organismos hospedeiros.

Como o *Genomic tRNA Database* realiza uma comparação com a frequência de códons no hospedeiro fornecida pelo *Codon Usage Database*. O algoritmo do SCAM, realiza uma leitura da página inicial, no formato HTML, do *Genomic tRNA Database* (Figura 14), procurando todas as informações sobre o hospedeiro *Homo sapiens*, que são necessárias a esse estudo, após encontrar esses dados ele é atribuído a uma variável indexada (*array*), verificando, posteriormente, qual informação sobre o hospedeiro consultado é a mais atual. Em seguida, o algoritmo lê uma nova página, também no formato HTML, do hospedeiro mais atual. Essa página contém todas as informações das frequências de anticódons do tRNA e as frequências de códons do *Homo sapiens* (Figura 15), após essa leitura dessa página o algoritmo reconhece os padrões e armazena no BDE todas as informações pertinentes.

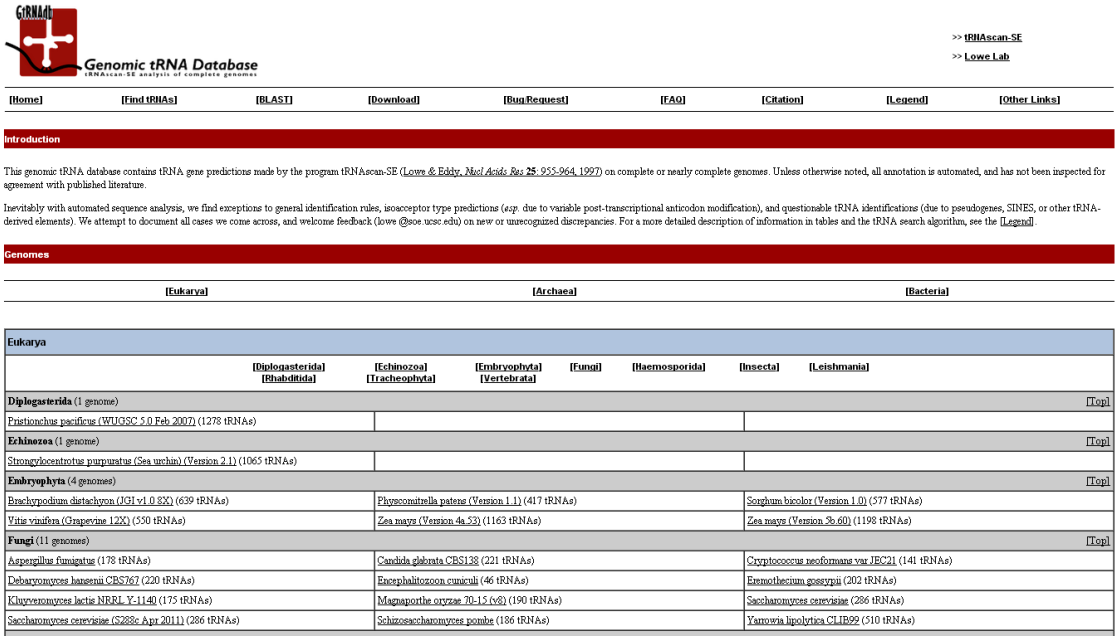

Figura 14 - Página inicial do *Genomic tRNA Database*

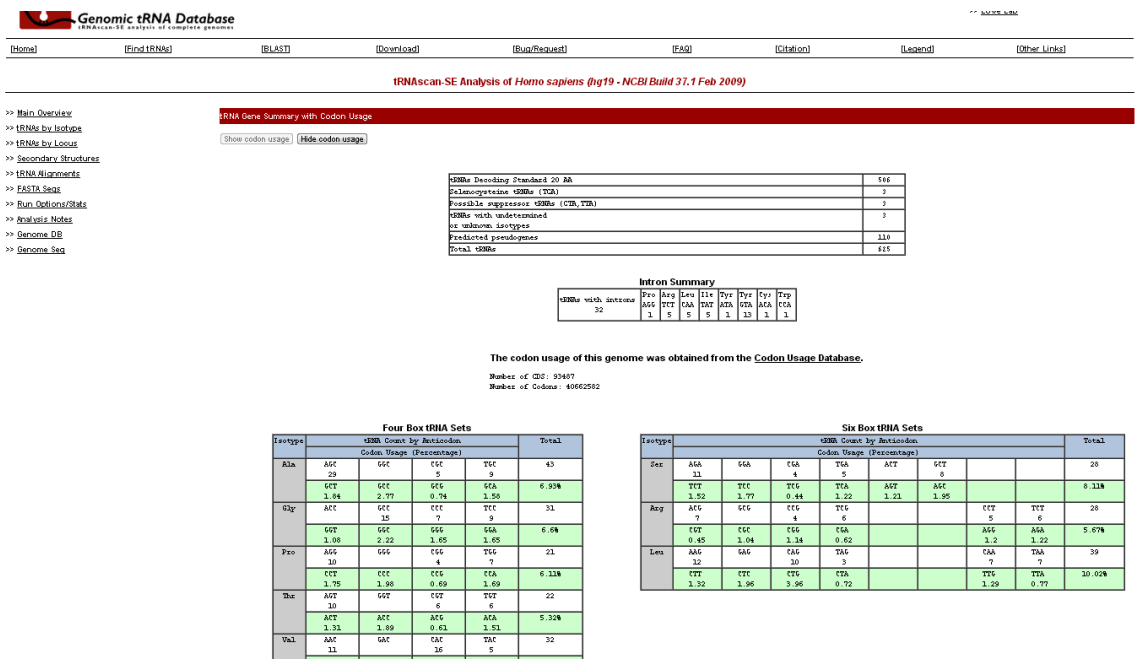

Figura 15 - Pagina com as informações sobre a frequência de códons

Estabeleceu um novo procedimento para gerar frequências de espécies cognatas de RNA transportador para cada códon, baseado em modelo e regras parametrizadas de compartilhamento de espécies de RNA transportador durante o processo de tradução de proteínas pelos códons sinônimos. O procedimento foi validado em relação à sensibilidade dos parâmetros em diferentes hospedeiros. Achando as distribuições de frequências de RNA transportador que apresentem a maior e a menor correlação com a

frequência de códons no hospedeiro, definindo o intervalo esperado de variação da distribuição buscada para ser considerado no estudo de sensibilidade.

A partir do cálculo das frequências de códons nas sequências codificantes (CDS) de genes virais, filtradas e armazenadas no BDE, determinou-se a correlação entre a frequência de códons nos genomas virais e nos genomas de hospedeiros, identificando-se códons com forte correlação negativa, isto é, códons muito frequentes no genoma viral e com pouca frequência no genoma do hospedeiro. Estes códons foram chamados de "códons candidatos para alvos terapêuticos (CCAT)". No caso em que existam dados consolidados de frequências de códons no genoma do vírus selecionado, o estudo verificou-se uma correlação com esses dados. Com base nos resultados obtidos, implementou-se uma nova funcionalidade para realizar o estudo da correlação cruzando a frequência de códons vírus-hospedeiro de espécies selecionadas por intermédio da interface WEB do BDE.

3.4. Modelagem e implementação do banco de dados.

O banco de dados foi modelado utilizando conceitos de SGBD *MySQL*, para o armazenamento das sequências codificantes, proveniente do *GenBank*, a frequência de códons de tRNA do *Genomic tRNA Database* e a frequência no uso de códons da espécie de hospedeiro do *Codon Usage Database*, foram criados ao todo 6 tabelas para auxiliar a aplicação.

- **Tabela BDE**
  - Número de Acesso
  - Numero da Versão do Acesso
  - Identificador Único da Sequência
  - Tipo do Vírus
  - Região Genômica (gene)
  - Título da Sequência (informação de é completa ou parcial)
  - Organismo do Vírus
  - Número de Acesso no *pubmed*
  - País de origem da sequência
  - Início da contagem dos códons

- Identificação proteica
- Tamanho da sequência
- Sequência codificante (CDS)
- Vetor de 64 elementos com o número de códons na sequência.

- **Tabela tRNA**
  - Códon
  - Valor
  - Total
- **Tabela de Códons**
  - Códon no formato do DNA (A, G, C, T)
  - Códon no formato do RNA (A, G, C, U)
  - Aminoácido formado pelo Códon
  - Sigla do aminoácido
  - Abreviação do aminoácido
- **Tabela Sequências completas**
  - Tipo do Virus
  - Identificador Único da Sequencia
  - Gene
  - Início da sequência
  - Término da sequência
- **Tabela Região Codificante**
  - Tipo do Vírus
  - Gene
  - Variável para identificar se possui a sequência no BDE
- **Tabela de atualização**
  - Data e hora da última verificação e atualização do sistema.

Na Figura 16 ilustra a modelagem do banco de dados no seu modelo lógico (BDL), dos dados coletados dos bancos de dados públicos.

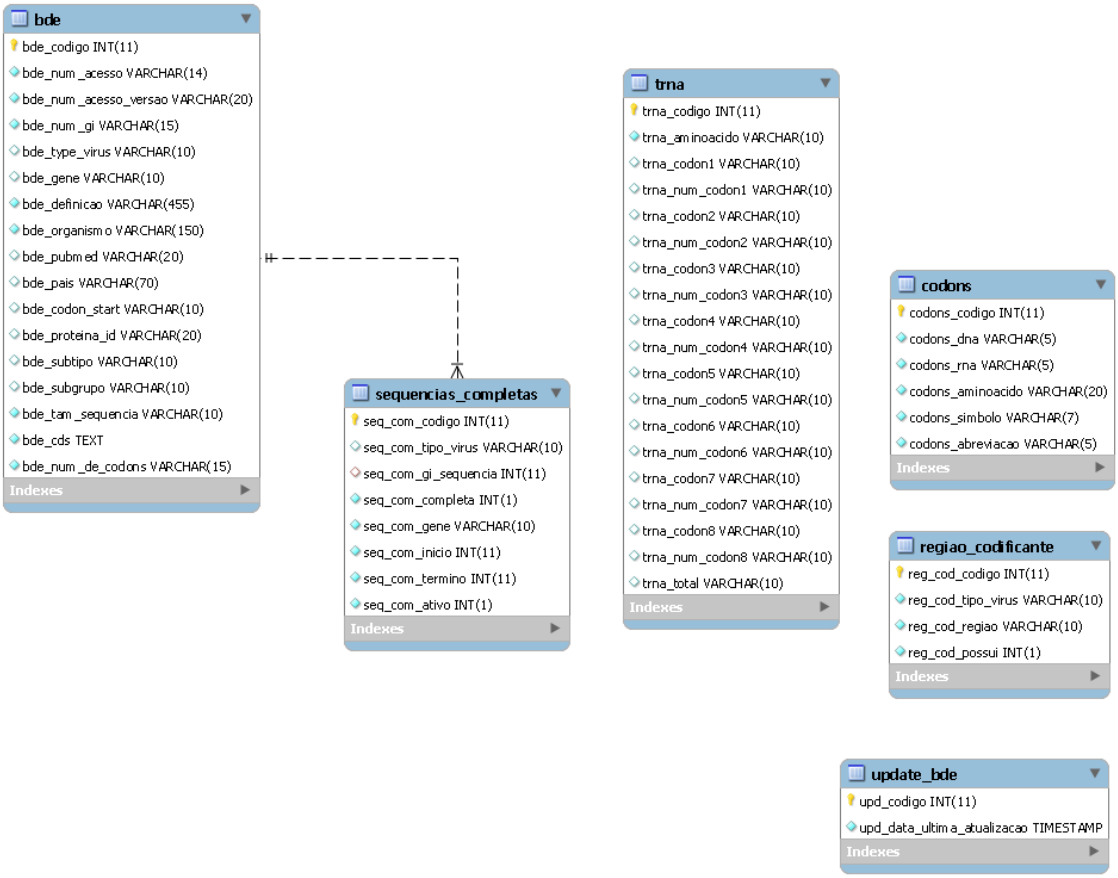

Figura 16 - Modelo lógico do banco de dados específico criado a partir dos da coleta dos dados do genbank, kazusa e genomic tRNA database.

3.5. Utilização dos dados de abundância de tRNA no processo de tradução para fins terapêuticos

Utilizando os dados de abundância genômica dos genes que codificam espécies de RNA transportador no hospedeiro, processada pelo modelo de compartilhamento, determinou-se a correlação entre a frequência de códons no genoma viral e a abundância relativa de espécies de RNA transportador, identificando-se códons com forte correlação negativa, isto é muito frequentes, mas com baixa abundância relativa de espécies de RNA transportador cognato ou específico. Estes códons foram chamados de "códons desfavorecidos no processo de tradução (CDPT)". Variando os parâmetros do modelo de compartilhamento de RNA transportador por códons sinônimos foi analisado o impacto deste modelo na seleção dos CDPT. Em particular, identificou-se um conjunto de CDPT comum a todas as distribuições de RNA transportador compreendidas entre as distribuições mais e menos correlacionadas com o genoma do

hospedeiro. Foi implementado uma nova funcionalidade ao portal WEB, para que os usuários possam realizar estudos de correlação cruzando a frequência de códons no vírus e a frequência de RNA transportador no hospedeiro selecionado no formulário principal da interface do portal.

Por meio da intercepção dos conjuntos de códons CCAT e CDPT foram identificados os códons que são os melhores alvos para uma terapia antiviral baseada no bloqueio seletivo das espécies cognatas de RNA transportador. Estes códons foram chamados de Códons Alvos de Terapia por Inibição de RNA transportador (CATIRT). Foi implementado uma nova funcionalidade para que os usuários possam realizar estudos identificando os CATIRTs para qualquer par de vírus-hospedeiro selecionado.

Por último visando caracterizar melhor o potencial dos CATIRTs como alvo terapêutico foi testado um indicador quantitativo e que estudou a sensibilidade desse indicador da configuração do modelo de compartilhamento de espécies de RNA transportador entre códons sinônimos durante o processo de tradução.

3.6. Ordenação dos códons e representação de código genético

Os códons foram ordenados de acordo com o tipo de sequências de nucleotídeos (purina ou pirimidina) nas diferentes posições de códons. As purinas vêm em primeiro lugar, o A antes de G, seguido de pirimidinas, o C antes de U. Assim, denotando por *ord(x)* o ordinal de uma dada base de *x*, temos *ord(A)* = 1, *ord(G)* = 2, *ord(C)* = 3, e *ord(U)* = 4. Isto implica que AAA é o primeiro, *ord(AAA)* = 1, enquanto que o último é UUU, *ord(UUU)* = 64 . Para um códon $b_1b_2b_3$ genérico, de tal modo que b_i = {A, G, C, U} para i = 1, 2, 3, temos $ord(b_1b_2b_3) = ord(b_3) + 4(ord(b_2)-1) + 16(ord(b_1)-1)$, desta forma, os primeiros 32 códons que começam com a purina e os últimos 32 códons com pirimidina. Além disso, este a ordenação de códon levar a um novo arranjo da tabela de conversão mostrado na Tabela 4 que se assemelha ao esquema de classificação do código genético proposto por Wilhelm e Nikolajewa (2004). O esquema da ordenação de códon adotada fornece melhores percepções sobre a simetria do box-estrutura subjacente que reflete os despedimentos no código genético. No entanto, neste caso, foi útil na identificação de regularidades de abundância de RNA transportador no genoma humano, bem como uma composição de códons atípica do HTLV.

Tabela 4 - Número de códons dos genes de tRNA em cognatos em Homo Sapiens por espécie códon.

n°	Codon	aa	G	g*	n°	Codon	Aa	G	g*
1	AAA	Lys	16	16	**33**	CAA	Gln	11	11
2	AAG		17	17	**34**	CAG		20	20
3	AAC	Asn	32	18	**35**	CAC	His	11	6,4
4	AAU		2	16	**36**	CAU		0	4,6
5	AGA	Arg	6	6	**37**	CGA	Arg	6	6
6	AGG		5	5	**38**	CGG		4	4
7	AGC	Ser	8	4,9	**39**	CGC	Arg	0	4,9
8	AGU		0	3,1	**40**	CGU		7	2,1
9	ACA	Thr	6	6	**41**	CCA	Pro	7	7
10	ACG		6	6	**42**	CCG		4	4
11	ACC	Thr	0	5,9	**43**	CCC	Pro	0	5,3
12	ACU		10	4,1	**44**	CCU		10	4,7
13	AUA	Ile	5	5	**45**	CUA	Leu	3	3
14	AUG	Met	20	20	**46**	CUG		10	10
15	AUC	Ile	3	9,6	**47**	CUC	Leu	0	7,2
16	AUU		14	7,4	**48**	CUU		12	4,8
17	GAA	Glu	13	13	**49**	UAA	Stop	-	-
18	GAG		13	13	**50**	UAG		-	-
19	GAC	Asp	19	10,1	**51**	UAC	Tyr	14	8,3
20	GAU		0	8,9	**52**	UAU		1	6,7
21	GGA	Gly	9	9	**53**	UGA	Stop	-	-
22	GGG		7	7	**54**	UGG	Trp	9	9
23	GGC	Gly	15	10,1	**55**	UGC	Cis	30	16,3
24	GGU		0	4,9	**56**	UGU		0	13,7
25	GCA	Ala	9	9	**57**	UCA	Ser	5	5
26	GCG		7	7	**58**	UCG		4	4
27	GCC	Ala	0	17,4	**59**	UCC	Ser	0	5,9
28	GCU		29	11,6	**60**	UCU		11	5,1
29	GUA	Val	5	5	**61**	UUA	Leu	7	7
30	GUG		16	16	**62**	UUG		7	7
31	GUC	Val	0	6,3	**63**	UUC	Phe	12	6,4
32	GUU		11	4,7	**64**	UUU		0	5,6

3.7. O uso de comparação de códons

O padrão de uso de códons de uma espécie *s* é, em primeira instância, dado pela lista das frequências genômicas dos códons, $O \leq C_{i,s} \leq 1, onde\ i = 1,2, \ldots 64$, de tal forma que $\sum_i C_{i,s} = 1$. A frequência genômica do códon *i* é calculada como a razão entre as contagens de códons, N_i, no genoma e o número total de códons onde $N = \sum_j N_j$, que é $C_i = N_i/N$. As frequências de códons de parada (*stop-codon*) são nas posições 49, 50 e 53, neste contexto eles foram consideradas nulas. Para medir o grau de semelhança entre os padrões de uso códons de duas espécies "a" e "b", introduziu-se o coeficiente de produto de ponto de similaridade *S*, que é dado por:

$$S_{a,b} = \frac{\sum_i C_{i,a} C_{i,b}}{\sqrt{(\sum_i C_{i,a}^2)(\sum_i C_{i,b}^2)}}$$

Esta teoria difere das maiorias teorias estudadas anteriormente com o uso de códons, devido a ela não se basear no processo de *Codon Usage Bias* (CUB), o método acima é o mais comum e utilizado para os estudos no uso de códons, pois ela não varia entre 0 e 1. O *s* posposto varia de 0 a 1, por que o *coseno* do ângulo é medido por um par de vetores de frequências de códons, a comparação dos vetores ($[C_{1,a,}C_{2,a,} \dots , C_{64,a}]$ *e* $[C_{1,b,}C_{2,b,} \dots , C_{64,b}]$), são realizados em um hiperespaço euclidiana com 64 dimensões, sendo que os dois vetores, não podem possuir valores negativos em sua composição.

3.8. Modelo de Tradução

Foram utilizadas as frequências de códons relativas para cada sinônimo de pares códons finais de pirimidina para estabelecer um equilíbrio entre o número de tradução de gene em espécies de RNA transportador e a frequência de códons por pares sinônimos. Denotam-se por g_{i} e g_{j} o número de genes de RNA transportador que descodifica o códon *i* como RNA transportador fraco e o *j* como RNA transportador forte, respectivamente, em um par de compartilhamento de RNA transportador mostrado na coluna *g* da Tabela 4. Seja $G = g_i + g_j$ ao qual denota-se o número total de genes de RNA transportador por par de códons. Denota-se, também, N_{i} e N_{j} como as contagens de códons genômicas para os códons *i* e *j*, respectivamente, que foram obtidos a partir dos dados do genoma humano. O par da frequência de códons relativa ao códon do RNA transportador fraco é dada por $f_i = N_i/(N_i + N_j)$. Assim, calcula-se o que chama-se de números de genes funcionais do RNA transportador como $g_i^* = f_i G$ e $g_j^* = G - g_i^*$, onde são mostrada na coluna g* da Tabela 4. Nota-se que os códons não formadores de pares de compartilhamento de RNA transportador, isto é, terminando com purinas A ou G, são repetidos na coluna g^* de acordo com a genômica RNA transportador do número de gene da coluna *g* da referida tabela. Os números de genes RNA transportador são utilizados para calcular relativamente a sua abundância, no qual é dada por: $t_{i,h} = g_i^* / \sum_j g_j^*$, que são necessários para o cálculo de *T-score*.

3.9. Calculo de pontuação terapêutica

Nos sistemas de produção a relação entre a demanda (d_i) e a oferta (s_i) para um recurso genérico *i* é uma medida de desigualdade dada por $u_i = 1 - d_i/s_i$ da cadeia de fornecimento para este recurso. O desequilíbrio é positivo quando $u_i > 0$, isto é, quando a oferta excede a demanda, e o desequilíbrio é negativo quando $u_i < 0$, isto é, quando a oferta não atende à demanda. Levando em conta que as moléculas de RNA mensageiro do vírus e do hospedeiro competem pelos mesmos recursos finitos para tradução no hospedeiro ($s_{i,h}$), o desempenho da cadeia de fornecimento será diferente desde as perspectivas do vírus ($u_{i,v} = 1 - d_{i,v}/s_{i,h}$) e do hospedeiro ($u_{i,h} = 1 - d_{i,h}/s_{i,h}$). Existem dois principais recursos de tradução, os ribossomos e as espécies de RNA transportador, pelos quais o vírus e o hospedeiro competem. Aqui nos focamos nas espécies de RNA transportador que decodificam cada códon *i*. As exigências são dadas pelas frequências de códons, que são $d_{i,v} = C_{i,v}$ e $d_{i,h} = C_{i,h}$, para o vírus e o hospedeiro, respectivamente, enquanto que o fornecimento é dado pela abundância relativa das espécies de tRNA cognatas, ou seja, $s_{i,h} = t_{i,h}$. A hipótese de viabilidade terapêutica se baseia na assumpção de que para algumas espécies de RNA transportador pode existir um desequilíbrio negativo para o vírus ($u_{i,v} < 0$), mas um desequilíbrio positivo para o hospedeiro ($u_{i,h} > 0$). Se existir um caso assim, ele deve sustentar que $\Delta u_i = u_{i,h} - u_{i,v} \gg 0$; assim, por meio do cálculo da diferença de desequilíbrio Δu_i para cada códon *i* em um determinado sistema de vírus-hospedeiro, e olhando para o valor máximo $max(\Delta u_i)$, poderia se identificar as espécies de RNA transportador que se inibidas terapeuticamente, iriam causar diminuição de $s_{i,h}$ (transformando-o em negativo), agravando o desequilíbrio para o vírus, mas sem afetar desequilíbrio no hospedeiro. A fim de dar prioridade a códons menos abundantes no genoma humano, minimizando assim o impacto sobre a célula hospedeira, Δu_i foi dividido pela frequência do códon *i* no genoma humano $C_{i,h}$, resultando, assim, na fórmula do índice T-score

$$T\text{-}score_i = \frac{\Delta u_i}{C_{i,h}}$$

Fazendo as substituições indicadas, o T-score vem dado de forma explícita como:

$$T\text{-}score_i = \frac{C_{i,v} - C_{i,h}}{t_{i,h} C_{i,h}}$$

De acordo com a definição de T-score, as espécies de tRNA que são os melhores alvos para inibição terapêutica (*Inhibition Therapy*) são aquelas que tem altos valores positivos de T-score, em comparação com a média de todas as espécies de RNA transportador.

4. Resultados e Discursões

Neste capítulo serão apresentados os resultados referentes à coleta das sequencias dos bancos de dados públicos, principal alvo desta pesquisa. Na sessão 4.1 será apresentada uma análise das sequências coletadas e armazenadas no BDE. Na sessão 4.2 será apresentada a ferramenta WEB com os gráficos demostrando o estudo da correlação entre frequência de códons em genomas virais e a abundância de espécies de RNA transportador cognato na célula hospedeira humana.

4.1. Análise dos dados coletados

O banco de dados possui atualmente 548.640 registros de sequências dos vírus HIV e HTLV coletadas do *GenBank*, com um volume total de aproximadamente 1 GB de armazenamento. Das sequências coletadas 99,05% correspondem ao vírus HIV e 0,95% correspondem ao vírus HTLV. Conforme o tipo sorológico dos vírus HIV e HTLV as sequencias foram estratificadas da seguinte forma HIV-1 (98,11%), HIV-2 (0,94%), HIV-3 (0,0007%), HTLV-1 (0,81%), HTLV-2 (0,13%), HTLV-3 (0,001%), HTLV-4 (0,0004%) (Figura 17). Considerando o comprimento da região gênica das sequências, 510.565 (93%) são parciais e 38.075 (7%) são completas.

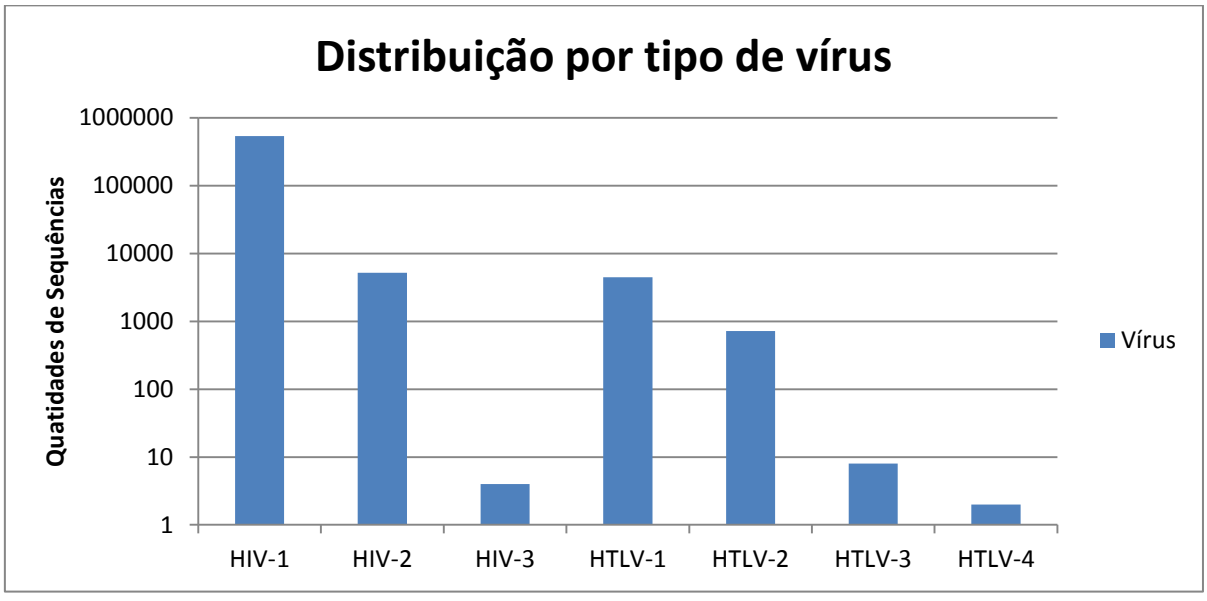

Figura 17 - Estratificação por classificação de vírus

Das 548.640 registros de sequências, 548.208 sequências tiveram informações das regiões genômicas (genes) e estão distribuídas da seguinte forma: genoma completo

(0,32%), *env* (32,37%), *gag* (12,53%), *gag-pol* (0,2%), *LTR* (0,01%), *nef* (4,99%), *pol* (42,64%), *pro* (0,12%), *px* (0,11%), *rev* (1%), *rex* (1,91%), *tat* (3,14%), *tax* (0,16%), *vif* (0,69%), *vpr* (0,71%), *vpu* (0,86%), *vpx* (0,004%) (Figura 18).

Com relação à distribuição por região genômica para o vírus HIV obtivemos os seguintes dados: genoma completo (0,32%), *env* (32,17%) *gag* (12,64%), *gag-pol* (0,2%), LTR (0,001%), *nef* (5.04%), *pol* (43,02%), pro (0,12%), *rev* (1,01%), *tat* (3,18%), *vif* (0,7%), *vpr* (0,72%), *vpu* (0,87%) e *vpx* (0,004%) (Figura 19), e para o vírus HTLV obtivemos os seguintes dados: genoma completo (0,62%), *env* (56,96%), *gag* (2,5%), *pol* (5,42%), *pro* (0,22%), px (12%), *rex* (3,79%), *tax* (17,26%) (Figura 20).

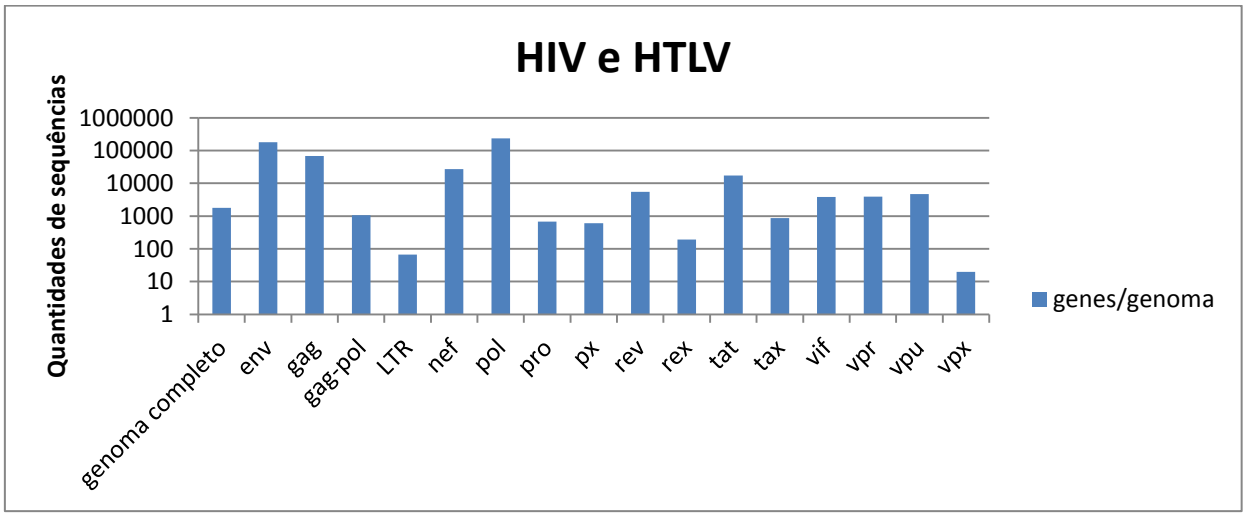

Figura 18 – Distribuição por regiões genômicas dos HIV e HTLV.

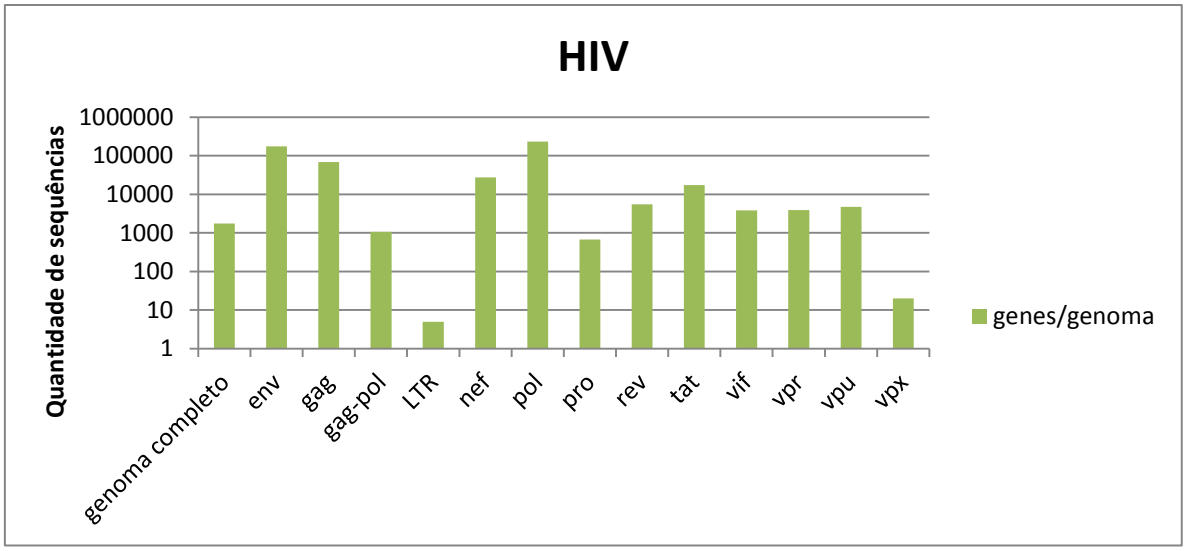

Figura 19 - Distribuição por regiões genômicas dos HIV.

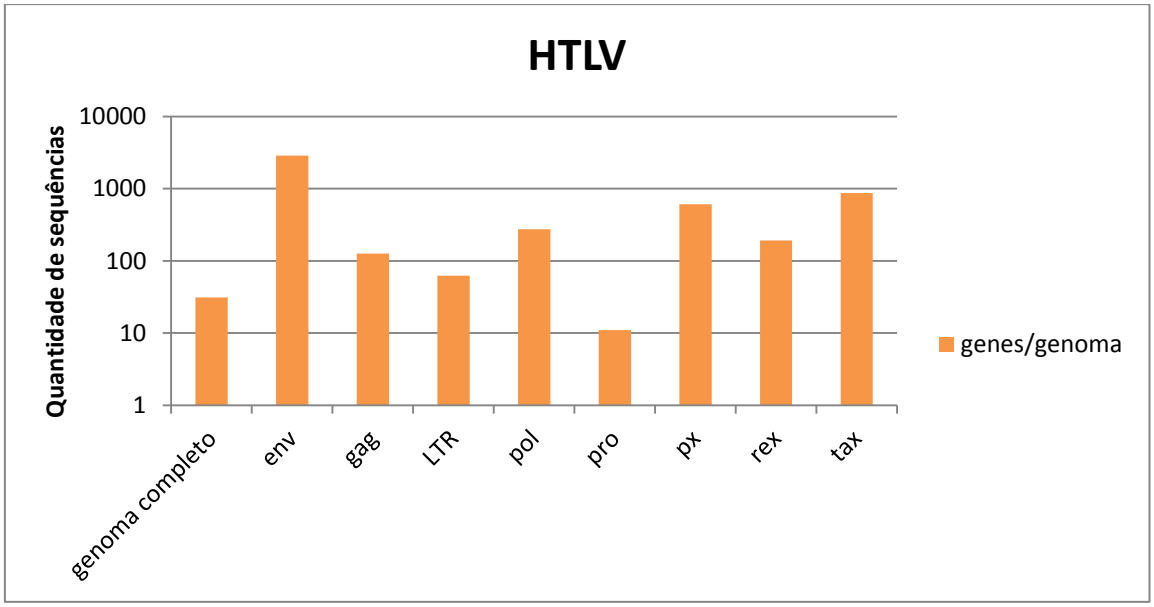

Figura 20 - Distribuição por regiões genômicas dos HTLV.

Em relação à distribuição geográfica, das 548.640 sequências coletadas, 548.453 (99,97%) possuem informações, distribuídas em 166 países diferentes. O Estados Unidos possuem mais de 1/4 (25,76%) dos isolados sequenciados e submetidos ao *GenBank* no mundo, seguido de Índia, Brasil, Quênia e África do Sul, acima de 20.000 sequências submetidas no *GenBank* (Figura 21).

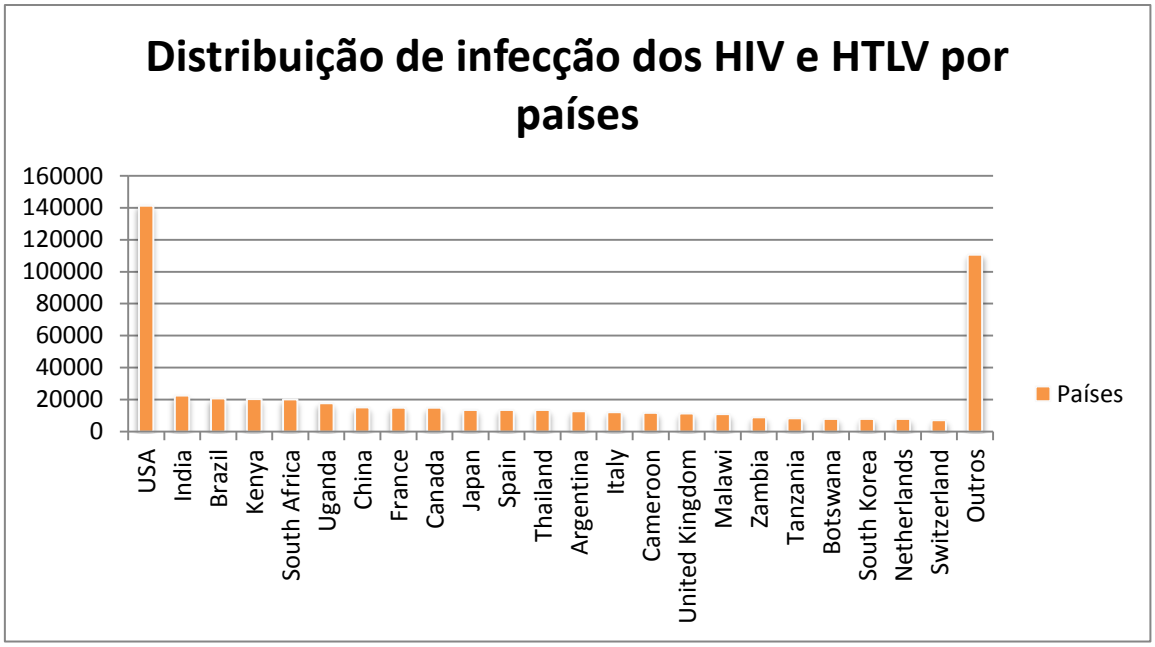

Figura 21 - Distribuição de sequências no *GenBank* dos HIV e HTLV por países.

4.2. WEB T-score: Therapeutic Score of tRNA Species

A ferramenta desenvolvida, além das sequências de isolados virais, disponibiliza informações mineradas do *Codon Usage Database*, contendo a frequência genômica de códons nos hospedeiros e do *Genomic tRNA Database*, contendo o número de cópias dos genes que codificam tRNA nos hospedeiros. Estas informações são armazenadas no BDE, no qual o usuário tem a possibilidade de escolher critérios de busca (Figura 22) de acordo com o objetivo da sua pesquisa. Após a escolha, o *software* gera uma resposta que é mostrada em outra interface, no qual aparecerão as frequências de códons do par de vírus-hospedeiro com frequências de espécies de RNA transportador, referentes à busca do usuário (Figuras Figura 23Figura 24Figura 25). A ferramenta WEB T-score: *Therapeutic Score of tRNA Species* está hospedada nos servidores da UNEB com o endereço eletrônico para acesso: http//www.t-score.uneb.br.

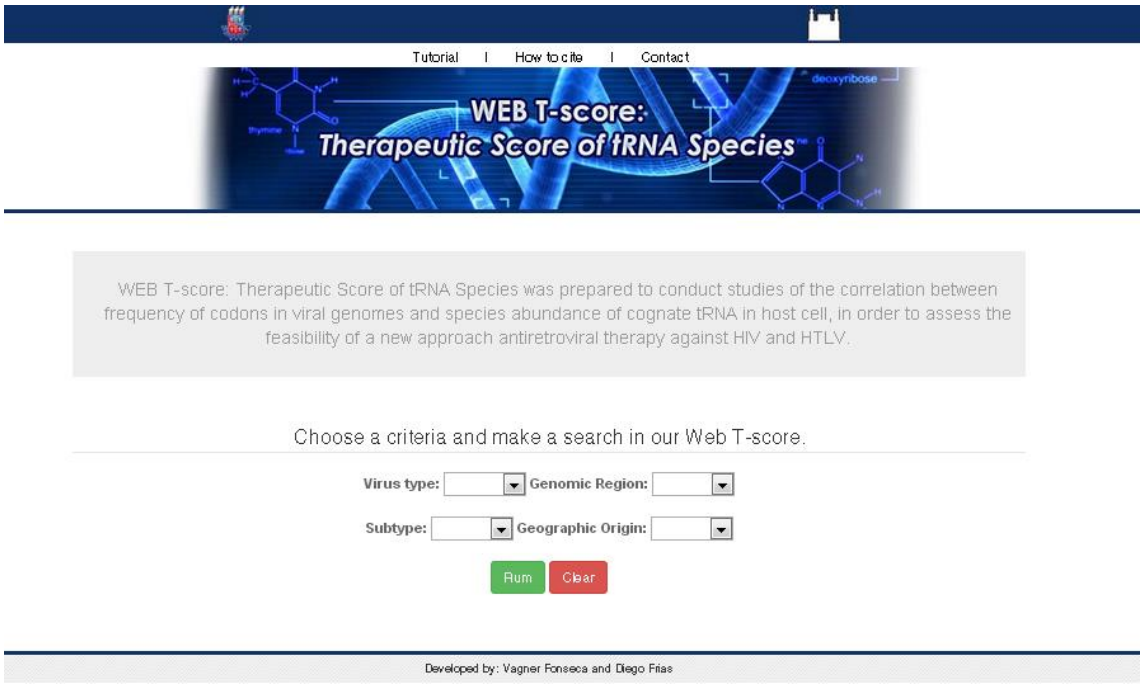

Figura 22 - Ferramenta WEB T-score: *Therapeutic Score of tRNA Species*.

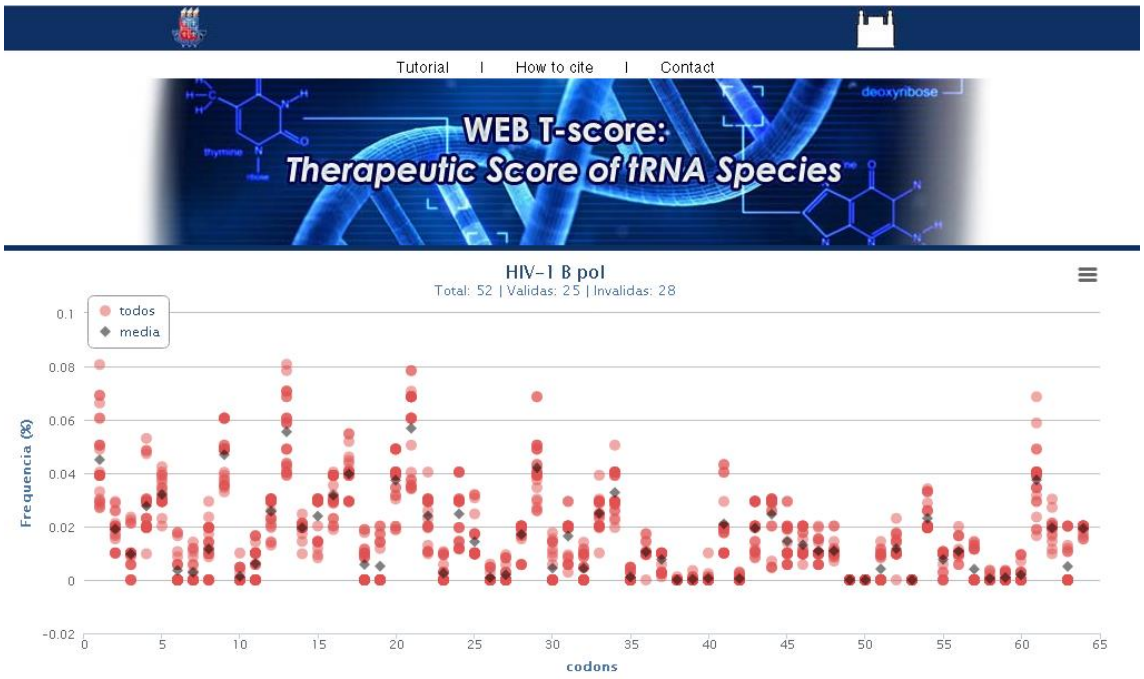

Figura 23 - Resultado da frequência de códons no gene pol do vírus HIV-1 subtipo B em qualquer região do mundo.

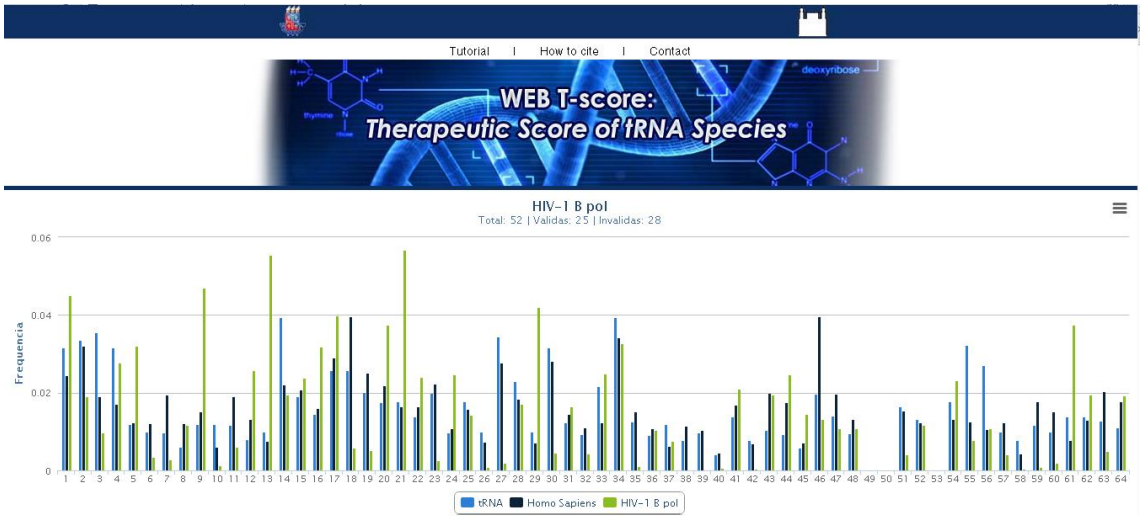

Figura 24 - Resultado da frequência média de códons no gene pol do vírus HIV-1 subtipo B em qualquer região do mundo sobreposta à frequência de códons no hospedeiro humano e à frequência de genes no genoma do hospedeiro que codificam os RNA transportador cognatos a cada códon.

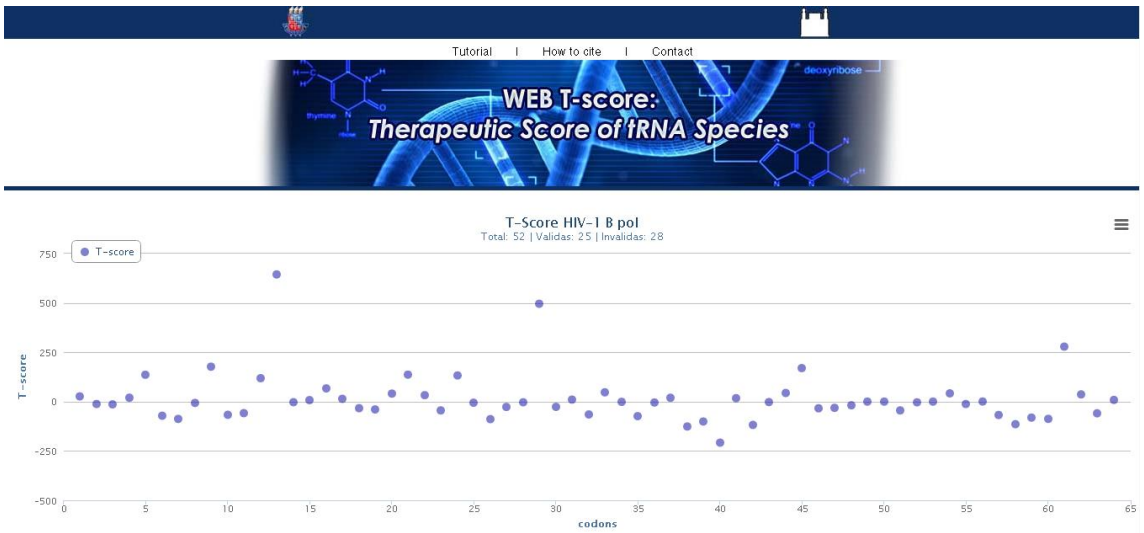

Figura 25 - Resultado do T-score para o gene pol do vírus HIV-1 subtipo B selecionado.

Todos os dados coletados sejam as sequências de nucleotídeo do HIV e HTLV e suas anotações tiveram como origem o *GenBank*, onde são submetidas os estudos realizados pelos pesquisadores. Estes pesquisadores desenvolvem suas pesquisas em diversos centros de pesquisas no mundo inteiro, os quais realizam sequenciamento e anotações sobre as sequências. A prática demonstra que ocorrem divergências nas anotações e até mesmo erros no sequenciamento.

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

Este capítulo apresentou detalhes sobre os dados minerados e a implementação da ferramenta WEB como experimento prático, com o intuito de validar o estudo proposto por esta pesquisa. Assim, pode-se concluir que a ferramenta foi validada com êxito, atingindo o objetivo desta pesquisa.

5. Considerações Finais

A pesquisa teve como objetivo o desenvolvimento do protótipo de uma ferramenta de bioinformática via WEB para o estudo da correlação entre frequência de códons em genomas virais e a abundância de espécies de RNA transportador cognato em célula hospedeira, com intuito de avaliar a viabilidade de uma nova abordagem terapêutica antirretroviral, em particular contra os vírus HIV e HTLV.

A principal tarefa deste projeto de pesquisa, consistiu em construir um banco de dados especifico (BDE) com sequências de genes virais e outros dados relativos ao hospedeiro humano e desenvolver uma ferramenta de bioinformática para subsídio ao estudo de uma nova abordagem de terapia antiviral baseada em interferência seletiva de espécies de RNA transportador. Para povoar o BDE foi desenvolvido um robô para atualização automática, mas controlada, a partir de diversos Bancos de Dados Públicos (BDP). Para acessar aos resultados dos estudos foi desenvolvido um portal WEB para disponibilizar gráficos com as correlações entre as frequências de códons nos genomas dos patógenos e dos hospedeiros e as frequências das espécies cognatas de RNA transportador no hospedeiro, calculadas segundo modelo parametrizado de compartilhamento de tRNA.

Para atingir o objetivo supracitado foi necessário cumprir algumas etapas: A primeira etapa foi compreender a estrutura e mecanismos de acesso automatizado aos BDP para *download* de sequência codificantes (CDS), das frequências genômicas de genes de espécies de tRNA (gtRNA_DB) e das frequências de códons no genoma dos hospedeiros (CUT).

A segunda etapa foi realizar um estudo das técnicas e estatísticas de reconhecimento de padrões para detecção e extração de sequências codificantes no quadro de leitura de sequências de DNA genômico ou RNA. Este método foi utilizado para excluir sequências com erros de anotação ou de sequenciamento, consideradas não válidas.

A terceira e última etapa de familiarização com o problema em estudo consistiu em compreender o papel e funcionamento do RNA transportador no processo de tradução de mRNA para a síntese de proteínas, o pareamento códon-anticódon no ribossomo e o pareamento não clássico, assim como entender a hipótese de compartilhamento de tRNA por códons sinônimos desenvolvida pela equipe de pesquisadores.

Para a elaboração do algoritmo capaz de minerar e coletar as informações armazenadas nos BDP, foram estudados três *framewoks* de bioinformática, sendo eles BioPerl, BioJava e BioPHP, optando-se por utilizar o BioPHP, por ser aquele que tínhamos maior experiência em programação.

A principal contribuição dessa pesquisa foi desenvolver uma nova ferramenta de bioinformática acessível via WEB que automatiza o *download* de sequências de genes virais e de dados dos organismos hospedeiros permitindo estender estudos que foram realizados anteriormente com um número limitado de casos, relacionados com a viabilidade de uma nova abordagem terapêutica baseada na inibição da replicação viral nas células infectadas, mediante a inibição seletiva de espécies de tRNA com alto índice terapêutico, T-score.

Os resultados do projeto abrem um leque de novas possibilidades para estudos similares com outras espécies de patógenos tais como: Malária, *E. coli*, Influenza A, Polivírus, Herpes e Hepatites de alta incidência e mortalidade na população mundial.

6. Referências

ABU-HANNA A. **Review of "Machine Learning" In:** Artificial Intelligence in Medicine. Elsevier. 1999;16:201‑4.

ADDRIANS, P. e ZANTINGE, D. **Data Mining.** Inglaterra: Addison-Wesley, 1996.

ALBERT, Bruce. *et al*. **Biologia Molecular da Célula.** 5ª edição, Porto Alegre: Artmed, 2010.

BALDI, P; BRUNAK, S. Bioinformatics: **The machine learning approach 2nd ed., chapter 7.** MIT Press, Cambridge, MA. 2001.

BARRE-SINOUSSI F, *et al*. **Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS).** Science 1983; 220:868-71.

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 25 de junho de 2013.

BORODOVSKY, M; MCININCH, J. **GenMark:** parallel gene recognition for both DNA strands. Comput Chem. 1993;17:123‑33.

BRACHNAD, R.J. e ANAND, T. **The process of knowledge discovery in databases.** In: FAYYAD, U.M. *et al.* Advances in Knowledge Discovery in Data Mining. Menlo Park: AAAI Press, 1996.

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.

CARELS Nicolas; VIDAL Ramon e FRÍAS, Diego. **Universal Features for the Classification of Coding and Non-Coding DNA Sequences.** Bioinformatics and Biology Insights, 2009:3 pag. 37-49. Disponível em: <http://www.la-press.com/universal-features-for-the-classification-of-coding-and-non-coding-dna-article-a1479>.

CARELS, Nicolas e FRÍAS, Diego. **Classifying Coding DNA with Nucleotide Statistics.** Bioinformatics and Biology Insights, 2009:3 pag. 141-154. Disponível em: <http://www.la-press.com/classifying-coding-dna-with-nucleotide-statistics-article-a1718>.

CARELS, Nicolas e FRIAS, Diego. **A Statistical Method without Training Step for the Classification of Coding Frame in Transcriptome Sequences.** Bioinformatics and Biology Insights, 2013:7, pag. 35-54. Disponível em: <http://www.la-press.com/classifying-coding-dna-with-nucleotide-statistics-article-a1718>.

CHINEN, J. e SHEARER, W. T. **Molecular virology and immunology of HIV infection.** Journal of Allergy and Clinical Immunology 2002; 110:189–198.

CUELLAR, M. L. **HIV infection-associated inflammatory musculoskeletal disorders.** Rheumatic Disease Clinics of North American 1998; 24: 403-421.

DILLY, R. **Data Mining: an introduction.** Belfast: Parallel Computer Centre, Queens University, 1999.

DINIZ, C.A. e LOUZADA-NETO, F. **Data Mining: uma introdução.** São Carlos: Associação Brasileira de Estatística, 2000.

ELEOPULOS, E.P.; *et al*. **A critique of the Montagnier evidence for the HIV/AIDS hypothesis.** Medical Hypoteses 2004; 63: 597-601.

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

FAYYAD, U.M. *et al.* **The KDD Process for Extracting Useful Knowledge from Volumes of Data.** In: ___. Advances in Knowledge Discovery in Data Mining. Menlo Park: AAAI Press, 1996a.

FAYYAD, U.M. *et al.* **Advances in Knowledge Discovery and Data Mining.** California: AAAI Press, 1996b.

FEUER G, GREEN PL.**Comparative biology of human T-cell lymphotropic vírus type 1 (HTLV-1) and HTLV-2.** Oncogene. 2005; 24: 5996-6004.

FRIAS, Diego; JUNIOR, Luis Carlos Alcantâra; GALVÃO, Bernardo. **Codon usage study of human HIV and HTLV supports promissory tRNA neutralization therapy viruses, from a translational logistic perspective.** Conferência AIDS, 2011.

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

GONSALEZ D, OLIVEIRA JS, HAAPALAINEN E, KERBAUY J. **Hairy cell leukemia: a histo-cytochemical and ultra-structural study.** São Paulo Med J 1998;116(2):1681-5.

GOUY M, GAUTIER C, ATTIMONELLI M, LANAVE C, DI PAOLA G. ACNUC - **a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage.** Computer applications in the biosciences: CABIOS 1:3 p. 167-72, 1985.

GOUY M, DELMOTTE S. **Remote access to ACNUC nucleotide and protein sequence databases at PBIL.** Biochimie 90:4, pg 555-62, 2008.

GUJARATI, D.N. **Econometria Básica.** Trad. Ernesto Yoshita. São Paulo: Makron Books, 2000.

HAND, D.J. **Data Mining: statistics and more?** The American Statistician, England, 52 (2): 112-118, 1998.

IKEHARA K, OMORI Y, ARAI R, et al. **A Novel Theory on the Origin of the Genetic Code: A GNC-SNS Hypothesis.** J Mol Evol. 2002;54:530–8.

JANEWAY, C.A; TRAVERS, P. **Immuno Biology – The immune system in health and disease.** CB 3.ed. 1997; p. 1:2 - 7:27.

KLEIN Florian, et al. **Broad neutralization by a combination of antibodies recognizing the CD4 binding site and a new conformational epitope on the HIV-1 envelope protein.** The Journal of Experimental Medicine (JEM), publicado em 23 de julho de 2012.

KROGH, A; MIAN, IS; HAUSSLER, D. **A hidden Markov model that finds genes in E. coli DNA.** Nucleic Acids Res. 1994;22:4768‑78.

LEHVASLAIHO H. **Introduction to Perl and BioPerl**, Institut Pasteur Tunis, 2007. Disponível em <http://www.pasteur.fr/~tekaia/BCGA/TALKS/Heikki_perl-bioperl.pdf>.

LEMOS, M. W**orkflow para Bioinformática.** PhD thesis, Departamento de Informática da PUC-Rio, 2004.

LEVINE, D.M. *et al.* **Estatística: teoria e aplicações.** Trad. Teresa C.P. de Souza. Rio de Janeiro: LTC Editora, 2000.

LI, Manqing, et al. **Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11**. Macmillan Publishers Limited, publicado em 23 de setembro de 2012.

LIPMAN, David J.; PEARSON, William R. **Rapid and sensitive protein similarity searches.** Science, 1985, 227 (4693) pp. 1435–41.

MANNILA, H. **Data mining: machine learning, statistics and databases.** International Conference on Statistics and Scientific Database Management, Estocolmo, 8, 1996.

MARTINS, G.A. **Estatística Geral e Aplicada.** São Paulo: Atlas, 2001.

MATTAR, F.N. **Pesquisa de Marketing.** São Paulo: Atlas, 1998.

MORETTIN, P.A. & TOLOI, C.M. **Séries Temporais.** 2.ª ed. São Paulo: Atual, 1987.

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.

PADOVANI, C.R. **Estatística na Metodologia da Investigação Científica.** Botucatu: UNESP, 1995.

PEREIRA, J.C.R. **Análise de Dados Qualitativos.** São Paulo: Edusp/Fapesp, 1999.

PEREIRA, Patrícia Reis. **Subtipos do HIV-1 e associação com características demográfico-epidemiológicas em pacientes atendidos em Hospital de referência em Porto Alegre, Brasil.** Universidade Federal Do Rio Grande Do Sul, 2010.

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.

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

QUEIROZ, M.S. **Representações Sociais.** Uma perspectiva multidisciplinar em pesquisa qualitativa. In: BRATA, R. B.; BRICEÑO-LEON, RJ. (Orgs.). Doenças Endêmicas: abordagens sociais, culturais e comportamentais. Rio de Janeiro, RJ: Editora Fiocruz, 2000 p. 27-46.

RAAPHORST, F.M. *et al.* **TCRBV CDR3 Diversity of CD4+ and CD8+ T-Lymphocytes in HIV-Infected Individuals.** Human Immunology 2002; 63: 51-60.

REPETTO, M. *et al.* **Oxidative stress in blood of HIV infected patients.** Clinica Chimica Acta 1996, 255: 107-117.

SADE, A.S. e SOUZA, J.M. **Prospecção de Conhecimento em Bases de Dados Ambientais.** Rio de Janeiro: UFRJ, 1996.

SHERMAN, M.P.; GREENE, W.C. **Slipping through the door: HIV entry into the nucleus.** Microbes and Infection 2002; 4: 67 – 63.

SLEASMAN, J. W. e GOODNOW, M. M. **HIV-1 Infection.** Journal of Allergy and Clinical Immunology 2003; 111: 582-592.

SPENCER, Paige S. SILLER, Efraín, ANDERSON John F. e BARRAL, José M. **Silent substitutions predictably alter translation elongation rates and protein folding efficiencies.** National Intitutes of Health J Mol Biol. 2012 September 21; 422(3): 328–335.

UCHIYAMA, T. YODOI J. SAGAWA, K. TAKATSUKI, K. UCHINO H. **Adult T-cell leukemia: clinical and hematologic features of 16 cases.** Blood. 1977; 50:481-492.

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.

WILHELM T, NIKOLAJEWA S. **A new classification scheme of the genetic code.** Journal of Molecular Evolution. 2004; 59(5):598–605.

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.

WONG, Emily HM. SMITH, David K. RABADAN, Raul. PEIRIS, Malik. POON, Leo LM. **Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus.** BMC Evolutionary Biology 2010, 10:253.

YADAVA, Anjali. OCKENHOUSE Christian F. **Effect of Codon Optimization on Expression Levels of a Functionally Folded Malaria Vaccine Candidate in Prokaryotic and Eukaryotic Expression Systems.** Infection and Immunity, Sept. 2003, p. 4961–4969.