MINISTÉRIO DA SAÚDE – FUNDAÇÃO OSWALDO CRUZ

ESCOLA NACIONAL DE SAÚDE PÚBLICA SÉRGIO AROUCA

EPIDEMIOLOGIA MOLECULAR DO VÍRUS DA IMUNODEFICIÊNCIA HUMANA DO TIPO I: MÉTODOS DE INFERÊNCIA FILOGENÉTICA

Por

Jorge Francisco da Cunha Pinto

Dissertação apresentada ao Mestrado em Saúde Pública – Área de concentração Epidemiologia Geral

Orientação: Prof. Dr. Cláudio José Struchiner

Rio de Janeiro, julho de 2004

Sumário

O crescimento das bases de dados moleculares referentes ao vírus da imunodeficiência humana do tipo I (HIV-1) aumentou progressivamente desde 1991. Pesquisadores do mundo inteiro têm se dedicado ao seqüenciamento de diferentes regiões do genoma do HIV visando elucidar o processo evolutivo viral. Supõe-se que este processo evolutivo esteja na base da pesquisa que determinará a produção de vacinas eficazes além de novas drogas para o combate da Aids. Neste trabalho, procuramos introduzir alguns aspectos da epidemiologia molecular do HIV-1 enfatizando a distribuição global dos seus subtipos e os métodos de inferência filogenética utilizados no estudo de sua evolução. Apresentamos, como aplicação dos métodos de inferência filogenética, um artigo intitulado "Epidemiologia Molecular do Sub-subtipo F1 do HIV-1", onde discutimos a epidemiologia molecular do sub-subtipo F1 buscando comparar as epidemias deste sub-subtipo no Brasil e na Romênia.

Palavras Chaves: inferência filogenética, epidemiologia molecular, subtipos de HIV-1.

Summary

The growth of the human immunodeficiency virus (HIV) sequence databases has been staggering since 1991. Researchers all over the world are sequencing different regions of the HIV genome to infer processes of HIV evolution. It has been supposed that this process is on the top of the researches that will contribute to vaccine development and the discovery of new drugs against Aids. In this work, we introduce some aspects of the molecular epidemiology of HIV-1 linking to its world diversity and global distribution. We also set forth the methods of inferring phylogenies. Finally, we present a paper named “Epidemiologia Molecular do Sub-subtipo F1 do HIV-1” where we discuss the molecular epidemic of the sub-subtype F1 of HIV-1 linking the Brazilian and Romanian epidemic of this sub-subtype.

Key words: inferring phylogenies, molecular epidemiology, HIV-1 subtypes.

Índice

Epidemiologia Molecular do Vírus da Imunodeficiência Humana do Tipo I: Métodos de Inferência Filogenética

I. Introdução

A identificação, em 1983, do vírus da imunodeficiência humana (HIV) como agente etiológico da Aids possibilitou uma enorme evolução no campo da epidemiologia molecular. Três genes estruturais e seis regulatórios codificam 15 proteínas virais cruciais ao entendimento do seu ciclo de replicação e suas relações com a evolução da Aids. A epidemiologia molecular pôde esclarecer a heterogeneidade do HIV incluindo as formas recombinantes e suas origens zoonóticas a partir de primatas não humanos (Hirsch *et al.*, 1989; McCutchan *et al.*, 1996; Gao *et al.*, 1999; Caride *et al.*, 2000; Corbet *et al.*, 2000; Carr *et al.*, 2001; Blackard *et al.*, 2002), além de sua extensa variabilidade genética em um mesmo indivíduo, particularmente na região hipervariável do seu envelope, demonstrando taxas evolutivas bastante aceleradas (Hu *et al.*, 1996; Zhu *et al.*, 1998; Shankarappa *et al.*, 1999; Holmes, 2003; Holmes, 2004).

O HIV replica continuamente mesmo nos períodos em que o indivíduo infectado encontra-se assintomático, o que impulsiona o surgimento de variantes virais resistentes a medicamentos capazes de escapar ao sistema imune levando o paciente ao óbito (Shankarappa *et al.*, 1999; Persaud *et al.*, 2003). Seu alto grau de mutação genética se deve principalmente a enzima transcriptase reversa que comete aproximadamente 0,2 erro por genoma durante cada ciclo de replicação facilitando a ocorrência posterior de erros na transcrição do DNA pela polimerase Pol II RNA. Uma nova geração surge a cada, aproximadamente, 2,5 dias e cerca de 10^{10} a 10^{12} novas partículas virais ocorrem a cada dia (Perelson *et al.*, 1996).

II. A organização genômica do HIV-1

Como todos os outros membros da família Retroviridae, os agentes etiológicos da Aids contêm um capsídeo viral composto pela proteína capsídica p24, a proteína nucleocapsídica p7/p9, o genoma diplóide de RNA e três enzimas virais: a transcriptase reversa, a protease e a integrase. O capsídeo viral circunda-se pela proteína da matriz (p17), posicionada na face interna do envelope. Este é composto de uma dupla camada lipídica, oriunda da membrana citoplasmática da célula hospedeira, ornamentada de espículas essenciais à invasão celular. Estas espículas são complexos protéicos compostos pela glicoproteína de superfície gp120 e a glicoproteína gp41, de transposição da membrana (Barre-Sinoussi, 1996; Fields *et al.*, 2001; Poignard *et al.*, 2001). O genoma dos retrovírus são diméricos consistindo de duas subunidades que possuem seqüências idênticas (ou quase idênticas) mantendo-se juntas por pareamento de bases. A importância disto é a possibilidade de um alto grau de recombinação viral durante a transcrição pela transcriptase reversa (Coffin, 1979).

Os retrovírus possuem três grupos de genes estruturais (figura 1): gag (antígeno grupo-específico), pol (polimerase) e env (envelope). O gene gag codifica a proteína precursora assemblina (p55) que será clivada pela protease para dar origem às proteínas da matriz (p17), do capsídeo (p24) e do nucleocapsídeo (p7/p9). O gene pol contém as instruções para a produção das enzimas (1) transcriptase reversa, (2) protease, responsável pela clivagem proteolítica dos produtos dos genes gag e pol sem a qual as partículas virais produzidas são desprovidas de infectividade (Kaplan et al., 1993; Kohl et al., 1988) e (3) a integrase, fundamental à integração do provírus ao genoma da célula hospedeira (Barre-Sinoussi, 1996; Fields *et al.*, 2001). O gene env traz as informações necessárias à produção da proteína precursora gp160, que será processada por uma protease celular para dar origem às duas proteínas do envelope: gp120 e gp4l. Estas proteínas vão desempenhar papel de maior relevância na patogenia da infecção.

As glicoproteínas do envelope viral desempenham a função de mediadoras da entrada do HIV na célula e são alvos do ataque humoral do hospedeiro. A gp120 é altamente glicosilada e compõe-se de cinco regiões constantes (C1 a C5) interpostas com cinco regiões variáveis (V1 a V5). A fusão viral à célula-alvo ocorre através de uma interação

seqüencial entre a gp120 e os receptores celulares do HIV: a molécula CD4 e os membros da família de receptores de quimiocina. A ligação da gp120 ao receptor CD4 não é suficiente para a invasão celular, mas gera uma alteração na estrutura da primeira que permite sua interação com um membro da família de receptores de quimiocina. As cepas com tropismo pelos macrófagos (cepas R5) utilizam o receptor CCR5, as com tropismo pelas células T (cepas X4) utilizam o receptor CXCR4 e aquelas com duplo tropismo (R5X4) podem utilizar um ou outro. Após a ligação da gp120 ao co-receptor, uma nova alteração na estrutura tridimensional da glicoproteína do envelope permite a fusão da gp41 à membrana celular e a penetração viral (Poignard *et al.*, 2001).

A importância dos co-receptores celulares do HIV pôde ser demonstrada com a descoberta de que o genótipo homozigoto para a deleção Δ32 no gene CCR5 confere resistência, mesmo que não absoluta, à infecção pelo HIV (Liu *et al.*, 1996) e de que o genótipo heterozigoto está associado com uma progressão clínica mais lenta (Dean *et al.*, 1996).

Os agentes etiológicos da Aids contêm, ainda, os genes acessórios tat, ver, vif, nef e vpr. O HIV-1 possui, além destes, o gene vpu e o HIV-2 possui o vpx. Embora não seja difícil compreender as funções dos genes virais estruturais e enzimáticos, o papel desempenhado pelos genes acessórios no ciclo de vida dos lentivírus é bem mais complexo (Broder *et al.*, 1999). Alguns estudos associam pacientes não progressores de com a deficiência do gene nef e o bloqueio da atividade deste gene no ciclo do HIV pode ser um alvo importante nos estudos relacionados às vacinas anti-HIV (Guimaraes *et al.*, 2002).

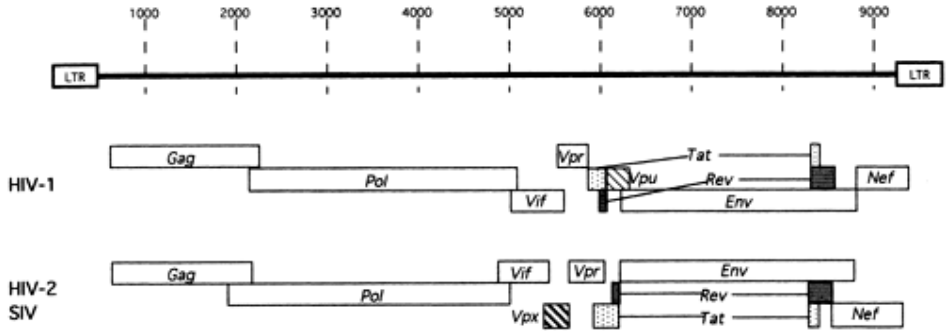

Figura 1 – Modelo simplificado da organização estrutural do HIV-1, HIV2 e SIV (Crandall, 1999).

III. A evolução molecular do HIV-1

Os estudos filogenéticos demonstram a relação histórica entre os genes através do tempo ao mesmo tempo em que tentam elucidar processos que afetem a dinâmica das seqüências populacionais nelas baseadas (Crandall & Templeton, 1999). A rápida evolução do HIV nos permite estudar as alterações filogenéticas ao longo de anos ou, mais simplesmente, de alguns meses (Hillis, 1999). Sua variabilidade genética pode alterar de forma importante a patogenia da doença, imunogenicidade e o potencial de transmissibilidade.

Sabendo-se que as taxas evolucionárias dos vírus RNA são consideravelmente maiores do que aquelas de seus hospedeiros, muitas são as explicações para sua variabilidade genética: a alta taxa de erro da transcriptase reversa em conjunto com a alta taxa de geração de populações virais (Ho *et al.*, 1995); a enorme diversidade de proteínas virais; a existência de fortes pressões ambientais seletivas, como a do sistema imune e a presença de agentes anti-retrovirais; o tropismo viral por células dos vários tecido e o aparecimento de recombinação entre os diversos genótipos (McCutchan *et al.*, 1996).

A constante emergência de variantes virais cria um repertório de seqüências genéticas heterogêneas, chamado de quasispecies, em indivíduos cronicamente infectados pelo HIV. O estudo longitudinal conduzido por Shankarappa e cols. (Shankarappa *et al.*, 1999) analisou a evolução das seqüências da região C2-V5 do gene env de nove pacientes por um período de tempo que variou de seis a 12 anos, desde a data da soroconversão. Propuseram, então, a existência de três fases distintas durante o estágio clínico assintomático: (1) uma fase inicial de duração variável na qual há aumento linear (aproximadamente 1% ao ano) tanto da divergência quanto da diversidade; (2) uma fase intermediária de duração aproximada de 1,8 anos caracterizada por um aumento contínuo da divergência, porém com estabilização, ou mesmo declínio, da diversidade; (3) uma fase tardia na qual se nota um alentecimento ou estabilização da divergência e uma manutenção da tendência de estabilização ou declínio da diversidade. Um desdobramento importante do entendimento dos padrões de evolução das seqüências virais, explicam os autores, poderá ser a habilidade em se prever com antecedência o início clínico da Aids através da freqüência com que o fenótipo X4 é detectado dentre o

repertório viral e dessa forma melhor definir o momento mais apropriado para o início da terapia anti-HIV.

Como resultado de sua grande diversidade genética, as seqüências conhecidas de HIV-1 podem ser classificadas do ponto de vista filogenético em grupos e subtipos. Tal classificação é reflexo do estudo das amostras que puderam ser coletadas e caracterizadas não devendo ser considerada como necessariamente representativas do espectro completo ou da real prevalência de cada subtipo (Hu *et al.*, 1996). Artefatos relacionados à amostragem, e mesmo eventuais dificuldades de isolamento de subtipos ou variantes específicas, fazem com que a maioria dessas investigações deva ser entendida como estudo de casos.

O HIV possui duas categorias principais baseadas na sua distribuição geográfica e na fonte animal de infecção humana: chimpanzé (Pan troglodytes) para HIV-1 (Gao *et al.*, 1999) e sooty mangabey (Cercocebus atys) para HIV-2 (Hirsch *et al.*, 1989). Para o HIV-1 foram descritos três grupos distantemente relacionados: o grupo M (principal) (Robertson, Anderson et al., 2000), o grupo N (para os não M e não O) (Simon *et al.*, 1998) e o grupo O (externo) (Gurtler *et al.*, 1994).

O grupo M apresenta linhagens distintas que foram designadas de subtipos e sub-subtipos (figura 2) e formas recombinantes (Carr *et al.*, 1996). A variabilidade genética que distancia os diferentes genótipos do grupo M entre si é de aproximadamente 30% para as seqüências do gene env e de 14% para as do gene gag (Crandall, 1999).

Os subtipos B e D são os mais próximos do ponto de vista filogenético e provavelmente derivam de um ancestral comum. De fato, a mais remota seqüência de HIV conhecida, colhida em 1959 na atual República Democrática do Congo, foi posicionada próxima ao nó ancestral dos subtipos B e D (Zhu *et al.*, 1998). Estima-se que o subtipo C seja globalmente o mais prevalente e que seja responsável por aproximadamente 48% de todas as infecções, em especial na Índia e no sul e sudeste africanos (Alaeus, 2000).

Tornando a epidemiologia molecular do HIV-1 ainda mais complexa, sabe-se que uma substancial parcela da pandemia é composta por formas recombinantes intersubtipo. A emergência de formas recombinantes pode ser considerada uma propriedade fundamental dos retrovírus em razão da natureza diplóide do seu genoma de RNA e da

possibilidade da transcriptase reversa atuar ora nesta, ora naquela fita em células infectadas por mais de uma variante viral (Anderson *et al.*, 1996; Robertson *et al.*, 2000; Blackard *et al.*, 2002). Dessa forma, os espécimes com recombinação intersubtipo são mais prevalentes em áreas onde múltiplos subtipos circulam. Algumas dessas formas recombinantes ocorrem em infecções isoladas e não desempenham um papel mais significativo na pandemia. Outras, entretanto, disseminaram-se como no caso da forma CRF01_AE, esporadicamente encontrada em países da África central, porém responsável pela epidemia explosiva da Tailândia e outros países do Sudeste asiático. Este subtipo é um mosaico entre o segmento gag do subtipo A e o segmento env de um subtipo cuja forma "pura" ainda não foi identificada (McCutchan *et al.*, 1996; Alaeus, 2000).

As seqüências anteriormente classificadas como subtipo I constituem formas recombinantes complexas (cpx) que envolvem pelo menos quatro subtipos e são atualmente chamadas de CRF04_cpx. O subtipo CRF02_AG, um complexo mosaico formado por segmentos intercalados dos subtipos A e G, é a variante que predomina na África central e centro-ocidental, onde é incriminada como a responsável por cerca de 50% a 70% de todas as infecções (Alaeus, 2000).

Sabe-se atualmente que algumas formas recombinantes já circulavam no início da epidemia. O próprio subtipo CRF02_AG, por exemplo, foi recentemente demonstrado em amostras colhidas ao longo do segundo semestre de 1985 em Kinshasa (Yang *et al.*, 2001).

A habilidade em justapor segmentos de diferentes subtipos em um único genoma pode abrir o caminho para a geração de cepas com propriedades biológicas, epidemiológicas ou imunológicas distintas (Anderson *et al.*, 1996; Robertson *et al.*, 2000; Blackard *et al.*, 2002). Em locais onde mais de um subtipo é prevalente, a rápida disseminação de uma das variantes em sobreposição à outra levanta a hipótese de que possa haver diferenças na transmissibilidade dos diversos subtipos. Um exemplo seria a rápida disseminação heterossexual do subtipo CRF01_AE na Tailândia, em sobreposição ao subtipo B.

O evento da recombinação também pode ocorrer quando há infecção por variantes diferentes do mesmo subtipo (Diaz *et al.*, 1995) e em casos de dupla infecção por

variantes de grupos diferentes de HIV-1 (Peeters *et al.*, 1999; Takehisa *et al.*, 1999). A descrição de variantes mosaicas entre os grupos M e O, como a descrita por Peeters e colaboradores (1999), além de demonstrar a viabilidade da recombinação entre variantes com acentuado grau de diversidade molecular, evidencia a possibilidade de erros de diagnóstico sorológico. O próprio grupo N do HIV-1, como vimos, é uma forma recombinante que ora contém segmentos do grupo M, ora do SIVcpz (Corbet *et al.*, 2000).

IV. A distribuição global dos subtipos de HIV-1

A freqüência com que os diferentes subtipos vêm sendo isolados em diversas regiões parece ser resultado da chamada "migração viral" (Myers, 1994). Dessa forma, na Europa e nas Américas a grande prevalência de infecções por subtipo B provavelmente é o resultado do chamado "efeito fundador", a introdução ao acaso deste subtipo em particular e sua posterior disseminação.

Foley e colaboradores (2000), por exemplo, estudaram as seqüências do gene env recuperadas de soros crioconservados, coletados de três pacientes em 1978 e de um em 1979 na cidade de São Francisco. Todas essas quatro seqüências históricas foram identificadas como subtipo B, com a mediana da distância nudeotídica de apenas 2,8%, sugerindo um recente ancestral comum. Uma vez que a taxa de substituição de nudeotídeos do gene env foi estimada em 0,6 a 1,1% ao ano, os investigadores inferem que as seqüências analisadas teriam um ancestral comum em 1975-1976. Para efeitos de comparação, as medianas das distâncias nudeotídicas do gene env de seqüências de 1983-1986, 1987-1989 e 1990-1996, seriam, respectivamente, de 11,9%, 14,3% e 16,4% (Foley *et al.*, 2000). Dessa forma, a epidemiologia molecular do HIV-1 obedece a dois fatores: sua diversificação genética inata e o fator de migração viral (Myers, 1994).

Em muitos países europeus observa-se amplo predomínio do subtipo B. Na França, por exemplo, apesar da ocorrência de múltiplos subtipos, apenas uma minoria de casos é atribuída a subtipos não-B (Couturier *et al.*, 2000; Fleury *et al.*, 2003). No continente africano, onde se acredita ser o início da epidemia de HIV/Aids, já foram encontrados todos os subtipos de HIV-1 e HIV-2, com aparente predomínio em muitas localidades do subtipo A (Louwagie *et al.*, 1995; Janssens *et al.*, 1997; Vidal *et al.*, 2000). No Sudeste asiático predomina o subtipo CRF01_AE. Nos países sul-americanos de língua espanhola submetidos a estudos de epidemiologia molecular há predomínio dos subtipos B e Fl (Russell *et al.*, 2000). Um estudo recente mostrou grande prevalência de formas recombinantes B /Fl na Argentina e, em casuística mais limitada, no Uruguai e Bolívia (Carr *et al.*, 2001). No Brasil, o mais afetado país da América do Sul e que mostra ter um mosaico de subepidemias regionais, encontram-se também vários

subtipos. O de maior prevalência é o B, mas encontramos também o F, C e os recombinantes B/C e B/F (Morgado *et al.*, 1998; Couto-Fernandez *et al.*, 1999; Bongertz *et al.*, 2000; Guimaraes *et al.*, 2001; Soares *et al.*, 2003).

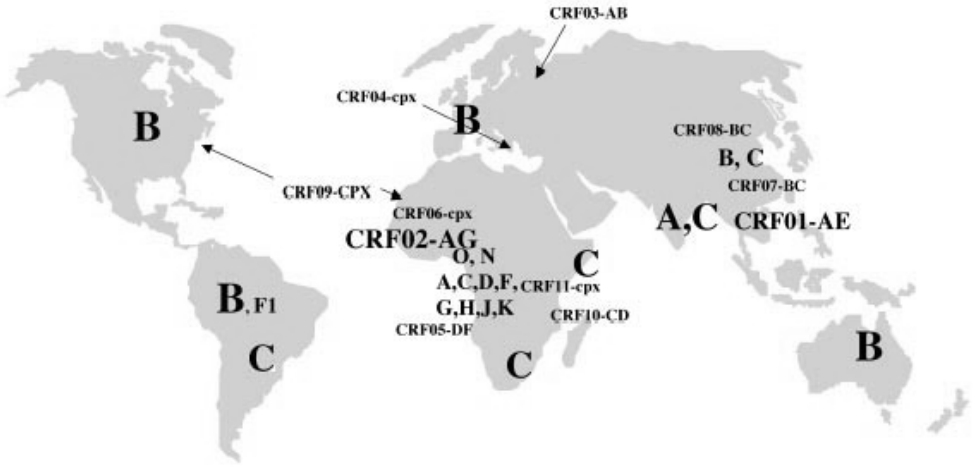

Figura 2 - Distribuição global dos subtipos de HIV-1.

V. Métodos empregados nos estudos filogenéticos

Diferentes tipos de dados podem ser utilizados na investigação das relações evolutivas entre genes e organismos. A maneira clássica de estimar a relação entre as espécies é comparar suas características morfológicas. Deste modo, a taxonomia é baseada principalmente na morfologia. As informações crescentes em biologia molecular, tais como seqüenciamento de nucleotídeos ou aminoácidos, podem também ser usadas para inferir relações filogenéticas. Os anos 80 foram marcados por grandes discussões sobre a importância dos dados moleculares nestas inferências (Patterson, 1987). Nos dias de hoje, entretanto, a utilização dos métodos aplicados ao estudo das relações entre genes já se encontra abalizada.

A biodiversidade existente nos dias atuais é o resultado de múltiplas aquisições genéticas incluindo mutações, reorganização de genomas e recombinações. De todas estas alterações, somente as mutações (pontuais, deleções ou inserções) são usadas pelos diferentes métodos moleculares de inferência filogenética.

Para se executar estes métodos, precisamos considerar a similaridade entre os genes estudados e assumirmos sua homologia. Embora supondo a existência de um ancestral comum, com o passar do tempo, é possível que duas seqüências nucleotídicas possuam variações suficientes para não conter informações suficientes sobre sua similaridade. Por isto, homologia só ocorre quando um ancestral comum é recente o suficiente para reter estas informações.

Quando duas seqüências são comparadas, podemos sempre calcular o percentual de similaridade contando o total de nucleotídeos ou aminoácidos idênticos entre elas. Quanto maior o grau de similaridade, maior a possibilidade de que sejam homólogas.

Comparações taxonômicas mostram que genes de espécies intimamente relacionadas se diferenciam por possuírem mutações pontuais entre si, geralmente na posição do terceiro códon que tem uma taxa de evolução mais rápida do que as do primeiro e segundo códon. As seqüências genéticas podem também possuir pequenas inserções ou deleções de nucleotídeos. Genes de espécies mais distantemente relacionadas diferem

entre si por um grande numero de mudanças de um mesmo tipo, mantendo regiões conservadas somente naquelas partes que codificam sítios catalíticos ou das proteínas do core. As diferenças entre espécies proximamente relacionadas serão mais facilmente acessadas através da análise de suas seqüências nucleotídicas e os relacionamentos mais distantes serão mais bem analisados comparando-se as seqüências de aminoácidos.

O processo de análise filogenética pode ser sumariado em cinco etapas. As primeiras duas são preparatórias para as etapas subseqüentes que envolvem a construção e a avaliação da árvore propriamente dita.

A primeira etapa é a da obtenção e alinhamento dos nucleotídeos ou das seqüências de aminoácidos de interesse. Como vimos acima, é fundamental a existência de sítios homólogos entre as seqüências estudadas e, por esta razão, as seqüências homólogas sob investigação são alinhadas de modo a formarem colunas como na figura 3.

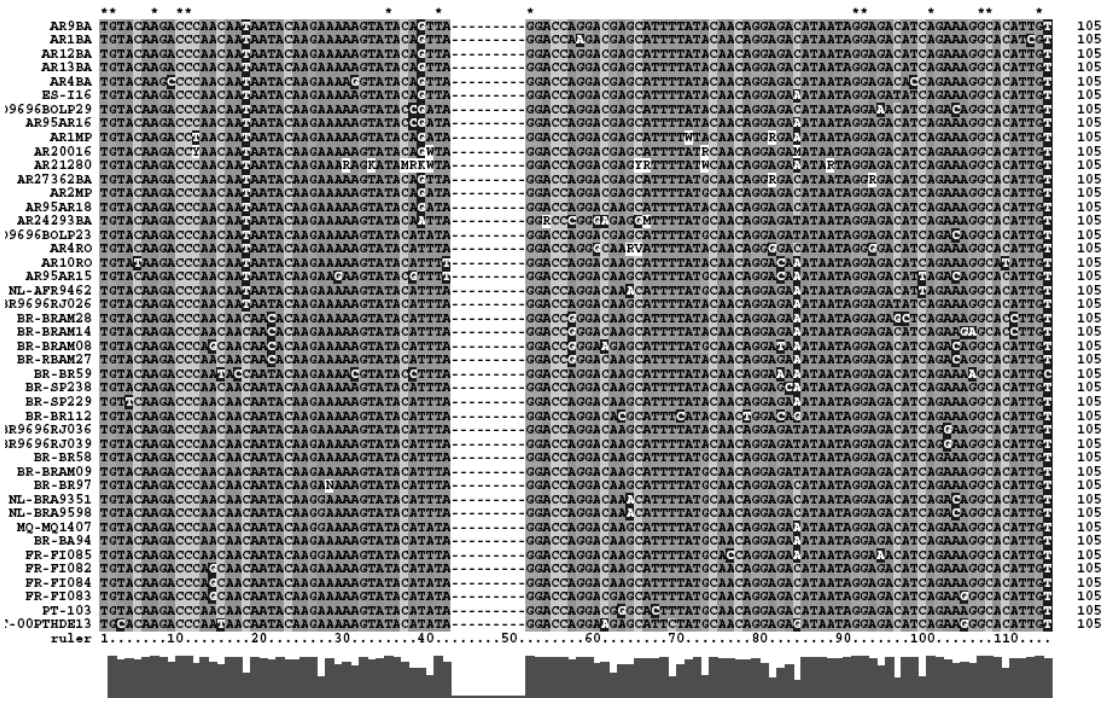

Figura 3 – Exemplo de alinhamento seqüencial múltiplo obtido pelo ClustalX(Thompson *et al.*, 1997) – os asterísticos no topo evidenciam as áreas conservadas, o histograma abaixo nos permite visualizar as áreas com maior similaridade.

Obter o alinhamento correto entre seqüências nucleotídicas é simples e pode ser feito até mesmo manualmente em um editor de textos. Quanto mais distante a relação entre as seqüências, mais difícil pode ser a execução do alinhamento. Deste modo, os alinhamentos são feitos normalmente com a utilização de programas de computadores que usam algoritmos particulares, seguidos de ajustes manuais.

A maioria dos algoritmos começa comparando a similaridade de todas as seqüências aos pares, alinhando, primeiramente, as duas com maior similaridade. As outras seqüências são, então, adicionadas progressivamente. O alinhamento continua de modo iterativo, com a adição de gaps na mesma posição para todos os membros de um mesmo agrupamento para se manter a homologia. Obter um bom alinhamento é crucial para a construção de uma árvore filogenética (AF) confiável, de modo que somente com um bom alinhamento poderemos submeter as seqüências aos programas específicos de inferência filogenética.

A segunda etapa será determinar a presença de um sinal filogenético. A maioria das análises de seqüências de DNA cai entre dois extremos: seqüências idênticas e seqüências que se tornaram tão divergentes que perderam a homologia. No primeiro extremo não há análise a ser feita e no segundo, o resultado que se pode obter não é confiável o suficiente para valer o esforço. Aqueles alinhamentos, conforme vimos acima, que possuam similaridade o suficiente para serem homólogas, poderão, de fato, serem objetos de estudos filogenéticos.

Uma vez que o alinhamento esteja completo, a etapa é a de decidir o método mais apropriado para a construção da AF. Finalmente, a árvore obtida deve ser examinada para se determinar o nível de confiabilidade a ser creditada em seus resultados (Hillis *et al.*, 1993).

A. A árvore filogenética.

A criação de árvores filogenéticas nos permite organizar nossos pensamentos quanto a uma seqüência genética e suas relações com outras afins. O exame destas árvores serve para determinar quão perto ou distante nossa seqüência de estudo se encontra de uma outra bem conhecida.

Uma AF é composta de linhas chamadas de ramos que se juntam para formar nós. Os nós nos extremos representam as taxa ou, no caso de seqüências, as próprias seqüências. O nó interno representa a seqüência imediatamente ancestral. A figura 4 ilustra uma árvore enraizada com seus ramos e nós.

Normalmente assumimos que o processo evolucionário é binário e resulta em um nó bifurcado que possui somente duas linhagens descendentes. Entretanto, nem sempre temos dados suficientes para inferir qual das espécies descende de um único ancestral comum, o que gera um nó multirramificado (politomia).

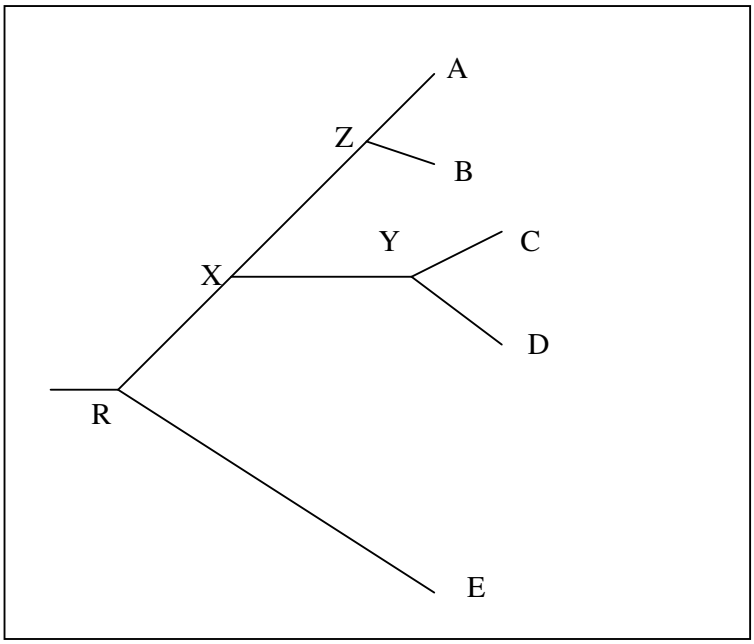

Figura 4 - Árvore enraizada – as extremidades dos ramos representam as seqüências em um subtipo, os quatro nós internos (R, X, Y e Z) representam as seqüências ancestrais.

Uma árvore desenraizada somente posiciona as seqüências umas em relação às outras sem, no entanto, mostrar a direção evolutiva. A árvore é dita enraizada se existe um nó particular, a raiz, para onde convergem todas as seqüências e pode ser obtida se uma ou mais unidades taxonômicas operacionais (UTO), isto é, nó terminal ou taxa, formarem um grupo externo que se acredita estarem o mais distantemente relacionado às UTO do grupo interno. O nó raiz é aquele que agrupa o grupo interno e externo representando, deste modo, o ancestral comum de ambos.

Uma AF pode ser enraizada mesmo sem que tenhamos certeza da UTO a escolher como grupo externo. Assumindo que a taxa de evolução é similar nas diferentes seqüências estudadas, disporemos a raiz no meio do caminho que junta as UTO mais dissimilares ou no ponto médio dos caminhos que juntam as conectadas a uma única margem (raiz central).

Ao enraizarmos uma árvore não devemos escolher um grupo externo relacionado de forma muito distante com o grupo interno. Isto pode resultar em sérios erros topológicos porque vários sítios podem estar saturados com mutações múltiplas e informações importantes podem ter sido apagadas. Também não devemos escolher um grupo externo que seja relacionado ao interno de forma muito próxima já que, neste caso, o grupo externo não seria verdadeiramente representativo. Muitas vezes o uso de mais de um grupo externo melhora, de modo geral, a topologia da árvore.

Nas figuras 5 a 7, mostramos as diversas formas de se publicar uma AF. As distâncias poderão estar relacionadas diretamente ao ramo, como no caso do cladograma da figura 7 ou, de forma indireta, com a adição de uma régua referência.

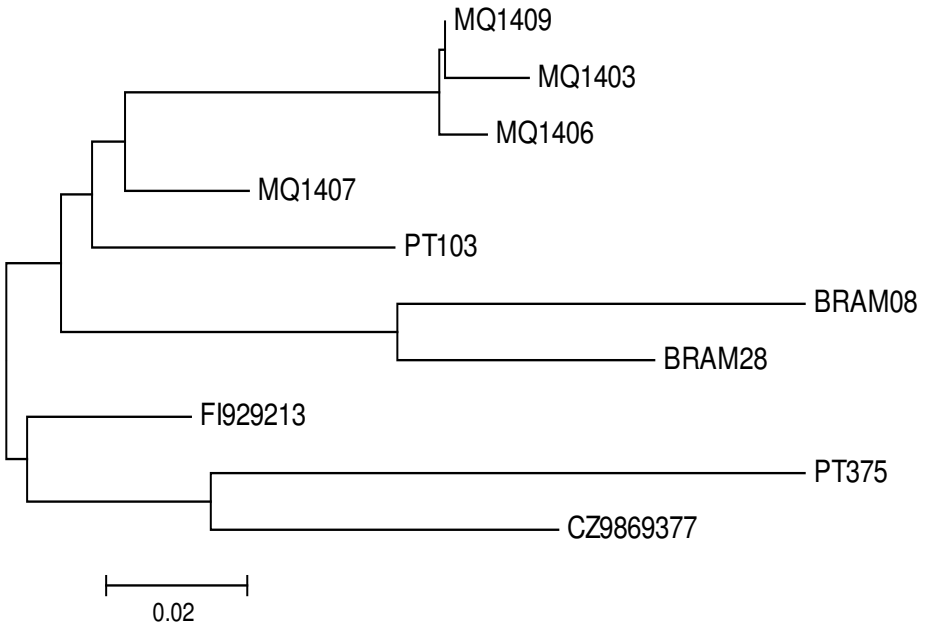

Figura 5 – Filograma obtido pelo método de agrupamento de vizinhos entre 10 seqüências. A distância pode ser medida com a régua de referência na base da árvore filogenética.

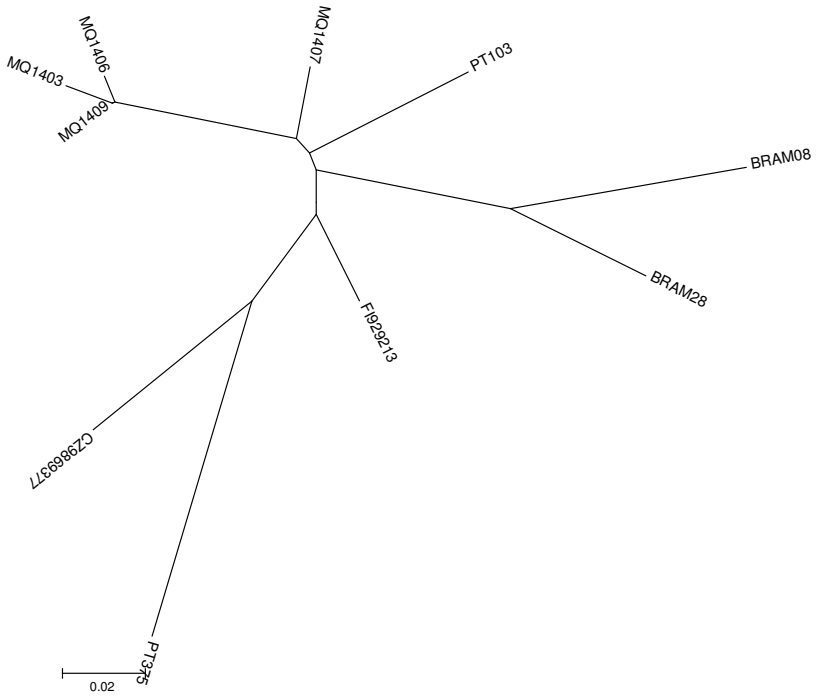

Figura 6 – Exemplo de árvore filogenética não enraizada obtida através do método de agrupamento de vizinhos.

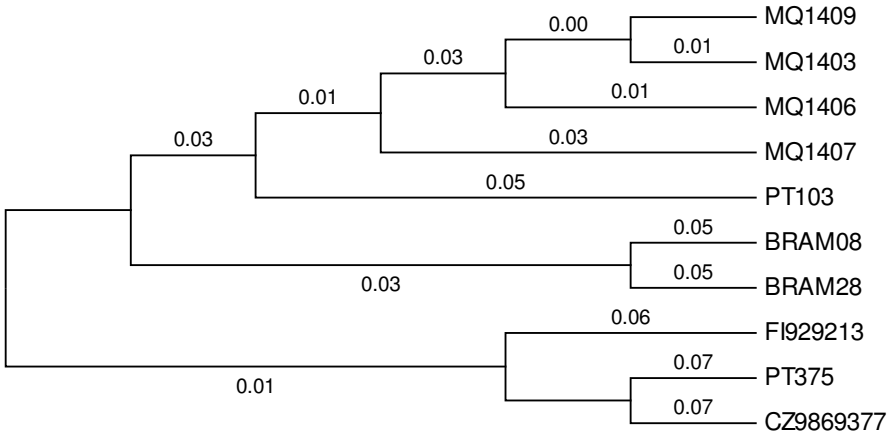

Figura 7 – Cladograma representativo do mesmo grupode seqüências das figuras anteriores. As distâncias foram aplicadas aos ramos.

B. Inferência de árvores filogenéticas.

A reconstrução filogenética a partir de um gene ou alinhamento de seqüência de aminoácidos nem sempre nos leva a conclusões acertadas. A reconstrução resulta numa AF inferida que pode diferir ou não da AF verdadeira. Muitos são os métodos utilizados na construção de AF já que não existe um modo correto único para isto.

Os métodos de construção de AF de dados moleculares podem ser agrupados primeiro por usarem estados de caracteres discretos ou matrizes de distâncias de pares dissimilares; segundo, por agruparem progressivamente as UTO, resultando numa única melhor árvore, ou por considerar todas as árvores teóricas possíveis.

Os **métodos de estado de caracteres** podem usar quaisquer caracteres discretos, como caracteres morfológicos, propriedades fisiológicas ou seqüência de dados. Cada posição no alinhamento é composta de um caractere e as posições dos nucleotídeos ou aminoácidos são chamadas de estado. Todos os caracteres são analisados separadamente, de forma independente uns dos outros. Os métodos de estados de caracteres guardam o estado original do caractere e, por isso, podem ser usado na reconstrução do estado do caractere dos nós ancestrais.

Em contraste, os **métodos de matrizes de distância** se iniciam pelo cálculo de algumas medidas de dissimilaridade entre cada par de UTO para produzir a matriz de distância aos pares e, assim, estimar a relação filogenética da UTO daquela matriz. Esses métodos parecem bem satisfatórios para a análise de seqüência de dados. Embora seja possível calcular as distâncias diretamente das seqüências alinhadas aos pares, resultados mais consistentes são obtidos quando todas as seqüências estão alinhadas. Métodos de matrizes de distância permitem escores para alinhamentos múltiplos.

Quando duas ou mais seqüências são divergentes, é provável que tenha ocorrido duas ou mais mutações consecutivas. Esses eventos múltiplos resultam em duas seqüências que são mais distantemente relacionadas do que se poderia deduzir pelas diferenças percentuais entre as seqüências. Quanto mais divergentes são as seqüências, maior o impacto dos eventos múltiplos. Modelos matemáticos nos permitem corrigir as

diferenças percentuais entre estas seqüências. Todo este processo é chamado de distância evolucionária ou distância genética, que é sempre maior que o calculado pela comparação direta entre as seqüências.

Métodos de distância descartam o estado original do caractere da seqüência. Como resultado, a informação necessária à reconstrução do nó ancestral é perdida. A maior vantagem destes métodos é que eles necessitam de menor atividade computacional para serem obtidos, o que facilita quando temos muitas seqüências a serem analisadas.

A tabela 2 lista os principais métodos de construção de AF classificados de acordo com a estratégia utilizada.

	Busca exaustiva	Agrupamento progressivo
Estado de caractere	Máxima parcimônia Máxima verossimilhança	
Matrizes de distância	Fitch-Margoliash	UPGMA* Agrupamento de visinhos

Tabela 1 - Métodos de construção de AF.

* Unweighted pair group method with arithmetic means.

i. Busca exaustiva

A busca exaustiva é utilizada para examinar possibilidades teóricas de obtenção de AF para um dado número de seqüências e o uso de certos critérios para escolha da melhor. Pode ser realizada através da máxima verossimilhança ou da máxima parcimônia.

a) Máxima verossimilhança

A máxima verossimilhança, em particular, possui as maiores vantagens como método de busca exaustiva e produz o maior número de árvores diferentes além de estimar, para cada uma delas, a probabilidade de representarem a AF real. Isto permite ao investigador suporte para comparar a melhor AF com a segunda melhor AF e estimar, também, o intervalo de confiança. Entretanto, quanto maior o número de seqüências

adicionadas ao estudo, maior o tempo e a atividade computacional, o que muitas vezes inviabiliza o método. O número de árvores bifurcadas para n UTO é dado pela equação 1.1:

$$(2n-3)!/\left(2^{n-2}(n-2)\right)! \qquad (1.1)$$

Como podemos perceber pela tabela 2, dificilmente poderemos lidar com dados contendo mais de 10 seqüências. Acima disto, devemos fracioná-los em extratos.

Número de UTO	Número de AF enraizadas
2	1
3	3
4	15
5	105
6	954
7	10395
8	135135
9	2027025
10	34459425

Tabela 2 - Número de AF possíveis para até 10 UTO.

b) Máxima Parsimônia (MP)

Objetiva encontrar a topologia da árvore para um grupo de seqüências que possa ser explicada com o menor número de mudanças de caracteres (mutações). O algoritmo da MP calcula a probabilidade da esperança de cada nucleotídeo (ou aminoácido) no nó ancestral (interno) e infere a chance da ocorrência de cada topologia para aquela probabilidade. Trata-se de um processo complexo, principalmente porque diferentes topologias de árvores requerem diferentes tratamentos matemáticos, o que demanda atividade computacional elevada.

c) Método Fitch-Margoliash

O método Fitch-Margoliash é um método de matrizes de distância que avalia todas as possibilidades de árvores para o menor tamanho de ramo usando um algoritmo específico que considera as distâncias entre os pares.

ii. Agrupamento progressivo

Os métodos de agrupamento progressivo demandam menos atividade computacional por examinar primeiro sub-árvores locais. São chamados de métodos de construção de AF por seguirem algoritmos específicos na construção de uma única árvore. Normalmente, as duas UTO mais intimamente relacionadas são combinadas de modo se tornar um grupo. Este grupamento é então tratado como uma única UTO que representa o ancestral das duas anteriores e deste modo a complexidade dos dados é reduzida. Este processo é repetido, agrupando-se sucessivamente as UTO proximamente relacionadas até que todas se combinem em uma única UTO.

Como só produz uma única AF, não é possível a obtenção dos estimadores de confiança pelo método de agrupamento progressivo, embora vários outros métodos estatísticos tenham sido desenvolvidos com esta finalidade. A maioria dos métodos de matrizes de distância usa o agrupamento progressivo para computar a melhor AF, enquanto os métodos de estado de caractere adotam a busca exaustiva.

a) Método de grupos pareados não ponderados com médias aritméticas (UPGMA)

É provavelmente o mais antigo e mais simples método de construção de AF por matrizes de distância. O agrupamento é feito pela procura do menor valor na matriz de distância do par. O novo agrupamento formado substitui a UTO que ele representa na matriz de distância. Este processo é repetido até todas UTO serem agrupadas. No UPGMA, a distância dos novos grupamentos formados é a média das distâncias da UTO original. Este processo assume que a taxa evolucionária do nó de dois grupamentos de

UTO a cada UTO original é a mesma. Assim, todo o processo de agrupamento assume que a taxa evolucionaria é idêntica para todos os ramos, o que significa dizer que nenhuma seqüência acumula mutações em um ritmo maior que a das outras. É óbvio que este pressuposto nem sempre é verdadeiro, o que nos mostra que o método nos dá AF falsas quando as taxas evolutivas são diferentes nos vários ramos.

b) Agrupamento de vizinhos (NJ)

O método de agrupamento de vizinhos (conhecido mesmo entre nós como *neighbor-joining*) constrói as AF por descobrir pares de vizinhos de forma seqüencial que são os pares de UTO ligados por um nó interior simples. O agrupamento utilizado por este algoritmo é completamente diferente dos descritos anteriormente porque não se preocupa em agrupar as UTO mais proximamente relacionadas, mas minimizar o tamanho de todos os nós internos e, assim, o tamanho de toda a árvore. Pode ser visto, então, como o método da parcimônia aplicado aos dados de matrizes de distância. O algoritmo do método de NJ se inicia com uma árvore semelhante a um arbusto, sem ramos internos. Inicialmente, introduz o primeiro ramo interno e calcula o tamanho da árvore resultante. O algoritmo liga sequencialmente os possíveis pares de UTO e, no final, junta o par que leva à menor árvore. O tamanho do ramo de um agrupamento de um par de vizinhos, X e Y, ao seu nó adjacente é baseado na distância média entre todas UTO e X para o ramo do X e todas UTO para Y para o ramo do Y, menos a distância media de todas as UTO remanescentes. O processo é assim repetido sempre juntando dois pares vizinhos de UTO baseando-se no menor ramo interno possível.

Este método se utiliza, de fato, do critério de mínima evolução (ME) e combina pares de seqüências minimizando os valores de S (melhor filogenia estimada) na equação 1.2, onde n é o número de UTO na árvore e V_i o i^{o} ramo.

$$S = \sum_{i=1}^{2n-3} V_i \qquad (1.2)$$

Não há como pressupor ser um método melhor do que o outro. Sugere-se que se aplique mais de um método aos mesmos dados. O método de máxima verossimilhança estima intrinsecamente os erros padrões para o tamanho dos ramos, o que confere suporte estatístico para cada tamanho de ramo e para toda a árvore. Para os outros métodos, a forma usual de se avaliar a qualidade da AF obtida é o de bootstraping.

C. Estimando a confiabilidade da árvore inferida

As duas técnicas mais amplamente empregadas na estimação da confiabilidade da árvore inferida são as análises de bootstrap e de jackknife. Estas técnicas se iniciaram com os trabalhos de (Mueller & Ayala, 1982) que usaram a abordagem jackknife para estimar a variância do tamanho dos ramos obtidos por UPGMA seguidos pelo trabalho de (Felsenstein, 1985) propondo o uso do bootstrap. Ambas as técnicas usam informações empíricas sobre a variação de um caractere para outro durante o processo evolucionário e, apesar de diferirem entre si, são da mesma família de técnicas.

A técnica de jackknife, a mais antiga das duas, escolhe uma observação de uma amostra por tempo determinado e faz a estimação. A variabilidade da estimação é dada por extrapolação da inferência de quão pequena é a variação que ela promove. O bootstrap, por sua vez, envolve a reamostragem com troca por uma amostra fictícia do mesmo tamanho que a amostra original.

i. A análise bootstrap

A análise bootstrap é amplamente utilizada como técnica de estimação de erros estatísticos em situações em que a distribuição na amostra original é desconhecida ou é de difícil derivação analítica.

Primeiro a seqüência de dados é trabalhada de forma a se obter um novo alinhamento a partir da amostra original através da escolha randômica de uma de suas colunas. Cada coluna no alinhamento pode ser escolhida mais de uma vez até que se obtenha uma nova amostra da seqüência, o que chamamos de replicação. Assim, no processo de

reamostragem, alguns caracteres não serão incluídos nunca, enquanto outros poderão ser incluídos repetidas vezes.

Segundo, para cada conjunto de dados replicados, uma árvore é construída e as diferenças entre as distâncias dos ramos nas novas árvores obtidas são computadas. Esta diferença entre as proporções é considerada a estimação da confiabilidade que dá suporte á árvore original.

Os valores de bootstrap dependerão do número de replicações feitas e poderão ser mostrados na AF de duas maneiras: a primeira, sumarizando os resultados da comparação numa árvore consensual e a segunda, sobrepondo os valores de bootstrap da árvore consensual na AF original.

ii. A análise jackknife

Serve como alternativa à análise de reamostragem especificamente usada para avaliação de confiabilidade de determinados descendentes numa AF. Aleatoriamente deleta-se a metade dos sítios de uma seqüência original de modo que a seqüência obtida é exatamente a metade da original. Esta reamostragem será feita várias vezes para gerar inúmeras novas amostras. Cada nova amostra será a base para uma reconstrução filogenética. A freqüência das sub-árvores são, então, contadas a partir das árvores reconstruídas. Se uma sub-árvore aparecer em todas as árvores reconstruídas, então o valor de jackknife será de 100%, o que confere maior confiabilidade.

Tanto para bootstrap ou para jackknife, valores inferiores a 70% deverão ser tratados com cautela.

D. Modelos de evolução de DNA

As seqüências divergem de um ancestral comum porque ocorrem mutações que em algum grau se transfere para as populações posteriores, por seleção ou por chance. De

forma a reconstruir o processo evolucionário através das AF precisamos pressupor um modelo de evolução.

Neste ponto, torna-se importante lembrar que as substituições nucleotídicas seguem alguns parâmetros. Incorporações de erros de reposição de purina por purina e pirimidina por pirimidina são, por razões estéricas, mais fáceis de ocorrer. O resultado desta mutação é o que chamamos de transição. Quando purina troca com pirimidina ou o reverso, chamamos de transversão.

Deste modo, existem quatro possibilidades de ocorrerem erros de transição ($A \leftrightarrow G, C \leftrightarrow T$) e oito possibilidades para os erros de transversão ($A \leftrightarrow C, A \leftrightarrow T, G \leftrightarrow C, G \leftrightarrow T$). Assim, se a mutação ocorre ao acaso, as transversões aconteceriam duas vezes mais facilmente que as transições. Entretanto, ao contrário do que matematicamente se poderia prever, na natureza se observa exatamente o oposto, com as transições ocorrendo duas vezes mais do que as transversões. Este é o parâmetro padrão utilizado nos modelos de substituição nucleotídica.

i. Modelo de Jukes-Cantor (JC69)

O modelo mais fácil de se considerar é aquele em que a probabilidade de um nucleotídeo ser trocado por um outro, numa determinada posição, num determinado tempo, ocorre de forma igual. Para inferir sobre esta probabilidade precisamos saber a taxa instantânea de troca para aquela posição. Este modelo simples, de um único parâmetro, é conhecido como modelo de Jukes-Cantor ou modelo de um parâmetro (Jukes & Cantor, 1969).

Se soubermos que existe um determinado nucleotídeo, G, em determinada posição, no tempo t = zero, podemos inferir qual a probabilidade desta posição continuar mantendo o mesmo nucleotídeo num tempo t posterior e qual é a probabilidade desta posição ter recebido um outro nucleotídeo, A, em troca. Isto é expresso como $P_{(GG)}(t)$ e $P_{(GA)}(t)$. Se a taxa de substituição por unidade de tempo é dada por α, então:

$$P_{(GG)}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \quad \text{e} \quad P_{(GA)}(t) = \frac{1}{4} + \frac{1}{4}e^{-4\alpha t} \qquad (1.3)$$

Quando t é próximo de zero, a probabilidade da posição não ter recebido uma troca é muito próxima de um.

Outros modelos importantes são os modelos gerais não reversíveis os quais citaremos alguns.

ii. Modelo de Kimura dois-parâmetros (K80)

Kimura (1980) introduziu um modelo que permite a ocorrência de desigualdade entre as taxas de transição/transversão (Kimura, 1980). A figura 8 mostra os dois parâmetros, α e β , que nos permite variar não somente a taxa total de substituição por unidade de tempo, como também a fração dessas substituições que são transições e as que são transversões. Devemos deixar claro que, para qualquer nucleotídeo, pode haver uma troca a uma taxa α, o que causa uma transição, e a uma taxa β, que causa transversão. A razão de transição/transversão, que chamaremos de *R* será α/(2β). A taxa total de troca será de α+2β.

Como podemos ver, o modelo é simétrico e, após tempo suficiente, a probabilidade será a mesma de termos uma purina ou uma pirimidina na posição estudada. Para as quatro bases possíveis, teríamos, então, como no modelo de Juckes-Cantor, a probabilidade de 1/4 para cada uma delas. Vale a pena salientar que o modelo JC69 é simplesmente um caso particular do modelo K80, onde α = β (*R* = 1/2).

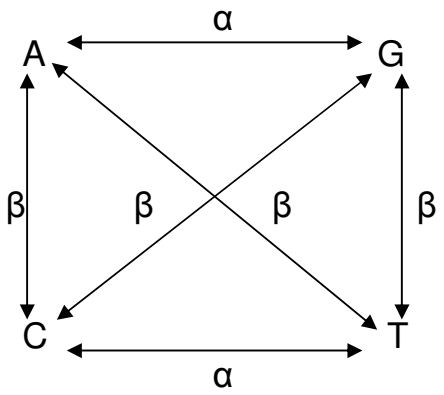

Figura 8 - Modelo K80

iii. Modelos de Tamura-Nei, F84 e HKY

O modelo K80 e o J69 oferecem grandes restrições para as seqüências nucleotídicas em detrimento de uma maior facilidade computacional. Seria interessante abrir mão do pressuposto de que todas as quatro bases teriam probabilidades iguais de freqüência. Com cinco parâmetros, dois dos principais modelos que relaxam esta condição são os F84, implementado por Felsenstein no pacote de filogenia do programa PHYLIP (Felsenstein, 1993) desde 1984, e o HKY, introduzido em 1985 por Hasegawa, Kishino e Yano. Ambos são similares e estendem o modelo K80 para aceitar freqüências assimétricas de bases com pequenas diferenças entre si. Em 1993, Tamura e Nei introduziram o modelo com seis parâmetros.

O modelo de Tamura-Nei idealiza que para cada posição nucleotídica temos a possibilidade de duas espécies de eventos. Se a posição possui uma purina (A ou G), ela possui uma probabilidade α_R por unidade de tempo de ocorrer um evento do tipo I, que é a troca da base de forma aleatória a partir de um conjunto de purinas. Também possui a probabilidade constante de que a troca ocorra a partir de um conjunto de todas as quatro bases, neste caso, evento do tipo II. Quando a posição nucleotídica possui uma pirimidina, existe a probabilidade constante α_Y de trocar a base de forma aleatória a partir de um conjunto de pirimidinas além da troca a partir de um conjunto de todas as quatro bases (igualmente, eventos do tipo I e II).

Supõe-se que o conjunto das quatro bases possua a mesma freqüência individual, isto é, π_A, π_C, π_G e π_T, o que se espera no modelo. Para os conjuntos de purina e de pirimidina também se espera que a relação entre as bases seja constante; deste modo, no conjunto de purinas, temos A e G na proporção $\pi_A : \pi_G$. Se observarmos a freqüência total de purinas no conjunto, então, $\pi_R = \pi_A + \pi_G$ e as freqüência de A e G serão, respectivamente, π_A / π_R e π_G / π_R . O mesmo ocorre para as pirimidinas, com proporção $\pi_C : \pi_T$ e freqüências relativas π_C / π_Y e π_T / π_Y.

Os dois tipos de eventos, num curto espaço de tempo *dt*, ocorrem com probabilidade $\alpha_R dt$ e βdt, quando a base de origem for uma purina e $\alpha_Y dt$ e βdt, quando for uma pirimidina.

Finalizando, se o modelo possui $\alpha_R = \alpha_Y$, este será o modelo F84. Se $\alpha_R / \alpha_Y = \pi_R / \pi_Y$, será o HKY. Como podemos observar, ambos são casos especiais do modelo Tamura-Nei.

iv. Modelo geral de tempo-reverso (GTR)

Se as freqüências do equilíbrio das bases são π_A, π_C, π_G e π_T, então o modelo é reversível se:

$$\pi_i \, \mathrm{Prob}(j|i,t) = \pi_j \, \mathrm{Prob}(i|j,t) \qquad (1.4)$$

Isto significa que a probabilidade de começarmos com i numa das pontas de um ramo da AF e terminarmos com j na outra é a mesma probabilidade de que ocorra o contrário. Assim, não há como dizer quem é ancestral ou quem é descendente de quem.

É muito mais uma conveniência matemática do que uma razão biológica para que os modelos de evolução de DNA sejam considerados reversíveis. Entretanto, alguns dos modelos reversíveis se aproximam tanto da realidade que nos permitem definir uma raiz para a AF (o que normalmente não seria possível). A tabela 3 nos mostra as taxas instantâneas de troca para a maioria dos modelos GTR.

de: \ à:	A	G	C	T
A	—	$\pi_G\alpha$	$\pi_C\beta$	$\pi_T\gamma$
G	$\pi_A\alpha$	—	$\pi_C\delta$	$\pi_T\varepsilon$
C	$\pi_A\beta$	$\pi_G\delta$	—	$\pi_T\eta$
T	$\pi_A\gamma$	$\pi_G\varepsilon$	$\pi_C\eta$	—

Tabela 3 - Modelo geral de tempo-reverso para evolução de DNA(Lanave *et al.*, 1984).

As taxas mostradas na tabela foram ajustadas de modo a permitir uma troca de base por unidade de tempo. Como π_i são as freqüências do equilíbrio das bases, então o total de

trocas será dado pela soma dos elementos da diagonal contrária, multiplicada, cada uma, pela probabilidade da ocorrência inicial daquela base. Temos assim:

$$2\pi_A\pi_G\alpha + 2\pi_A\pi_C\beta + 2\pi_A\pi_T\gamma + 2\pi_G\pi_C\delta + 2\pi_G\pi_T\varepsilon + 2\pi_C\pi_T\eta = 1 \qquad (1.5)$$

Como em todos os modelos de alterações de DNA, se todos os oito parâmetros livres da matriz **A** forem especificados poderemos calcular a matriz de probabilidade de transição **P** para qualquer ramo de tamanho t, como se segue:

$$A = \begin{bmatrix} -(\pi_G\alpha + \pi_C\beta + \pi_T\gamma) & \pi_A\alpha & \pi_A\beta & \pi_A\gamma \\ \pi_G\alpha & -(\pi_A\alpha + \pi_C\delta + \pi_T\varepsilon) & \pi_G\delta & \pi_G\varepsilon \\ \pi_C\beta & \pi_C\delta & -(\pi_A\beta + \pi_G\delta\pi_G\delta + \pi_T\eta) & \pi_C\eta \\ \pi_T\gamma & \pi_T\varepsilon & \pi_T\eta & -(\pi_A\gamma + \pi_G\varepsilon + \pi_C\eta) \end{bmatrix} \qquad (1.6)$$

A princípio, a matriz de probabilidade **P** pode ser computada pela exponenciação da matriz **A**. Assim:

$$P(t) = e^{At} \qquad (1.7)$$

Na prática, isto deve ser feito de forma numérica, já que não há fórmula conveniente para os elementos de **P**(t).

VI. Programas computacionais utilizados em inferência filogenética

Existe uma quantidade significativa de pacotes computacionais para uso em estudos filogenéticos. Apresentaremos aqui somente os mais utilizados sem a pretensão, entretanto, de ensiná-los.

A. Programas computacionais de alinhamento de seqüências

i. BioEdit

O programa BioEdit é um editor de seqüências que roda em ambiente Windows versões 95, 98, NT, 2000 e XP e que pode ser utilizado para editá-las, alinhá-las manipulá-las e analisá-las. Pode ser obtido em http://www.mbio.ncsu.edu/BioEdit/bioedit.html e o manual em http://www.mbio.ncsu.edu/BioEdit/BioDoc.pdf, gratuitamente.

Este programa pode importar arquivos diretamente do "GenBank" em formato FASTA e sua interface gráfica facilita a edição manual das seqüências.

ii. ClustalX

O ClustalX é a interface gráfica do ClustalW para uso em ambiente Windows (Thompson *et al.*, 1994; Thompson *et al.*, 1997). Fornece um ambiente integrado para a realização de alinhamentos de múltiplas seqüências, além da análise de seus resultados. As seqüências são mostradas numa janela e recursos de cores e marcações permitem visualizar melhor os sítios conservados. Menus são ativados por simples cliques facilitando a seleção de opções para o alinhamento múltiplo. É possível a utilização do processo de copiar-e-colar e a seleção de sub-amostras para um novo realinhamento. A qualidade do alinhamento pode ser verificada através da observação de histograma colocado em sua base gráfica.

O alinhamento obtido pode ser exportado em vários formatos, facilitando a utilização de outros programas de inferência filogenética.

Pode ser obtido de forma gratuita em http://www-igbmc.u-strasbg.fr/BioInfo.

iii. DAMBE

O Data Analysis in Molecular Biology and Evolution (DAMBE) é um programa computacional que permite a obtenção, organização, manipulação, alinhamento e análise das seqüências de dados moleculares. Também pode ser utilizado para o cálculo das distâncias genéticas ou reconstruções filogenéticas (Xia & Xie, 2001).

Facilita muito a organização dos dados, principalmente ao tornar evidente seqüências repetidas ou de baixa qualidade.

Pode ser obtido em http://aix1.uottawa.ca/~xxia/software/software.htm.

B. Programas computacionais de construção de árvores filogenéticas

i. PHYLIP

Este é um pacote de programas para inferência filogenética distribuído gratuitamente em http://evolution.genetics.washington.edu/phylip.html escrito para ser utilizado por várias plataformas computacionais.

ii. MrBayes

É um programa para reconstrução de árvores filogenéticas pelo método baysiano. É mais efetivo quando utilizado em conjunto com o PAUP*

Encontra-se disponível em http://brahms.biology.rochester.edu/software.html.

iii. PAUP*

O PAUP* (phylogenetic analysis using parsimony* and other methods) é um programa computacional distribuído por Sinauer Associates. Apesar de ser uma versão beta, não é de livre distribuição e, segundo seu autor, foi disponibilizado para compra nesta fase por se tratar do programa de inferência filogenética que possui menos bugs (Swofford, 2003).

Os dados devem estar em formato NEXUS, o que pode ser obtido formatando-se as seqüências alinhadas pelo ClustalX.

Na interface Windows, recebe os comandos em linha. Alguns desses comandos (scripts) podem ser obtidos pela internet.

iv. MEGA

A principal proposta do MEGA (Molecular evulutionary genetics analysis) é de ser um programa computacional de comparação de seqüências moleculares de fácil compreensão à comunidade científica (Kumar *et al.*, 2001). Permite a construção de AF bem como definir as distâncias genéticas entre populações de seqüências previamente definidas pelo usuário. Tem sido usado na versão 2.1, entretanto, já se encontra disponível a versão 3 beta 6 (Kumar *et al.*, 2004).

C. Visualizadores de árvores filogenéticas

i. TreeView

TreeView é um programa de visualização e impressão de AF (Page, 1996). Lê a maioria das árvores em formato NEXUX, tais como aqueles produzidos pelo PAUP* e seus componentes, e do estilo de PHYLIP (inclusive aqueles produzidos pelo fastDNAml e o ClustalW). Está disponível para interface Macintosh e Windows.

Pode ser obtido gratuitamente em http://taxonomy.zoology.gla.ac.uk/rod/treeview.html.

D. Programas acessórios

i. MODELTEST

MODELTEST é uma calculadora simples escrita em ANSI C e compilada para rodar em Macintosh e Windows. É projetado comparar diferentes modelos aninhados de substituição do DNA em uma estrutura hipoteticamente hierárquica. Calcula a razão de probabilidade estatística $\delta = - 2 \log \Delta$ (onde Δ é a razão de verossimilhança) e a associa

a um p-valor numa distribuição quiquadrada com q graus de liberdade a fim rejeitar ou não diferentes hipóteses nulas sobre o processo de substituição do DNA. Calcula também a estimativa do critério de informação de Akaike (AIC) associada a cada contagem da probabilidade.

É utilizado após se rodar dentro de PAUP* um script que determinará o modelo de substituição a ser empregado na inferência filogenética posterior. Para a versão mais recente do PAUP*, beta 10, é necessário correções no script ou utilizar a versão 3.5 que corrige este problema e já está disponível gratuitamente em http://darwin.uvigo.es.

VII. Epidemiologia Molecular do Sub-subtipo F1 do HIV-1

Apresentamos, a seguir, um exercício de aplicação dos métodos de análise filogenética baseado na aquisição de seqüências do sub-subtipo F1 do HIV-1 depositadas na base de dados do Laboratório Nacional de Los Alamos (LANL). Este trabalho aborda a epidemia de HIV/Aids no Brasil e na Romênia dando importância aos aspectos da evolução molecular. O estudo visa estabelecer a relação evolutiva entre os dois países. Executamos as técnicas de inferência filogenética descritas acima e acreditamos que o pesquisador minimamente familiarizado com a questão da epidemiologia molecular possa reproduzi-lo ou ainda empreender novos estudos a partir de seqüências próprias. Em anexo, apresentamos o artigo completo.

A. MÉTODOS

Trabalhamos com seqüências obtidas da base de dados do Laboratório Nacional de Los Alamos (LANL), Estados Unidos, através de seu sítio eletrônico e ferramenta de busca em linha (http://hiv-web.lanl.gov/). Solicitamos todas as seqüências de HIV-1 que correspondessem aos seguintes parâmetros: sub-subtipo F1, que contivessem a região genômica V3, de qualquer região geográfica. As seqüências obtidas foram analisadas com o programa DAMBE (Xia & Xie, 2001) para a verificação das repetições. O alinhamento múltiplo foi realizado com o programa ClustalX versão 1.81(Thompson *et al.*, 1997; Chenna *et al.*, 2003). As seqüências alinhadas foram cortadas no tamanho da seqüência BR7494 e formatadas para uso no PAUP* - Phylogenetic Analysis Using Parsimony (*and Other Methods), versão 4.0b10 (Swofford, 2003), das bases 7078 a 7400 (da seqüência do HXB2), correspondentes à alça C2-V3 do gene env. As inferências filogenéticas foram feitas pelo método de aproximação de vizinhos (NJ) (Saitou & Nei, 1987) usando como modelo de evolução o "modelo geral de tempos reversíveis" com distribuição gama (GTR+G) (Lanave *et al.*, 1984; Rodriguez *et al.*, 1990), escolhido através da análise das seqüências pelo Modeltest 3.06 (Posada & Crandall, 1998). Os parâmetros do modelo sugerido foram: freqüência A = 0,4037; freqüência C = 0,1846; freqüência G = 0,1961e freqüência T = 0,2156. Os valores da matriz de classificação R foram: R(a) [A-C] = 1,6989; R(b) [A-G] = 3,3751; R(c) [A-T]

= 0,8474; R(d) [C-G] = 0,5380; R(e) [C-T] = 3,3751; R(f) [G-T] = 1. A proporção de sítios invariáveis foi de zero e o parâmetro da forma da distribuição gama de sítios heterogêneos variáveis de 0,6568. A estimação estatística da confiabilidade das árvores obtidas foi feita pelo método de bootstrap (Hall, 2001). As árvores filogenéticas foram formatadas para leitura no Tree-View versão 1.6.0 (Zhai *et al.*, 2002).

As medidas de dispersão entre os pares foram calculadas para os agrupamentos geográficos subdivididos em: brasileiras (BR), africanas (AF), romenas (RO) e demais localidades (OU). O cálculo das distâncias nucleotídicas foi obtido utilizando-se o modelo de Kimura 2-parâmetros (Kimura, 1980), com erro padrão estimado por bootstrap (1000 replicações), implementado no programa MEGA versão 2.1 (Kumar *et al.*, 2001). As médias das distâncias inter grupos foram também analisadas e comparadas utilizando o mesmo modelo.

B. RESULTADOS

Obtivemos 202 seqüências que foram analisadas no DAMBE. Excluímos as repetidas e selecionamos 80 taxa para o estudo. Incluímos ainda antes do alinhamento, a seqüência de referência SIV (acesso AF003038), que usamos como grupo externo, além da primeira seqüência brasileira identificada do subtipo F, BR7944, que serviu também como base para o corte após o alinhamento. Trabalhamos, assim, com 26 taxa brasileiras, 27 taxa romenas, 14 taxa africanas e 13 taxa de outras localidades. Havia, ainda, duas outras seqüências, as AR9515 e PTHDE13, que foram posteriormente retiradas por se tratar a primeira de vírus recombinante B/F e a segunda por não apresentar boa confiabilidade, o que comprometeria a qualidade do estudo. O filograma obtido encontra-se na figura 9. A média geral das distâncias nucleotídicas foi de 0,143, com erro padrão de 0,013. As distâncias inter e intragrupos podem ser vistas nas tabelas 4 e 5 (erro padrão estimado por bootstrap com 1000 replicações e 97396 reproduções aleatórias).

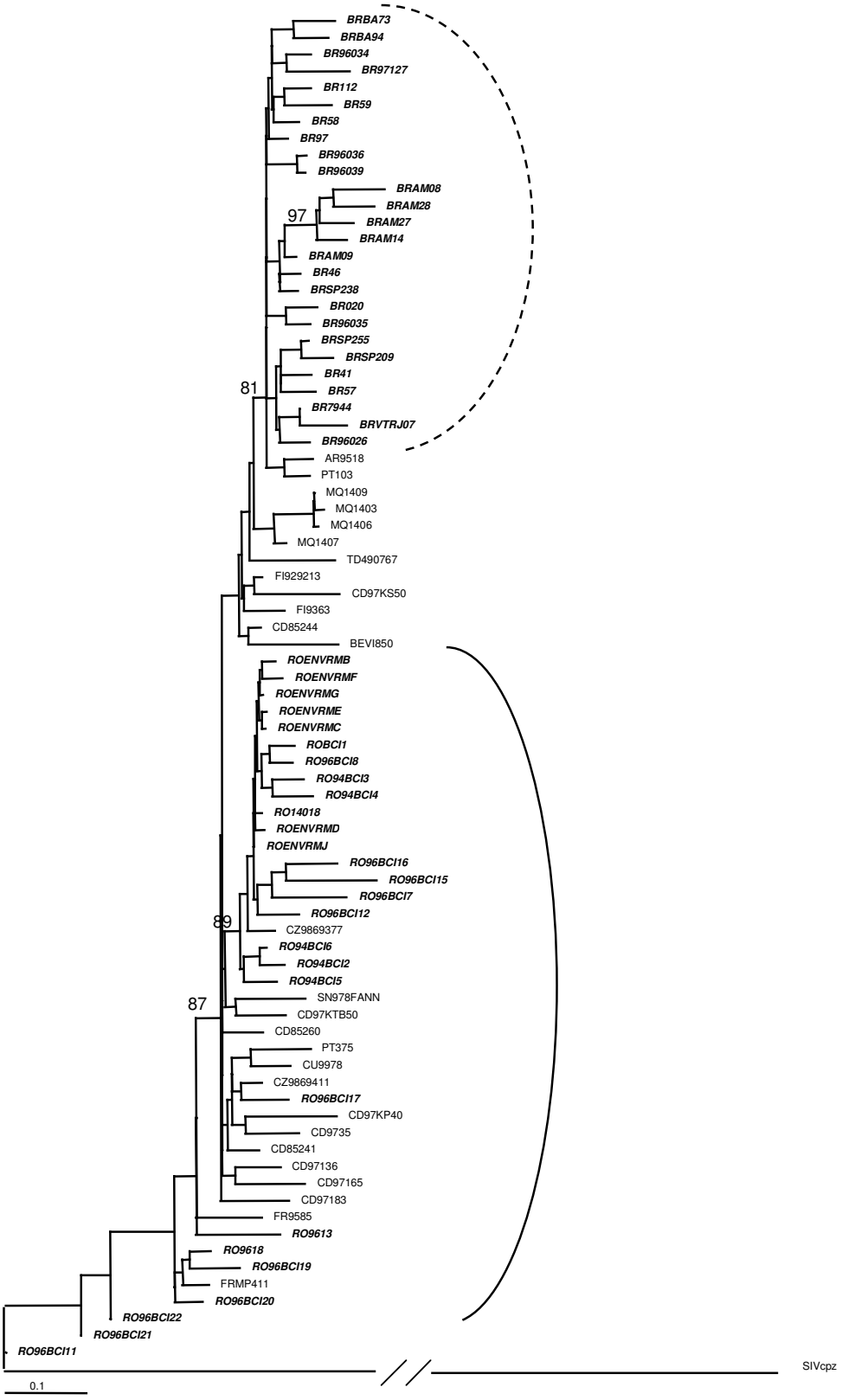

Figura 9 - Filograma obtido pelo método NJ. As seqüências brasileiras e romenas encontram-se destacadas. Valores de bootstrap aplicados aos principais ramos.

Grupo	d	S.E.
BR	0,099	0,010
OU	0,124	0,016
RO	0,104	0,011
AF	0,129	0,014

Tabela 4 - Distâncias intra subtipo estimadas pelo método de Kimura dois parâmetros (SE = erro padrão estimado pelo método de bootstrap com 1000 replicações)

	BR	OU	RO	AF
BR	*	[0,013]	[0,017]	[0,014]
OU	0,125	*	[0,016]	[0,014]
RO	0,139	0,140	*	[0,014]
AF	0,134	0,134	0,130	*

Tabela 5 - Distância entre a média dos grupos estimada pelo método de Kimura dois parâmetros. Erro padrão entre colchetes estimado pelo método de bootstrap (1000 replicações).

C. DISCUSSÃO

O reconhecimento de tipos e subtipos de HIV nos permite entender o crescimento da epidemia nas várias regiões geográficas e estabelecer estratégias específicas de prevenção.

No continente africano, onde se acredita ser o início da epidemia de HIV/Aids, já foram encontrados todos os subtipos de HIV-1 e HIV-2. No Brasil, o mais afetado país da América do Sul e que mostra ter um mosaico de subepidemias regionais, encontram-se também vários subtipos. O de maior prevalência é o B, mas encontramos também o F, C e os recombinantes B/C e B/F (Morgado *et al.*, 1998; Morgado *et al.*, 1998; Guimaraes *et al.*, 2002; Guimaraes *et al.*, 2002; Soares *et al.*, 2003).

A prevalência global de infecções por subtipo F é relativamente baixa. Até 1999 as seqüências representativas desta variante eram divididas em três grupos: F1, F2 e F3,

todas demonstráveis em solo africano (Triques *et al.*, 1999). As seqüências recuperadas de pacientes brasileiros, bem como as que ocorrem na Romênia, agrupavam-se com seqüências africanas da variante F1.

A introdução do subtipo F no Brasil ocorreu provavelmente em época posterior à introdução do subtipo B em razão da diversidade genética menos ampla destas seqüências. Este subtipo ganhou importância em crescimento e é a segunda variante mais comum, respondendo por boa parcela das infecções entre usuários de drogas endovenosas da cidade de São Paulo (Bongertz *et al.*, 2000) além de ser encontrado em cerca da metade dos casos de Manaus (Vicente *et al.*, 2000).

Casos isolados de seqüências representativas dos subtipos D e A também foram recentemente descritas no país, mas sempre que eram apropriadamente estudados, demonstravam ser vírus recombinantes ou infecções simultâneas por mais de um subtipo (Couto-Fernandez *et al.*, 1999; Ramos *et al.*, 1999).

O subtipo F foi também descrito entre 1989-1990 na Romênia, em crianças que viviam em orfanatos (Hersh *et al.*, 1991). Os estudos epidemiológicos demonstraram que a maioria das crianças se infectou por transmissão horizontal, ou por transfusão sangüínea ou por material médico-cirúrgico infectado. As taxas de dispersão nucleotídicas entre os indivíduos estavam entre 0,9% e 3,6% (Dumitrescu *et al.*, 1994). Com a adição de um número maior de seqüências, estas taxas passaram a variar entre 9,3% e 11,5% (tabela 4), o que demonstra serem geneticamente relacionadas, porém com um maior tempo de evolução entre elas. Estes dados sugerem uma única introdução comum de subtipo F nas seqüências estudas provavelmente com a contaminação de um adulto através de contato sexual em viagem fora da Romênia. As seqüências brasileiras se comportam de forma semelhante às romenas, com taxas de dispersão nucleotídicas variando entre 8,9% e 10,9% (tabela 4).

Observando o filograma da figura 9 verificamos que a topologia da árvore mostra os dois grupos de seqüências em ramos separados, com um ancestral comum relativamente distante. Se houvesse uma relação epidemiológica direta entre ambos, os agrupamentos sairiam integralmente um do outro. Podemos afirmar, deste modo, que existe uma relação evolucionária entre os subtipos F brasileiros e romenos, ainda que de modo limitado.

Tomando como base as taxas de dispersões nucleotídicas entre os grupos estudados, tanto o grupo brasileiro quanto o romeno possuem maior proximidade ao africano do que entre si, sugerindo introduções independentes de um ancestral comum.

D. Conclusão

O conhecimento da evolução genética do HIV é importante não somente para o entendimento dos mecanismos evolucionários básicos como também para a orientação de estratégias de controle e erradicação da Aids. Ainda que países distantes geográfica e culturalmente, o Brasil e a Romênia mantêm relações epidemiológicas evolutivas entre os HIV-1 do subtipo F. A análise filogenética das seqüências recuperadas da base de dados do Laboratório Nacional de Los Alamos (LANL) revela que, apesar de se agruparem em ramos diferentes, as seqüências brasileiras e romenas possuem um ancestral comum introduzidos num passado recente.

VIII. Conclusões gerais

A epidemia de HIV/Aids determinou uma investida sem precedentes na história das doenças infecciosas na busca do seu entendimento. Este esforço nos leva às pesquisas clínicas, imunológicas, epidemiológicas e de biologia celular, matemática e molecular. Para cada nova descoberta num dos campos envolvidos, observamos novas respostas de pesquisadores de disciplinas antes tão díspares.

De acordo com a teoria evolucionária, todos os organismos vivos possuem um ancestral comum. A rápida evolução do HIV e dos vírus RNA, em geral, os torna alvos ideais para aplicação de métodos filogenéticos. Através destes métodos, podemos entender um pouco mais a cerca de suas origens, diversidade e transmissibilidade.

Os estudos filogenéticos demonstraram a existência de um berço comum, no continente africano, onde se supõe o HIV tenha sua origem zoonótica a partir de primatas não humanos. Pudemos, ainda, a partir desses estudos, verificar a presença de vírus recombinantes e entender detalhadamente algumas formas de transmissão, especialmente as transmissões entre diferentes grupos em risco e as múltiplas transmissões para um mesmo indivíduo.

A análise filogenética também se mostra necessária como estimadora de taxas de mutações, probabilidade de substituições e outros parâmetros relacionados à presença de resistência anti-retroviral.

A criação de bancos de dados, com coleções significativas de seqüências genéticas do HIV disponibilizadas pela internet, associado a um maior poder computacional dos dias atuais aumentaram a colaboração entre os pesquisadores de todo o mundo facilitando a concepção de novos métodos filogenéticos.

A relação entre países distantes pode ser estabelecida através de estudos filogenéticos onde a busca de um ancestral comum pode explicar melhor este relacionamento. Ainda que geográfica e culturalmente distantes o Brasil e a Romênia mantêm relações epidemiológicas evolutivas entre os HIV-1 do subtipo F. A análise filogenética das

seqüências recuperadas da base de dados do Laboratório Nacional de Los Alamos (LANL) revelou que apesar de se agruparem em ramos diferentes, as seqüências brasileiras e romenas possuem um ancestral comum com introduções independentes num passado recente, provavelmente proveniente do continente africano.

IX. Referencias Bibliográficas

Alaeus, A. (2000). "Significance of HIV-1 genetic subtypes." Scand J Infect Dis **32**(5): 455-63.

Anderson, R. M., Schwartlander, B., McCutchan, F. & Hu, D. (1996). "Implications of genetic variability in HIV for epidemiology and public health." Lancet **347**(9018): 1778-9.

Barre-Sinoussi, F. (1996). "HIV as the cause of AIDS." Lancet **348**(9019): 31-5.

Blackard, J. T., Cohen, D. E. & Mayer, K. H. (2002). "Human immunodeficiency virus superinfection and recombination: current state of knowledge and potential clinical consequences." Clin Infect Dis **34**(8): 1108-14.

Bongertz, V., Bou-Habib, D. C., Brigido, L. F., Caseiro, M., Chequer, P. J., Couto-Fernandez, J. C., Ferreira, P. C., Galvao-Castro, B., Greco, D., Guimaraes, M. L., Linhares de Carvalho, M. I., Morgado, M. G., Oliveira, C. A., Osmanov, S., Ramos, C. A., Rossini, M., Sabino, E., Tanuri, A. & Ueda, M. (2000). "HIV-1 diversity in Brazil: genetic, biologic, and immunologic characterization of HIV-1 strains in three potential HIV vaccine evaluation sites. Brazilian Network for HIV Isolation and Characterization." J Acquir Immune Defic Syndr **23**(2): 184-93.

Broder, S., Merigan, T. C. & Bolognesi, D. (1999). Textbook of AIDS medicine. Baltimore, Williams & Wilkins.

Caride, E., Brindeiro, R., Hertogs, K., Larder, B., Dehertogh, P., Machado, E., de Sa, C. A., Eyer-Silva, W. A., Sion, F. S., Passioni, L. F., Menezes, J. A., Calazans, A. R. & Tanuri, A. (2000). "Drug-resistant reverse transcriptase genotyping and phenotyping of B and non-B subtypes (F and A) of human immunodeficiency virus type I found in Brazilian patients failing HAART." Virology **275**(1): 107-15.

Carr, J. K., Avila, M., Gomez Carrillo, M., Salomon, H., Hierholzer, J., Watanaveeradej, V., Pando, M. A., Negrete, M., Russell, K. L., Sanchez, J., Birx, D. L., Andrade, R., Vinoles, J. & McCutchan, F. E. (2001). "Diverse BF recombinants have spread widely since the introduction of HIV-1 into South America." Aids **15**(15): F41-7.

Carr, J. K., Salminen, M. O., Koch, C., Gotte, D., Artenstein, A. W., Hegerich, P. A., St Louis, D., Burke, D. S. & McCutchan, F. E. (1996). "Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand." J Virol **70**(9): 5935-43.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. & Thompson, J. D. (2003). "Multiple sequence alignment with the Clustal series of programs." Nucleic Acids Res **31**(13): 3497-500.

Coffin, J. M. (1979). "Structure, replication, and recombination of retrovirus genomes: some unifying hypotheses." J Gen Virol **42**(1): 1-26.

Corbet, S., Muller-Trutwin, M. C., Versmisse, P., Delarue, S., Ayouba, A., Lewis, J., Brunak, S., Martin, P., Brun-Vezinet, F., Simon, F., Barre-Sinoussi, F. & Mauclere, P. (2000). "env sequences of simian immunodeficiency viruses from chimpanzees in Cameroon are strongly related to those of human immunodeficiency virus group N from the same geographic area." J Virol **74**(1): 529-34.

Couto-Fernandez, J. C., Morgado, M. G., Bongertz, V., Tanuri, A., Andrade, T., Brites, C. & Galvao-Castro, B. (1999). "HIV-1 subtyping in Salvador, Bahia, Brazil: a city with African sociodemographic characteristics." J Acquir Immune Defic Syndr **22**(3): 288-93.

Couturier, E., Damond, F., Roques, P., Fleury, H., Barin, F., Brunet, J. B., Brun-Vezinet, F. & Simon, F. (2000). "HIV-1 diversity in France, 1996-1998. The AC 11 laboratory network." Aids **14**(3): 289-96.

Crandall, K. A. (1999). The evolution of HIV. Baltimore, MD, Johns Hopkins University Press.

Crandall, K. A. & Templeton, A. R. (1999). Statistical Approaches to Detecting Recombination. The evolution of HIV. K. A. Crandall. Baltimore, MD, Johns Hopkins University Press**:** 153 - 176.

Dean, M., Carrington, M., Winkler, C., Huttley, G. A., Smith, M. W., Allikmets, R., Goedert, J. J., Buchbinder, S. P., Vittinghoff, E., Gomperts, E., Donfield, S., Vlahov, D., Kaslow, R., Saah, A., Rinaldo, C., Detels, R. & O'Brien, S. J. (1996). "Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study." Science **273**(5283): 1856-62.

Diaz, R. S., Sabino, E. C., Mayer, A., Mosley, J. W. & Busch, M. P. (1995). "Dual human immunodeficiency virus type 1 infection and recombination in a dually exposed transfusion recipient. The Transfusion Safety Study Group." J Virol **69**(6): 3273-81.

Dumitrescu, O., Kalish, M. L., Kliks, S. C., Bandea, C. I. & Levy, J. A. (1994). "Characterization of human immunodeficiency virus type 1 isolates from children in Romania: identification of a new envelope subtype." J Infect Dis **169**(2): 281-8.

Felsenstein, J. (1985). "Confidence limits on phylogenies: An approach using the bootstrap." Evolution **34**: 152-161.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package). Seattle, Distributed by the author**:** version 3.5c.

Fields, B. N., Knipe, D. M., Howley, P. M. & Griffin, D. E. (2001). Fields' virology. Philadelphia, Lippincott Williams & Wilkins**:** 2 v. (xix, 3087, 72 p.).

Fleury, H., Recordon-Pinson, P., Caumont, A., Faure, M., Roques, P., Plantier, J. C., Couturier, E., Dormont, D., Masquelier, B. & Simon, F. (2003). "HIV type 1 diversity in France, 1999-2001: molecular characterization of non-B HIV type 1 subtypes and potential impact on susceptibility to antiretroviral drugs." AIDS Res Hum Retroviruses **19**(1): 41-7.

Foley, B., Pan, H., Buchbinder, S. & Delwart, E. L. (2000). "Apparent founder effect during the early years of the San Francisco HIV type 1 epidemic (1978-1979)." AIDS Res Hum Retroviruses **16**(15): 1463-9.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M. & Hahn, B. H. (1999). "Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes." Nature **397**(6718): 436-41.

Guimaraes, M. L., Bastos, F. I., Telles, P. R., Galvao-Castro, B., Diaz, R. S., Bongertz, V. & Morgado, M. G. (2001). "Retrovirus infections in a sample of injecting drug users in Rio de Janeiro City, Brazil: prevalence of HIV-1 subtypes, and co-infection with HTLV-I/II." J Clin Virol **21**(2): 143-51.

Guimaraes, M. L., dos Santos Moreira, A., Loureiro, R., Galvao-Castro, B. & Morgado, M. G. (2002). "High frequency of recombinant genomes in HIV type 1 samples from Brazilian southeastern and southern regions." AIDS Res Hum Retroviruses **18**(17): 1261-9.

Guimaraes, M. L., Moreira, A. S. & Morgado, M. G. (2002). "Polymorphism of the Human Immunodeficiency Virus Type 1 in Brazil: genetic characterization of the nef gene and implications for vaccine design." Mem Inst Oswaldo Cruz **97**(4): 523-6.

Gurtler, L. G., Hauser, P. H., Eberle, J., von Brunn, A., Knapp, S., Zekeng, L., Tsague, J. M. & Kaptue, L. (1994). "A new subtype of human immunodeficiency virus type 1 (MVP-5180) from Cameroon." J Virol **68**(3): 1581-5.

Hall, B. G. (2001). Phylogenetic trees made easy: a how-to manual for molecular biologists. Sunderland, Mass., Sinauer Associates.

Hersh, B. S., Popovici, F., Apetrei, R. C., Zolotusca, L., Beldescu, N., Calomfirescu, A., Jezek, Z., Oxtoby, M. J., Gromyko, A. & Heymann, D. L. (1991). "Acquired immunodeficiency syndrome in Romania." Lancet **338**(8768): 645-9.

Hillis, D. M. (1999). Phylogenetics and the Study of HIV. The evolution of HIV. K. A. Crandall. Baltimore, MD, Johns Hopkins University Press**:** 105 - 121.

Hillis, D. M., Allard, M. W. & Miyamoto, M. M. (1993). "Analysis of DNA sequence data: phylogenetic inference." Methods Enzymol **224**: 456-87.

Hirsch, V. M., Olmsted, R. A., Murphey-Corb, M., Purcell, R. H. & Johnson, P. R. (1989). "An African primate lentivirus (SIVsm) closely related to HIV-2." Nature **339**(6223): 389-92.

Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M. & Markowitz, M. (1995). "Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection." Nature **373**(6510): 123-6.

Holmes, E. C. (2003). "Molecular clocks and the puzzle of RNA virus origins." J Virol **77**(7): 3893-7.

Holmes, E. C. (2004). "The phylogeography of human viruses." Mol Ecol **13**(4): 745-56.

Hu, D. J., Dondero, T. J., Rayfield, M. A., George, J. R., Schochetman, G., Jaffe, H. W., Luo, C. C., Kalish, M. L., Weniger, B. G., Pau, C. P., Schable, C. A. & Curran, J. W. (1996). "The emerging genetic diversity of HIV. The importance of global surveillance for diagnostics, research, and prevention." Jama **275**(3): 210-6.

Janssens, W., Buve, A. & Nkengasong, J. N. (1997). "The puzzle of HIV-1 subtypes in Africa." Aids **11**(6): 705-12.

Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. Mammalian protein metabolism. H. N. Munro. New York, Academic Press**:** v. <2-4 >.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." J Mol Evol **16**(2): 111-20.

Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001). "MEGA2: molecular evolutionary genetics analysis software." Bioinformatics **17**(12): 1244-5.

Kumar, S., Tamura, K. & Nei, M. (2004). "MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and Sequence Alignment." Briefings in Bioinformatics **5**(2): in press.

Lanave, C., Preparata, G., Saccone, C. & Serio, G. (1984). "A new method for calculating evolutionary substitution rates." J Mol Evol **20**(1): 86-93.

Liu, R., Paxton, W. A., Choe, S., Ceradini, D., Martin, S. R., Horuk, R., MacDonald, M. E., Stuhlmann, H., Koup, R. A. & Landau, N. R. (1996). "Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection." Cell **86**(3): 367-77.

Louwagie, J., Janssens, W., Mascola, J., Heyndrickx, L., Hegerich, P., van der Groen, G., McCutchan, F. E. & Burke, D. S. (1995). "Genetic diversity of the envelope glycoprotein from human immunodeficiency virus type 1 isolates of African origin." J Virol **69**(1): 263-71.

McCutchan, F. E., Salminen, M. O., Carr, J. K. & Burke, D. S. (1996). "HIV-1 genetic diversity." Aids **10 Suppl 3**: S13-20.

Morgado, M. G., Guimaraes, M. L., Gripp, C. B., Costa, C. I., Neves, I., Jr., Veloso, V. G., Linhares-Carvalho, M. I., Castello-Branco, L. R., Bastos, F. I., Kuiken, C., Castilho, E. A., Galvao-Castro, B. & Bongertz, V. (1998). "Molecular epidemiology of HIV-1 in Brazil: high prevalence of HIV-1 subtype B and identification of an HIV-1 subtype D infection in the city of Rio de Janeiro, Brazil. Evandro Chagas Hospital AIDS Clinical Research Group." J Acquir Immune Defic Syndr Hum Retrovirol **18**(5): 488-94.

Morgado, M. G., Guimaraes, M. L., Neves Junior, I., dos Santos, V. G., Linhares-de-Carvalho, M. I., Castello-Branco, L. R., Bastos, F. I., Castilho, E. A., Galvao-Castro, B. & Bongertz, V. (1998). "Molecular epidemiology of HIV in Brazil: polymorphism of the antigenically distinct HIV-1 B subtype strains. The Hospital Evandro Chagas AIDS Clinical Research Group." Mem Inst Oswaldo Cruz **93**(3): 383-6.

Mueller, L. D. & Ayala, F. J. (1982). "Estimation and interpretation of genetic distance in empirical studies." Genet Res **40**(2): 127-37.

Myers, G. (1994). "Tenth anniversary perspectives on AIDS. HIV: between past and future." AIDS Res Hum Retroviruses **10**(11): 1317-24.

Page, R. D. (1996). "TreeView: an application to display phylogenetic trees on personal computers." Comput Appl Biosci **12**(4): 357-8.

Patterson, C. (1987). Molecules and morphology in evolution: conflict or compromise? Cambridge [Cambridgeshire]; New York, Cambridge University Press.

Peeters, M., Liegeois, F., Torimiro, N., Bourgeois, A., Mpoudi, E., Vergne, L., Saman, E., Delaporte, E. & Saragosti, S. (1999). "Characterization of a highly replicative intergroup M/O human immunodeficiency virus type 1 recombinant isolated from a Cameroonian patient." J Virol **73**(9): 7368-75.

Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. & Ho, D. D. (1996). "HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time." Science **271**(5255): 1582-6.

Persaud, D., Zhou, Y., Siliciano, J. M. & Siliciano, R. F. (2003). "Latency in human immunodeficiency virus type 1 infection: no easy answers." J Virol **77**(3): 1659-65.

Poignard, P., Saphire, E. O., Parren, P. W. & Burton, D. R. (2001). "gp120: Biologic aspects of structural features." Annu Rev Immunol **19**: 253-74.

Posada, D. & Crandall, K. A. (1998). "MODELTEST: testing the model of DNA substitution." Bioinformatics **14**(9): 817-8.

Ramos, A., Tanuri, A., Schechter, M., Rayfield, M. A., Hu, D. J., Cabral, M. C., Bandea, C. I., Baggs, J. & Pieniazek, D. (1999). "Dual and recombinant infections: an integral part of the HIV-1 epidemic in Brazil." Emerg Infect Dis **5**(1): 65-74.

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., Gao, F., Hahn, B. H., Kalish, M. L., Kuiken, C., Learn, G. H., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Salminen, M., Sharp, P. M., Wolinsky, S. & Korber, B. (2000). "HIV-1 nomenclature proposal." Science **288**(5463): 55-6.

Rodriguez, F., Oliver, J. L., Marin, A. & Medina, J. R. (1990). "The general stochastic model of nucleotide substitution." J Theor Biol **142**(4): 485-501.

Russell, K. L., Carcamo, C., Watts, D. M., Sanchez, J., Gotuzzo, E., Euler, A., Blanco, J. C., Galeano, A., Alava, A., Mullins, J. I., Holmes, K. K. & Carr, J. K. (2000). "Emerging genetic diversity of HIV-1 in South America." Aids **14**(12): 1785-91.

Saitou, N. & Nei, M. (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Mol Biol Evol **4**(4): 406-25.

Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C. R., Learn, G. H., He, X., Huang, X. L. & Mullins, J. I. (1999). "Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection." J Virol **73**(12): 10489-502.

Simon, F., Mauclere, P., Roques, P., Loussert-Ajaka, I., Muller-Trutwin, M. C., Saragosti, S., Georges-Courbot, M. C., Barre-Sinoussi, F. & Brun-Vezinet, F. (1998). "Identification of a new human immunodeficiency virus type 1 distinct from group M and group O." Nat Med **4**(9): 1032-7.

Soares, M. A., De Oliveira, T., Brindeiro, R. M., Diaz, R. S., Sabino, E. C., Brigido, L., Pires, I. L., Morgado, M. G., Dantas, M. C., Barreira, D., Teixeira, P. R., Cassol, S. & Tanuri, A. (2003). "A specific subtype C of human immunodeficiency virus type 1 circulates in Brazil." Aids **17**(1): 11-21.

Swofford, D. L. (2003). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland, Massachusetts, Sinauer Associates.

Takehisa, J., Zekeng, L., Ido, E., Yamaguchi-Kabata, Y., Mboudjeka, I., Harada, Y., Miura, T., Kaptu, L. & Hayami, M. (1999). "Human immunodeficiency virus type 1 intergroup (M/O) recombination in cameroon." J Virol **73**(8): 6810-20.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." Nucleic Acids Res **25**(24): 4876-82.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.

Triques, K., Bourgeois, A., Saragosti, S., Vidal, N., Mpoudi-Ngole, E., Nzilambi, N., Apetrei, C., Ekwalanga, M., Delaporte, E. & Peeters, M. (1999). "High diversity of HIV-1 subtype F strains in Central Africa." Virology **259**(1): 99-109.

Vicente, A. C., Otsuki, K., Silva, N. B., Castilho, M. C., Barros, F. S., Pieniazek, D., Hu, D., Rayfield, M. A., Bretas, G. & Tanuri, A. (2000). "The HIV epidemic in the Amazon Basin is driven by prototypic and recombinant HIV-1 subtypes B and F." J Acquir Immune Defic Syndr **23**(4): 327-31.

Vidal, N., Peeters, M., Mulanga-Kabeya, C., Nzilambi, N., Robertson, D., Ilunga, W., Sema, H., Tshimanga, K., Bongo, B. & Delaporte, E. (2000). "Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa." J Virol **74**(22): 10498-507.

Xia, X. & Xie, Z. (2001). "DAMBE: software package for data analysis in molecular biology and evolution." J Hered **92**(4): 371-3.

Yang, C., Dash, B., Hanna, S. L., Frances, H. S., Nzilambi, N., Colebunders, R. C., St Louis, M., Quinn, T. C., Folks, T. M. & Lal, R. B. (2001). "Predominance of HIV type 1 subtype G among commercial sex workers from Kinshasa, Democratic Republic of Congo." AIDS Res Hum Retroviruses **17**(4): 361-5.

Zhai, Y., Tchieu, J. & Saier, M. H., Jr. (2002). "A web-based Tree View (TV) program for the visualization of phylogenetic trees." J Mol Microbiol Biotechnol **4**(1): 69-70.

Zhu, T., Korber, B. T., Nahmias, A. J., Hooper, E., Sharp, P. M. & Ho, D. D. (1998). "An African HIV-1 sequence from 1959 and implications for the origin of the epidemic." Nature **391**(6667): 594-7.

X. Anexos

EPIDEMIOLOGIA MOLECULAR DO SUB-SUBTIPO F1 DO HIV-1

SUMÁRIO

Os vírus da imunodeficiência humana do tipo 1 e 2 (HIV-1 e HIV-2) foram os primeiros lentivírus humanos identificados. Acredita-se que tenha sido introduzido na população humana através de duas espécies de primatas africanos. Desde sua identificação como agente etiológico da Aids ainda na primeira metade dos anos 80, inúmeros são os estudos relacionados ao seu genoma. A caracterização dos subtipos do HIV-1 nos possibilita entender a dinâmica evolutiva das populações virais, auxiliando nas estratégias de prevenção, bem como nos estudos relacionados à obtenção de uma vacina eficaz. Analisamos, neste estudo, seqüências do HIV-1, sub-subtipo F1. Realizamos o estudo filogenético e o relacionamos à epidemia de HIV/Aids do Brasil e da Romênia.

Palavras chaves: evolução molecular, HIV-1, subtipos de HIV-1.

SUMMARY

The human immunodeficiency viruses types 1 and 2 (HIV-1 and HIV-2) were the first identified human lentiviruses. It is widely believed that these pathogens entered the human population from at least two different African primate species. Since the identification of HIV-1 and HIV-2 as the etiologic agents of AIDS in the early 1980s, both viruses have been intensively characterized by full or partial sequencing of their genomes. We can now understand the evolutionary dynamic of the viral populations given the characterization of the various HIV-1 subtypes helping us to establish strategies of prevention and the achievement of effective vaccine. In this paper, we analyze the F1 sub-subtypes of HIV-1 linking the HIV/AIDS epidemic in Brazil and Romania.

Key words: molecular evolution, HIV-1, HIV-1 subtypes.

INTRODUÇÃO

Os vírus, especialmente os que possuem genoma RNA, são os organismos ideais para estudos de dinâmicas evolucionárias. Sua alta taxa de substituição nucleotídica faz com que o processo epidemiológico modelador de sua diversidade ocorra na mesma escala de tempo em que as mutações são adicionadas às novas populações virais (Holmes, 2004). A diversidade genética do vírus da imunodeficiência humana (HIV) possui como causas principais exatamente a alta taxa replicativa e a baixa fidelidade da transcriptase reversa (TR) (Crandall, 1999).

O HIV possui duas categorias principais baseadas na sua distribuição geográfica e na fonte animal de infecção humana: chimpanzé (Pan troglodytes) para HIV-1 (Gao *et al.*, 1999) e sooty mangabey (Cercocebus atys) para HIV-2 (Hirsch *et al.*, 1989). Para o HIV-1 foram descritos três grupos distantemente relacionados: o grupo M (principal) (Robertson *et al.*, 2000), o grupo N (para os não M e não O) (Simon *et al.*, 1998) e o grupo O (externo) (Gurtler *et al.*, 1994). O grupo M apresenta linhagens distintas que foram designadas de subtipos e sub-subtipos (figura 1) e formas recombinantes (Carr *et al.*, 2001; Guimaraes *et al.*, 2002).

A introdução de uma ou poucas variantes em áreas geográficas diferentes é responsável pelo "efeito fundador" de cada subtipo nesses locais: no Brasil, podemos observar o efeito fundador dos subtipos B, F e C. Em 1995, Bandea e colaboradores discutiram as relações entre as epidemias do sub-subtipo F1 (doravante designado subtipo F) no Brasil e na Romênia. Utilizaram seqüências do gene do envelope, região C2-V3, de amostras obtidas em crianças da Romênia, seqüências representativas do Brasil, inclusive com a primeira F identificada no país, e da República dos Camarões. A análise filogenética mostrou que as seqüências brasileiras e romenas pertenciam a um grupamento muito próximo, apesar de distintos. As distâncias dentro do subtipo F brasileiro variavam de 5% a 13,8%, o que se encontrava dentro dos limites estabelecidos de dispersão intra-subtipo, entre as da Romênia, 0,9% a 6,5%, também dentro dos limites. As medidas de dispersões entre os dois grupos, 7,5% a 12,9%, evidenciavam ser o mesmo subtipo F circulando nos dois países (Bandea *et al.*, 1995).

Face à posição geográfica do Brasil e da Romênia e de seu intercâmbio cultural e comercial relativamente pequeno, torna-se intrigante a ocorrência e alta prevalência do

mesmo subtipo em áreas geográficas tão díspares. Nosso estudo visa reproduzir, à luz de ferramentas mais aprimoradas e de um maior número de seqüências da alça V3 do HIV-1, o mapeamento epidemiológico molecular do subtipo F.

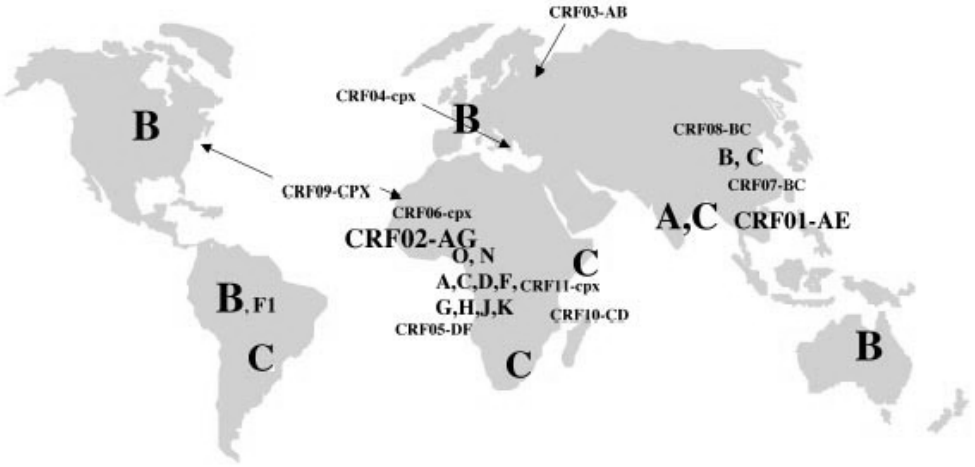

Figura 1 – distribuição mundial dos subtipos de HIV-1 (Peeters & Delaporte, 1999).

Organização Estrutural do HIV

Três genes estruturais e seis reguladores codificam 15 proteínas virais cruciais ao entendimento do ciclo de replicação do HIV e suas relações com a patogenia da Aids. Os estudos moleculares, através da construção de árvores evolutivas, buscam entender desde a origem do vírus ao estudo de transmissão em casos particulares, passando pelo estudo dos compartimentos reservatórios (Holder & Lewis, 2003). A filogenia é uma ferramenta imprescindível para se melhor compreender a relação histórica entre os genes através do tempo e espera-se que a inferência filogenética possa elucidar processos relacionados à dinâmica das seqüências populacionais nelas baseadas (Crandall & Templeton, 1999). A rápida evolução do HIV, se por um lado dificulta o desenvolvimento de estratégias terapêuticas e profiláticas, demonstra ser ideal como objeto para a aplicação dos métodos filogenéticos, permitindo-nos verificar alterações ao longo de anos ou de poucos meses (Hillis, 1999).

Os retrovírus possuem três grupos de genes estruturais (figura 2): gag (antígeno grupo-específico), pol (polimerase) e env (envelope). O gene gag codifica a proteína precursora assemblina (p55) que será clivada pela protease para dar origem às proteínas da matriz (p17), do capsídeo (p24) e do nucleocapsídeo (p7/p9). O gene pol contém as instruções para a produção das enzimas (1) transcriptase reversa, (2) protease, responsável pela clivagem proteolítica dos produtos dos genes gag e pol sem a qual as partículas virais produzidas são desprovidas de infectividade (Kaplan et al., 1993; Kohl et al., 1988) e (3) a integrase, fundamental à integração do provírus ao genoma da célula hospedeira (Barre-Sinoussi, 1996; Fields *et al.*, 2001). O gene env traz as informações necessárias à produção da proteína precursora gp160, que será processada por uma protease celular para dar origem às duas proteínas do envelope: gp120 e gp4l. A gp120 é altamente glicosilada e compõe-se de cinco regiões constantes (C1 a C5) interpostas com cinco regiões variáveis (V1 a V5).

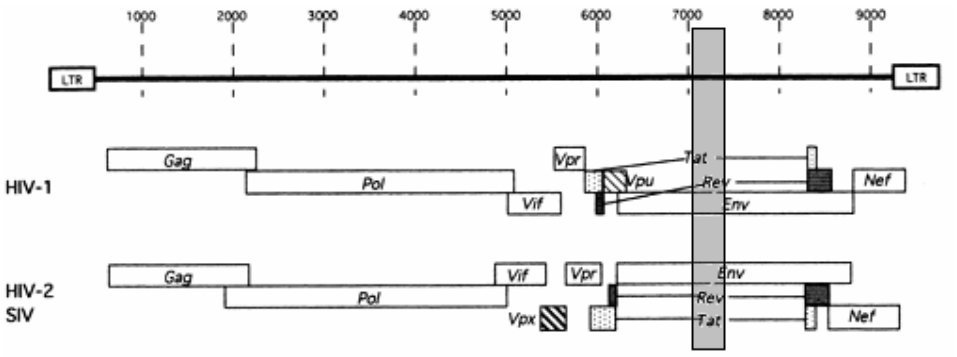

Figura 2 - Organização genômica dos principais retrovírus de primatas – Em destaque a região do envelope, utilizada neste estudo (modificado de Crandall, 1999).

Métodos

Trabalhamos com seqüências obtidas da base de dados do Laboratório Nacional de Los Alamos (LANL), Estados Unidos, através de seu sítio eletrônico e ferramenta de busca em linha (http://hiv-web.lanl.gov/). Solicitamos todas as seqüências de HIV-1 que correspondessem aos seguintes parâmetros: sub-subtipo F1, região genômica V3, de qualquer região geográfica. As seqüências obtidas foram analisadas com o programa DAMBE (Xia & Xie, 2001) para a verificação das repetições. O alinhamento múltiplo foi realizado com o programa ClustalX versão 1.81(Thompson *et al.*, 1997; Chenna *et al.*, 2003). As seqüências alinhadas foram cortadas no tamanho da seqüência BR7494 e formatadas para uso no PAUP* - Phylogenetic Analysis Using Parsimony (*and Other Methods), versão 4.0b10 (Swofford, 2003), das bases 7078 a 7400 (referente a seqüência HXB2), correspondentes à alça C2-V3 do gene env. As inferências filogenéticas foram feitas pelo método de aproximação de vizinhos (NJ) (Saitou & Nei, 1987) usando como modelo de evolução o "modelo geral de tempos reversíveis" com distribuição gama (GTR+G) (Lanave *et al.*, 1984; Rodriguez *et al.*, 1990), escolhido através da análise das seqüências pelo Modeltest 3.06 (Posada & Crandall, 1998). Os parâmetros do modelo sugerido foram: freqüência A = 0,4037; freqüência C = 0,1846; freqüência G = 0,1961e freqüência T = 0,2156. Os valores da matriz de classificação R foram: R(a) [A-C] = 1,6989; R(b) [A-G] = 3,3751; R(c) [A-T] = 0,8474; R(d) [C-G] = 0,5380; R(e) [C-T] = 3,3751; R(f) [G-T] = 1. A proporção de sítios invariáveis foi de zero e o parâmetro da forma da distribuição gama de sítios heterogêneos variáveis de 0,6568. A estimação estatística da confiabilidade das árvores obtidas foi feita pelo método de bootstrap, o mais utilizado para isto (Hall, 2001). As árvores filogenéticas foram formatadas para leitura no Tree-View versão 1.6.0 (Zhai *et al.*, 2002).

As medidas de dispersão entre os pares foram calculadas para os agrupamentos geográficos subdivididos em: brasileiras (BR), africanas (AF), romenas (RO) e demais localidades (OU). O cálculo das distâncias nucleotídicas foi obtido utilizando-se o modelo de Kimura 2-parâmetros (Kimura, 1980), com erro padrão estimado por bootstrap (1000 replicações), implementado no programa MEGA versão 2.1 (Kumar *et al.*, 2001). A média das distâncias inter grupos foram também analisadas e comparadas.

A exceção do PAUP*, todos os programas utilizados são de acesso livre, obtidos facilmente através da internet.

RESULTADOS

Obtivemos 202 seqüências que foram analisadas no DAMBE. Excluímos as repetidas e selecionamos 80 taxa para o estudo. Incluímos ainda antes do alinhamento, a seqüência de referência SIV (acesso AB029818), que usamos como grupo externo, além da primeira seqüência brasileira identificada do subtipo F, BR7944, que serviu também como base para o corte após o alinhamento. Trabalhamos, assim, com 26 taxa brasileiras, 27 taxa romenas, 14 taxa africanas e 13 taxa de outras localidades. Havia, ainda, duas outras seqüências, as AR9515 e PTHDE13, que foram posteriormente retiradas por se tratar a primeira de vírus recombinante B/F e a segunda por não apresentar boa confiabilidade, o que comprometeria a qualidade do estudo. A média geral das distâncias nucleotídicas foi de 0,143, com erro padrão de 0,013. As distâncias inter e intragrupos podem ser vistas nas tabelas 2 e 3 (erro padrão estimado por bootstrap com 1000 replicações e 97396 reproduções aleatórias).

Acesso	Nome	País	Ano	Tamanho
U68522	AR9516	AR	1993	258
U37043	AR9515	AR	1995	870
U37033	AR9518	AR	1995	864
X96527	BEVI850	BE		858
AF034031	BR97127	BR	1997	345
AF113560	BR112	BR		305
AF005494	BR020	BR	1993	Completo
AF113562	BR41	BR		305
AF113564	BR46	BR		318
AF113566	BR57	BR		318
AF113567	BR58	BR		321
AF113568	BR59	BR		321
L19237	BR7944	BR		282
AF034003	BR96026	BR	1996	345
AF034006	BR96034	BR	1996	345
AF034007	BR96035	BR	1996	345
AF034008	BR96036	BR	1996	345
AF034009	BR96039	BR	1996	345
AF113575	BR97	BR		305
AF076314	BRAM08	BR		321
AF076315	BRAM09	BR		345
AF076317	BRAM14	BR		345
AF076321	BRAM28	BR		303
Y18756	BRBA73	BR		345
Y18758	BRBA94	BR		345
U31588	BRSP209	BR		367
U31593	BRSP238	BR		313
U31594	BRSP255	BR		342
AF076320	BRAM27	BR		345
AF153457	BRVTRJ07	BR		345
AF260444	CD85241	CD	1985	460
AF260445	CD85244	CD	1985	481
AF260447	CD85260	CD	1985	485
AJ404119	CD97136	CD	1997	552
AJ404128	CD97165	CD	1997	576
AJ404159	CD97183	CD	1997	549
AJ404073	CD9735	CD	1997	567
AJ404075	CD97KP40	CD	1997	549
AJ404107	CD97KS50	CD	1997	543
AJ404146	CD97KTB50	CD	1997	549
AF224009	CZ9869377	CZ	1998	327
AF224010	CZ9869411	CZ	1998	330
AJ272677	SN978FANN	SN	1997	581
AF219369	FI929213	FI	1992	486
AF075703	FI9363	FI	1993	Completo
Z95465	FR9585	FR	1995	519
AJ249238	FRMP411	FR	1996	Completo
AF425562	CU9978	CU	1999	309
U67709	MQ1403	MQ		325
U67710	MQ1406	MQ		325
U67707	MQ1407	MQ		325
U67706	MQ1409	MQ		325
AJ296264	PT103	PT		508
AJ296250	PT375	PT		508
AY161873	PTHDE13	PT		407
Z83293	RO94BCI2	RO	1994	483
Z83297	RO94BCI3	RO	1994	483
Z83298	RO94BCI4	RO	1994	486
Z83299	RO94BCI5	RO	1994	486
Z83300	RO94BCI6	RO	1994	483
L19571	RO14018	RO		336
Z83285	RO96BCI11	RO	1996	492
Z83286	RO96BCI12	RO	1996	486
Z83288	RO96BCI15	RO	1996	486
Z83289	RO96BCI16	RO	1996	483
Z83290	RO96BCI17	RO	1996	492
Z83292	RO96BCI19	RO	1996	489
Z83294	RO96BCI20	RO	1996	483
Z83295	RO96BCI21	RO	1996	492
Z83296	RO96BCI22	RO	1996	501
Z83301	RO96BCI7	RO	1996	507
Z83302	RO96BCI8	RO	1996	483
Z83284	ROBCI1	RO	1994	486
L19572	ROENVRMB	RO		336
L19573	ROENVRMC	RO		336
L19574	ROENVRMD	RO		336
L19575	ROENVRME	RO		335
L19576	ROENVRMF	RO		335
L19577	ROENVRMG	RO		336
L19579	ROENVRMJ	RO		336
AF003038	SIV	NL	1998	1104

Tabela 1 - Acessos e nomes das seqüências obtidas da base de dados do Laboratório Nacional de Los Alamos (LANL), Estados Unidos.

Grupo	d	S.E.
BR	0,099	0,010
OU	0,124	0,016
RO	0,104	0,011
AF	0,129	0,014

Tabela 2 - Distâncias intra subtipos estimadas pelo método de Kimura dois parâmetros (SE = erro padrão estimado pelo método de bootstrap com 1000 replicações)

	BR	OU	RO	AF
BR	*	[0,013]	[0,017]	[0,014]
OU	0,125	*	[0,016]	[0,014]
RO	0,139	0,140	*	[0,014]
AF	0,134	0,134	0,130	*

Tabela 3 - Distância entre a média dos subtipos estimada pelo método de Kimura dois parâmetros. Erro padrão entre colchetes estimado pelo método de bootstrap (1000 replicações).

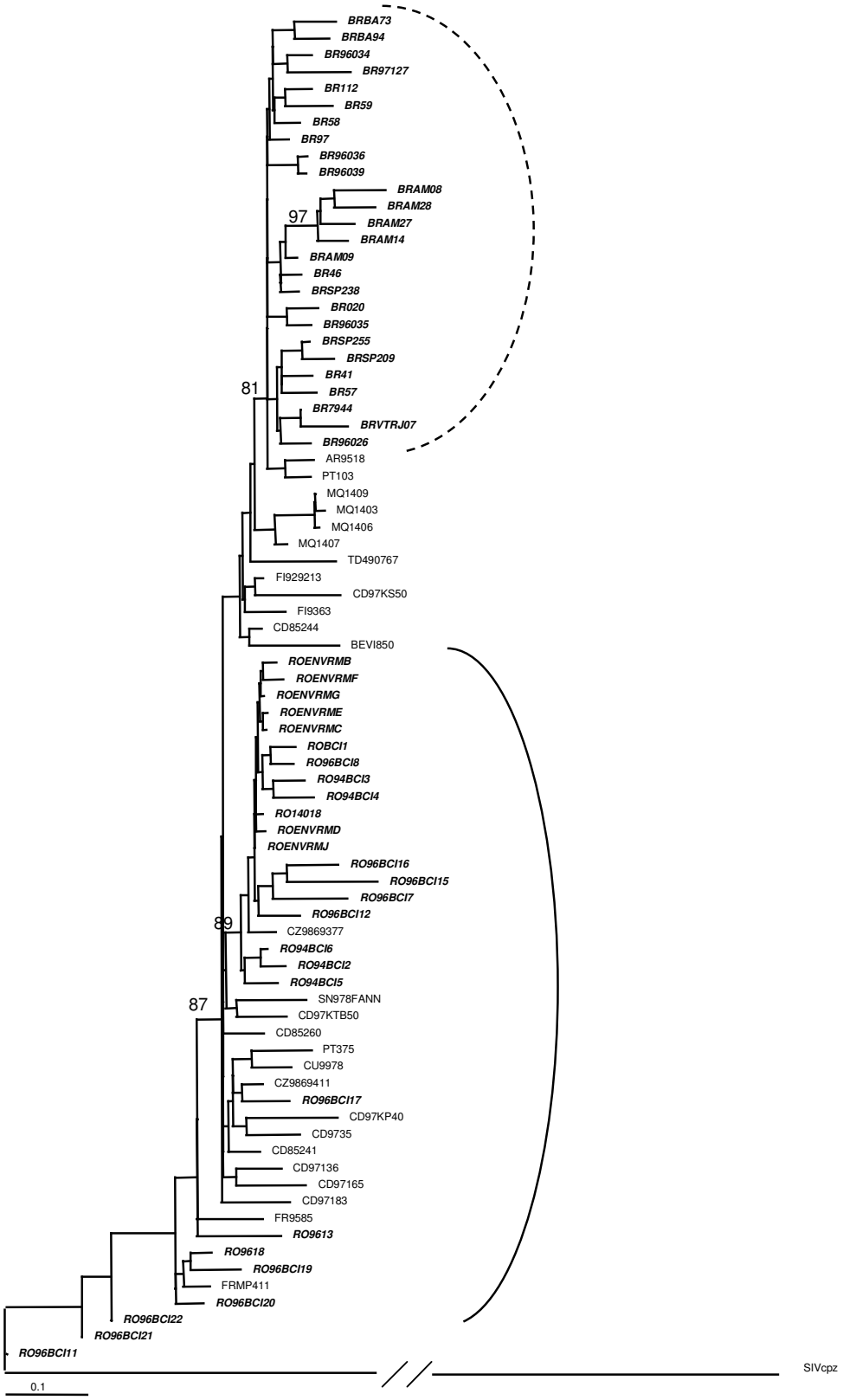

Figura 3 – Filograma obtido pelo método de aproximação de vizinhos (NJ). As seqüências brasileiras e romenas encontram-se destacadas. Valores de bootstrap aplicados aos principais ramos.

DISCUSSÃO

O reconhecimento de tipos e subtipos de HIV nos permite entender o crescimento da epidemia nas várias regiões geográficas e estabelecer estratégias específicas de prevenção.

No continente africano, onde se acredita ser o início da epidemia de HIV/Aids, já foram encontrados todos os subtipos de HIV-1 e HIV-2. No Brasil, o mais afetado país da América do Sul e que mostra ter um mosaico de subepidemias regionais, encontram-se também vários subtipos. O de maior prevalência é o B, mas encontramos também o F, C e os recombinantes B/C e B/F (Morgado *et al.*, 1998; Carr *et al.*, 2001; Guimaraes *et al.*, 2002; Guimaraes *et al.*, 2002; Soares *et al.*, 2003).

A prevalência global de infecções por subtipo F é relativamente baixa. Até 1999 as seqüências representativas desta variante eram divididas em três grupos: F1, F2 e F3, todas demonstráveis em solo africano (Triques *et al.*, 1999). As seqüências recuperadas de pacientes brasileiros, bem como as que ocorrem na Romênia, agrupavam-se com seqüências africanas da variante F1.

A introdução do subtipo F no Brasil ocorreu provavelmente em época posterior à introdução do subtipo B em razão da diversidade genética menos ampla destas seqüências. Este subtipo ganhou importância em crescimento e é a segunda variante mais comum, respondendo por boa parcela das infecções entre usuários de drogas endovenosas da cidade de São Paulo (Bongertz *et al.*, 2000) além de ser encontrado em cerca da metade dos casos de Manaus (Vicente *et al.*, 2000).

Casos isolados de seqüências representativas dos subtipos D e A também foram recentemente descritas no país, mas sempre que eram apropriadamente estudados, demonstravam ser vírus recombinantes ou infecções simultâneas por mais de um subtipo (Couto-Fernandez *et al.*, 1999; Ramos *et al.*, 1999).

O subtipo F foi também descrito entre 1989-1990 na Romênia, em crianças que viviam em orfanatos (Hersh *et al.*, 1991). Os estudos epidemiológicos demonstraram que a maioria das crianças se infectou por transmissão horizontal, ou por transfusão sangüínea ou por material médico-cirúrgico infectado. As taxas de dispersão nucleotídicas entre os

indivíduos estavam entre 0,9% e 3,6% (Dumitrescu *et al.*, 1994). Com a adição de um número maior de seqüências, estas taxas variam entre 9,3% e 11,5% (tabela 2), mostrando serem geneticamente relacionadas, porem com um maior período de evolução. Estes dados sugerem uma única introdução comum de subtipo F nas seqüências estudas provavelmente com a contaminação de um adulto através de contato sexual em viagem fora da Romênia. As seqüências brasileiras se comportam de forma semelhante às romenas, com taxas de dispersão nucleotídicas variando entre 8,7% e 10,7% (tabela 2).

Observando o filograma da figura 3 verificamos que a topologia da árvore mostra os dois grupos de seqüências em ramos separados, com um ancestral comum relativamente distante. Se houvesse uma relação epidemiológica direta entre ambos, os agrupamentos sairiam integralmente um do outro. Podemos afirmar, deste modo, que existe uma relação evolucionária entre os subtipos F brasileiros e romenos, ainda que de modo limitado.

Tomando como base as taxas de dispersões nucleotídicas entre os grupos estudados, tanto o grupo brasileiro quanto o romeno possuem maior proximidade ao africano do que entre si, sugerindo introduções independentes de um ancestral africano comum.

CONCLUSÕES

O conhecimento da evolução genética do HIV é importante não somente para o entendimento dos mecanismos evolucionários básicos como também para a orientação de estratégias de controle e erradicação da Aids. Ainda que países distantes geográfica e culturalmente, o Brasil e a Romênia mantêm relações epidemiológicas evolutivas entre os HIV-1 do subtipo F. A análise filogenética das seqüências recuperadas da base de dados do Laboratório Nacional de Los Alamos (LANL) revela que apesar de se agruparem em ramos diferentes, as seqüências brasileiras e romenas possuem um ancestral comum introduzidos num passado recente. Deste modo, a epidemia de HIV/Aids evolui, nestes países, de forma distinta, apesar de possuírem um ancestral comum.

BIBLIOGRAFIA

Bandea, C. I., Ramos, A., Pieniazek, D., Pascu, R., Tanuri, A., Schochetman, G. & Rayfield, M. A. (1995). "Epidemiologic and evolutionary relationships between Romanian and Brazilian HIV-subtype F strains." Emerg Infect Dis **1**(3): 91-3.

Barre-Sinoussi, F. (1996). "HIV as the cause of AIDS." Lancet **348**(9019): 31-5.

Bongertz, V., Bou-Habib, D. C., Brigido, L. F., Caseiro, M., Chequer, P. J., Couto-Fernandez, J. C., Ferreira, P. C., Galvao-Castro, B., Greco, D., Guimaraes, M. L., Linhares de Carvalho, M. I., Morgado, M. G., Oliveira, C. A., Osmanov, S., Ramos, C. A., Rossini, M., Sabino, E., Tanuri, A. & Ueda, M. (2000). "HIV-1 diversity in Brazil: genetic, biologic, and immunologic characterization of HIV-1 strains in three potential HIV vaccine evaluation sites. Brazilian Network for HIV Isolation and Characterization." J Acquir Immune Defic Syndr **23**(2): 184-93.

Carr, J. K., Avila, M., Gomez Carrillo, M., Salomon, H., Hierholzer, J., Watanaveeradej, V., Pando, M. A., Negrete, M., Russell, K. L., Sanchez, J., Birx, D. L., Andrade, R., Vinoles, J. & McCutchan, F. E. (2001). "Diverse BF recombinants have spread widely since the introduction of HIV-1 into South America." Aids **15**(15): F41-7.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. & Thompson, J. D. (2003). "Multiple sequence alignment with the Clustal series of programs." Nucleic Acids Res **31**(13): 3497-500.

Couto-Fernandez, J. C., Morgado, M. G., Bongertz, V., Tanuri, A., Andrade, T., Brites, C. & Galvao-Castro, B. (1999). "HIV-1 subtyping in Salvador, Bahia, Brazil: a city with African sociodemographic characteristics." J Acquir Immune Defic Syndr **22**(3): 288-93.

Crandall, K. A. (1999). The evolution of HIV. Baltimore, MD, Johns Hopkins University Press.

Crandall, K. A. & Templeton, A. R. (1999). Statistical Approaches to Detecting Recombination. The evolution of HIV. K. A. Crandall. Baltimore, MD, Johns Hopkins University Press**:** 153 - 176.

Dumitrescu, O., Kalish, M. L., Kliks, S. C., Bandea, C. I. & Levy, J. A. (1994). "Characterization of human immunodeficiency virus type 1 isolates from children in Romania: identification of a new envelope subtype." J Infect Dis **169**(2): 281-8.

Fields, B. N., Knipe, D. M., Howley, P. M. & Griffin, D. E. (2001). Fields' virology. Philadelphia, Lippincott Williams & Wilkins**:** 2 v. (xix, 3087, 72 p.).

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M. & Hahn, B. H. (1999). "Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes." Nature **397**(6718): 436-41.

Guimaraes, M. L., dos Santos Moreira, A., Loureiro, R., Galvao-Castro, B. & Morgado, M. G. (2002). "High frequency of recombinant genomes in HIV type 1 samples from Brazilian southeastern and southern regions." AIDS Res Hum Retroviruses **18**(17): 1261-9.

Guimaraes, M. L., Moreira, A. S. & Morgado, M. G. (2002). "Polymorphism of the Human Immunodeficiency Virus Type 1 in Brazil: genetic characterization of the nef gene and implications for vaccine design." Mem Inst Oswaldo Cruz **97**(4): 523-6.

Gurtler, L. G., Hauser, P. H., Eberle, J., von Brunn, A., Knapp, S., Zekeng, L., Tsague, J. M. & Kaptue, L. (1994). "A new subtype of human immunodeficiency virus type 1 (MVP-5180) from Cameroon." J Virol **68**(3): 1581-5.

Hall, B. G. (2001). Phylogenetic trees made easy: a how-to manual for molecular biologists. Sunderland, Mass., Sinauer Associates.

Hersh, B. S., Popovici, F., Apetrei, R. C., Zolotusca, L., Beldescu, N., Calomfirescu, A., Jezek, Z., Oxtoby, M. J., Gromyko, A. & Heymann, D. L. (1991). "Acquired immunodeficiency syndrome in Romania." Lancet **338**(8768): 645-9.

Hillis, D. M. (1999). Phylogenetics and the Study of HIV. The evolution of HIV. K. A. Crandall. Baltimore, MD, Johns Hopkins University Press**:** 105 - 121.

Hirsch, V. M., Olmsted, R. A., Murphey-Corb, M., Purcell, R. H. & Johnson, P. R. (1989). "An African primate lentivirus (SIVsm) closely related to HIV-2." Nature **339**(6223): 389-92.

Holder, M. & Lewis, P. O. (2003). "Phylogeny estimation: traditional and Bayesian approaches." Nat Rev Genet **4**(4): 275-84.

Holmes, E. C. (2004). "The phylogeography of human viruses." Mol Ecol **13**(4): 745-56.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." J Mol Evol **16**(2): 111-20.

Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001). "MEGA2: molecular evolutionary genetics analysis software." Bioinformatics **17**(12): 1244-5.

Lanave, C., Preparata, G., Saccone, C. & Serio, G. (1984). "A new method for calculating evolutionary substitution rates." J Mol Evol **20**(1): 86-93.

Morgado, M. G., Guimaraes, M. L., Neves Junior, I., dos Santos, V. G., Linhares-de-Carvalho, M. I., Castello-Branco, L. R., Bastos, F. I., Castilho, E. A., Galvao-Castro, B. & Bongertz, V. (1998). "Molecular epidemiology of HIV in Brazil: polymorphism of the antigenically distinct HIV-1 B subtype strains. The Hospital Evandro Chagas AIDS Clinical Research Group." Mem Inst Oswaldo Cruz **93**(3): 383-6.

Peeters, M. & Delaporte, E. (1999). "[Genetic diversity of HIV infection worldwide and its consequences]." Med Trop (Mars) **59**(4 Pt 2): 449-55.

Posada, D. & Crandall, K. A. (1998). "MODELTEST: testing the model of DNA substitution." Bioinformatics **14**(9): 817-8.

Ramos, A., Tanuri, A., Schechter, M., Rayfield, M. A., Hu, D. J., Cabral, M. C., Bandea, C. I., Baggs, J. & Pieniazek, D. (1999). "Dual and recombinant infections: an integral part of the HIV-1 epidemic in Brazil." Emerg Infect Dis **5**(1): 65-74.

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., Gao, F., Hahn, B. H., Kalish, M. L., Kuiken, C., Learn, G. H., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Salminen, M., Sharp, P. M., Wolinsky, S. & Korber, B. (2000). "HIV-1 nomenclature proposal." Science **288**(5463): 55-6.

Rodriguez, F., Oliver, J. L., Marin, A. & Medina, J. R. (1990). "The general stochastic model of nucleotide substitution." J Theor Biol **142**(4): 485-501.

Saitou, N. & Nei, M. (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Mol Biol Evol **4**(4): 406-25.

Simon, F., Mauclere, P., Roques, P., Loussert-Ajaka, I., Muller-Trutwin, M. C., Saragosti, S., Georges-Courbot, M. C., Barre-Sinoussi, F. & Brun-Vezinet, F. (1998). "Identification of a new human immunodeficiency virus type 1 distinct from group M and group O." Nat Med **4**(9): 1032-7.

Soares, M. A., De Oliveira, T., Brindeiro, R. M., Diaz, R. S., Sabino, E. C., Brigido, L., Pires, I. L., Morgado, M. G., Dantas, M. C., Barreira, D., Teixeira, P. R., Cassol, S. & Tanuri, A. (2003). "A specific subtype C of human immunodeficiency virus type 1 circulates in Brazil." Aids **17**(1): 11-21.

Swofford, D. L. (2003). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland, Massachusetts, Sinauer Associates.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." Nucleic Acids Res **25**(24): 4876-82.

Triques, K., Bourgeois, A., Saragosti, S., Vidal, N., Mpoudi-Ngole, E., Nzilambi, N., Apetrei, C., Ekwalanga, M., Delaporte, E. & Peeters, M. (1999). "High diversity of HIV-1 subtype F strains in Central Africa." Virology **259**(1): 99-109.

Vicente, A. C., Otsuki, K., Silva, N. B., Castilho, M. C., Barros, F. S., Pieniazek, D., Hu, D., Rayfield, M. A., Bretas, G. & Tanuri, A. (2000). "The HIV epidemic in the Amazon Basin is driven by prototypic and recombinant HIV-1 subtypes B and F." J Acquir Immune Defic Syndr **23**(4): 327-31.

Xia, X. & Xie, Z. (2001). "DAMBE: software package for data analysis in molecular biology and evolution." J Hered **92**(4): 371-3.

Zhai, Y., Tchieu, J. & Saier, M. H., Jr. (2002). "A web-based Tree View (TV) program for the visualization of phylogenetic trees." J Mol Microbiol Biotechnol **4**(1): 69-70.