Luis Henrique Teixeira Caetano

Um Ambiente de Apoio e Análise à Identificação Humana Através do DNA Mitocondrial

Maceió - AL, Brasil

15 de Maio de 2006

Proc: 2004042967-7.
Documento recebido
em: 11 / 08 / 2006
Maria Santana Barbosa
Agente do Funcionário

Luis Henrique Teixeira Caetano

Um Ambiente de Apoio e Análise à Identificação Humana Através do DNA Mitocondrial

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo curso de Mestrado em Modelagem Computacional de Conhecimento do Instituto de Computação da Universidade Federal de Alagoas.

Orientadora Prof. Dra. Eliana Silva de Almeida

Co-orientador:
Prof. Dr. Luiz Antônio Ferreira da Silva

Instituto de Computação
Universidade Federal de Alagoas

Maceió - AL, Brasil

15 de Maio de 2006

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária Responsável: Helena Cristina Pimentel do Vale

C128u Caetano, Luis Henrique Teixeira.
Um ambiente de apoio e análise à identificação humana através do DNA Mitocondrial / Luis Henrique Teixeira Caetano. – Maceió, 2006.
144f. : il.

Orientadora: Eliana Silva de Almeida.
Co-Orientador: Luiz Antônio Ferreira da Silva.
Dissertação (mestrado em Modelagem Computacional de Conhecimento) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2006.

Bibliografia: f. 135-140.

1. Bioinformática. 2. DNA – Alinhamento de seqüência. 3. Sistema de recuperação da informação. 4. Banco de dados. 5. Banco de dados populacional. 6. DNAmt – Perfil. 7. Genética forense. 8. DNA forense. 9. Identificação humana. I. Título.

CDU: 004.78:575.113.1

Dissertação apresentada, como requisito parcial, para a obtenção do grau de Mestre pelo curso de Mestrado em Modelagem Computacional de Conhecimento, do Instituto de Computação da Universidade Federal de Alagoas, aprovada pela comissão examinadora que abaixo assina:

Prof. Dra. Eliana Silva de Almeida
UFAL - Instituto de Computação
Orientadora

Prof. Dr. Luiz Antônio Ferreira da Silva
UFAL - Instituto de Ciências
Biológicas e da Saúde
Co-orientador

Prof. Dr. Luiz Marcos Garcia Gonçalves
UFRN - Dep. de Engenharia de
Computação e Automoção
Examinador

Prof. Dr. Evandro de Barros Costa
UFAL - Instituto de Computação
Examinador

Resumo

A análise do DNA Mitocondrial já provou ser uma poderosa ferramenta de exclusão nos estudos de casos forenses. Ela está sendo utilizada, amplamente, para caracterizar espécimens biológicas forenses, particularmente quando há insuficiente DNA Nuclear em amostras para tipagem. Cabelos, ossos, dentes, dentre outras amostras biológicas que estejam severamente degradadas e ou em estado de decomposição, poderão ser submetidas à análise do seu DNA Mitocondrial.

Cientistas forenses analisam as variações genéticas das regiões HVI e HVII do DNA Mitocondrial para ajudar na identificação de pessoas desaparecidas e em casos criminais. Banco de dados populacionais estão sendo gerados e utilizados para estimar e determinar a raridade de perfis de DNA Mitocondrial obtidos em casos forenses.

Os laboratórios de DNA forense têm grande dificuldade em adotar *softwares* específicos, que se adaptem a sua rotina de trabalho, para dar suporte à gerência da grande quantidade de dados que são obtidos das análises de seqüências do DNA Mitocondrial. Conseqüentemente, esta carência tem provocado erros na produção de perfis de DNA Mitocondrial para serem analisados e utilizados em casos forenses.

A presente dissertação descreve a implementação de um sistema de Bioinformática que vem contribuir para o processo de automatização das análises, do armazenamento e da comparação dos dados de perfis de DNA Mitocondrial para a identificação humana.

A ferramenta proposta é desenvolvida com base nos padrões descritos pela literatura, considerando a consistência no alinhamento de seqüências do DNA Mitocondrial, gerando o haplótipo e anotando seus polimorfismos com o uso da nomenclatura apropriada, como também facilita a inspeção e validação de polimorfismos devido aos erros que poderão estar presentes na sua seqüência de DNA.

A análise dos polimorfismos é uma ferramenta de grande valia na prevenção de erros que possam ser gerados durante todo o processo de manipulação, extração, amplificação, seqüenciamento e no alinhamento de seqüências, do DNA Mitocondrial. Isto nos habilita a criar, armazenar e estimar a freqüência de perfis de DNA Mitocondrial, em banco de dados forenses e populacionais, com qualidade e eficácia.

O sistema de Bioinformática proposto é desenvolvido para ser utilizado no ambiente da *Web* contribuindo para a integração, padronização e a comunicação entre os laboratórios nacionais de DNA forense. Permitindo, assim, o compartilhamento de perfis de DNA Mitocondrial a serem comparados entre os diversos estados do país, usufruindo de um banco de dados comum.

Abstract

The Mitochondrial DNA analysis has proven to be a powerful exclusionary tool in forensic casework. It is being used widely to characterize forensic biological specimens, particularly when there is insufficient nuclear DNA in samples for typing. Hair shafts, bones, teeth and other samples that are severely decomposed may be subjected to Mitocondrial DNA analysis.

Forensic scientists analyse the HVI and HVII Mitocondrial DNA genetic variations to help resolve identity in missing persons and criminal cases. Population databases are being generated and used to estimate and determine the rarity of Mitocondrial DNA profiles obtained in forensic cases.

Forensic DNA labs have great difficulties in adopting specific softwares, which adapt to their work routine, for managing the massive amount of sequence data that are generated. Eventually, this lack has caused many errors on the production of Mitocondrial DNA profiles to be analysed and used for forensic purposes.

The present dissertation describes the implementation of a Bioinformatic system that contributes to the process of automated analysis, data storage and comparison of Mitochondrial DNA profiles for human identification.

The development of the proposed tool is based on the patterns described by the literature, taking in account the consistency of aligning Mitochondrial DNA sequences, which generates the haplotype and annotates its polymorphisms by the appropriate nomenclature, and also facilitates the inspection and validation of the polymorphisms due to the errors that can be present on the raw DNA sequence.

The analysis of polymorphisms is a tool of great value to help prevent errors that may be generated during the entire process of manipulation, extraction, amplification, sequencing and aligning Mitochondrial DNA sequences. Hence this enables us to create, store and estimate the frequency of Mitochondrial DNA profiles, in forensic and population databases, with efficiency and high quality.

The Bioinformatic system proposed is a web-based application that contributes to integrate, standardize and enable the communication between the national forensic DNA laboratories, thus allowing them to share Mitochondrial DNA profiles to be compared among the states of Brazil, through a mutual database.

Dedicatória

À memória de minha avó Letícia,
eternamente querida ...

Agradecimentos

Em primeiro lugar, gostaria de agradeçer aos meus pais Lúcia e Luiz Carlos Caetano pela minha concepção, criação e por todos os princípios morais e éticos que me foram transmitidos. Em particular, agradeço ao meu pai pela construção do acesso às inúmeras possibilidades de aprendizagem e experiências de vida que me foram oferecidas. Agradeço também, em especial e com muito amor, à minha mãe que sempre me incentivou e me deu forças para lutar pelos meus objetivos e por tudo que a vida tem de melhor a oferecer.

Ao meu avô José Caetano (o Zé) pela sua alegria, energia, espírito jovem e o seu convívio diário nesta reta final dos trabalhos.

À todos os colegas de trabalho do laboratório de DNA forense da UFAL que sempre me receberam de abraços abertos. Em particular, à Adriana e ao Dalmo pela atenção e por tudo que me transmitiram sobre o estudo do DNA Mitocondrial humano.

Aos meus colegas de mestrado Liliane, Alan Pedro, Agnaldo, Frederico e o Glauber pelo companheirismo e a amizade durante esta jornada de dois anos. Dentre também, aos colegas do Instituto da Computação: o Tenório, Marcinho (Gabarito) e o Leandro (o Bill), pela força e atenção.

Aos Professores Dr. Evandro Costa, Dr. João Soletti e Dr. Henrique Pacca pela dedicação na contribuiçao da minha formação profissional.

À Prof. Dra. Eliana Silva de Almeida, amiga e orientadora, por toda a paciência, disponibilidade, sabedoria e pela confiança depositada em mim para a realização deste trabalho.

À FAPEAL pela subvenção de recursos para a efetivação desta pesquisa.

Por último, gostaria de agradecer ao amigo e Prof. Dr. Luiz Antônio Ferreira da Silva pela idealização deste trabalho, a oportunidade que me foi dada de estar encarregado e responsável pela elaboração e produção deste projeto, pela confiança depositada, a atenção, a paciência e a dedicação na minha formação.

... e viva a comunidade de *Softwares* Livre! Pois sem estas ferramentas a realização desta dissertação seria muito mais difícil.

Sumário

Lista de Figuras

Lista de Tabelas

1 *Introdução*

"Toda grande jornada
começa com um pequeno passo."
Provérbio Chinês

Neste capítulo são apresentados o objetivo e a estrutura deste trabalho ...

1.1 O DNA Forense

A mais moderna metodologia aplicada mundialmente na identificação humana e, talvez, a mais importante contribuição da ciência para o combate ao crime e à impunidade, é a identificação humana pelo estudo do DNA (*deoxyribonucleic acid*). A utilização do DNA é mais conhecida devido à divulgação de casos famosos na mídia como um instrumento para determinação de paternidade. No entanto, o estudo do DNA tem aplicação muito mais ampla, podendo servir à resolução de casos criminais, identificar restos mortais ou pessoas desaparecidas.

O uso da análise do DNA Forense na resolução de casos que requerem a identificação de pessoas está sendo utilizado nos tribunais com sucesso, sendo comparado e até superando a impressão digital (SILVA; PASSOS, 2002) introduzida a mais de um século como prova investigativa. As polícias têm utilizado as evidências de DNA Forense a não mais do que uma década, que emergiu como uma das ferramentas mais poderosas disponíveis para agentes de execução da lei e para a administração de justiça.

A análise do DNA é a próxima geração de identificação humana na ciência de investigação policial e é também considerada como mais uma ferramenta primordial para a segurança pública. Através do DNA pode-se identificar pessoas através de amostras biológicas como, por exemplo, a partir do sangue ou sêmen desidratados, ossos, bulbo capilar, saliva, pele, suor, esfregaços anais, orais ou vaginais, mesmo em quantidades microscópicas. Esta sua abrangência tem redirecionado os estudos nas academias de formação de polícias em vários países.

O estudo do DNA Mitocondrial também passou a fazer parte do arsenal biotecnológico utilizado para a identificação humana[1]. Sua herança é exclusivamente materna: passando da mãe para os filhos, das filhas para os netos e assim sucessivamente. Através desta mais recente metodologia, um indivíduo pode ser identificado a partir da comparação do seu DNA Mitocondrial com aqueles de seus parentes genéticos maternos. A análise do DNA Mitocondrial é a metodologia de escolha para a identificação humana através de restos humanos antigos, por exemplo: a partir de ossos e dentes que resistiram à sua degradação ao longo dos anos, e também, como em investigações criminais onde, por exemplo, as únicas evidências são pêlos sem bulbo ou quando há somente restos mortais altamente degradados obtidos como evidências biológicas. Essa metodologia é ainda utilizada em

[1]Em 1996 é utilizado pela primeira vez a evidência do DNA Mitocondrial na corte judicial nos EUA. Paul Ware é condenado por estupro e assassinato de uma garota de 4 anos, depois de ter encontrado um *match* do seu perfil de DNA Mitocondrial com a da amostra de um cabelo encontrado no corpo da criança.

estudos antropológicos e evolutivos. Investigações a partir deste tipo de DNA consistem em seqüenciar, alinhar e comparar as regiões hipervariáveis, do genoma mitocondrial, de evidências biológicas e de indivíduos referenciais à luz do conhecimento.

A partir da descoberta das regiões hipervariáveis do DNA, investigações genéticas têm possibilitado a resolução de disputas envolvendo direito de famílias e de nacionalidade. O emprego das inovações surgidas em consequência da evolução do conhecimento científico nessa área estendeu-se a investigações criminais, e hoje, profissionais com conhecimento aprofundados de biologia molecular, genética e estatística (expertos em identificação humana pelo DNA) fazem parte dos quadros técnico-científicos de serviços relacionados com a segurança pública, tornando-se assim altamente importante para os poderes judiciário e executivo, devido ao seu poder de descriminação.

O tipo de herança do DNA Mitocondrial pode ser muito útil em testes de identificação humana porque permite uma comparação direta de seqüências de parentes com a mesma linhagem materna, sem ambiguidade causada por recombinação meiótica e mistura de genes nucleares. De fato, quando a amostra sequenciada é comparada com uma pessoa de referência, a probabilidade de um parentesco materno pode ser calculada.

O conhecimento da freqüência com que certos tipos de DNA Mitocondrial ocorrem numa dada população é de crucial importância para a aplicação de seus marcadores forenses já que os tipos de seqüências de DNA Mitocondrial estão fortemente correlacionadas com origens geográficas, fenótipos ou etnias.

A identificação humana através do DNA Mitocondrial se dá pelo cumprimento de várias etapas, com diferentes processos de análises, realizados a partir de uma amostra biológica em questão. As etapas envolvidas podem ser divididas em duas fases. A primeira se dá pelo trabalho de laboratório, envolvendo processos para a extração e amplificação do DNA, com a manipulação do material da amostra e com produtos bioquímicos. Já a segunda fase se dá pelo uso da computação para dar suporte nos processos de sequenciamento do DNA, alinhamento de seqüências, análise de qualidade, armazenamento e comparação de dados, cálculos estatísticos e probabilísticos etc.

Nesta dissertação estuda-se algumas técnicas para o alinhamento de seqüências de DNA. Este estudo é realizado para a escolha da técnica mais apropriada para o alinhamento de seqüências do DNA Mitocondrial humano em estudos de casos forenses. Veremos que a etapa do alinhamento é uma das mais cruciais e problemáticas, ainda hoje, no estudo do DNA Mitocondrial, bem como também é tratada a forma de anotar o resultado dessa análise para ser armazenado e comparado no banco de dados.

Um dos maiores problemas enfrentados pela identificação humana através do DNA Mitocondrial é a falta de ferramentas específicas para alinhar, gerar o haplótipo[2], verificar erros e, em seguida, armazenar estas informações em bancos de dados para depois serem analisadas, de maneira comparativa, em busca de similaridades. Diante dessa carência a comunidade forense tem tido dificuldades para gerar bancos de dados populacionais pois o seu uso é indispensável em estudos de casos forenses para realizar cálculos estatísticos e probabilísticos. Estes problemas são aqui tratados e uma solução é apresentada com o desenvolvimento de um sistema computacional.

[2]Grupo de alelos (polimorfismos do DNA) situados em regiões muito próximas, sendo transmitidos em conjunto, como uma unidade, para os descendentes.

1.2 Motivação

No cenário da Bioinformática a pesquisa está amplamente voltada para o estudo do genoma humano, especificamente na análise de novos genes. Nela, encontram-se linhas de pesquisa voltadas para desvendar proteínas e suas funções. Estudos voltados para a determinação da estrutura 3D da proteína contribuem para revelar a sua real funcionalidade (GIBAS; JAMBECK, 2001; LESK, 2002; BARNES; GRAY, 2003).

No entanto, existe pouca atenção voltada para o estudo do DNA forense. Geneticistas forenses dependem, muitas vezes, da utilização de ferramentas de Bioinformática que sejam genéricas e voltadas para análises genômicas (BENTON, 1996).

A motivação deste trabalho partiu do laboratório de DNA forense da UFAL. Lá detectou-se a escassez de ferramentas específicas para tratar das análises do DNA Mitocondrial forense, com o objetivo de identificar amostras de pessoas desaparecidas.

A Bioinformática é bastante conhecida por possuir como um dos seus primordiais problemas a gerência da grande quantidade de dados que são gerados a partir dos seus estudos genômicos. A problemática não poderia ser diferente em se tratando do DNA Mitocondrial Forense.

O estudo do DNA Mitocondrial forense necessita, como ferramenta básica, uma base de dados de perfis populacionais, visto que o estudo da identificação de pessoas através do DNA é baseado e constituído em cálculos probabilísticos e estatísticos, quanto maior for o tamanho dessa base mais precisa será a estimava deste cálculo. No entanto, a construção desta base de dados precisa ser cuidadosamente elaborada, desde a etapa da tipagem da amostra em questão (fase de laboratório) até a sua fase de análises da sua seqüência, o qual é realizada através de softwares de Bioinformática.

A prática atual realizada pelos geneticistas forenses para organizar estas informações se resumem a armazenar seus dados em editores de texto e/ou em planilhas. Tal abordagem é bastante deficiente no que diz respeito a precariedade da formatação que estes dados estão sujeitos a receber para serem, eventualmente, comparados entre si. Sem contar que esta formatação poderá variar entre os laboratórios, o que comprometeria seriamente o compartilhamento destes dados entre eles para serem analisados.

É também visível a falta de integração das ferramentas utilizadas para análise do alinhamento de sequências com as ferramentas de armazenamento de dados. Este problema ocasiona o geneticista forense a ter que editar e copiar o resultado do alinhamento de suas sequências, armazenando-os (digitando) em ferramentas inapropriadas, sem nenhum controle do tipo de dado, para então serem comparados manualmente.

Realizar a identificação humana através do DNA Mitocondrial necessita de uma avaliação rigorosa dos passos para gerar, validar, armazenar, comparar e estimar o peso de perfis de DNA Mitocondrial em estudos de casos forenses. Portanto, este trabalho foi realizado com a finalidade de construir o ferramental necessário de análise e armazenamento, em um banco de dados relacional, integrando e automatizando o seu uso para serem utilizadas no ambiente da *Web*.

1.3 Objetivos

O trabalho ora proposto possui os seguintes objetivos:

1. Desenvolver uma ferramenta para o alinhamento específico de seqüências do DNA Mitocondrial.
2. Gerar o haplótipo do perfil de DNA Mitocondrial no seu devido formato de anotação.
3. Elaborar uma ferramenta para a verificação de erros no haplótipo.
4. Construir um banco de dados relacional para comportar e gerenciar as três bases de dados, quais são: de pessoas desaparecidas, reclamantes e de perfis populacionais.
5. Automatizar o processo de armazenamento e comparação de perfis.
6. Estimar o peso de uma evidência ao acaso.

1.4 Estrutura do texto

A dissertação está dividida em seis capítulos. Neste primeiro capítulo é descrito todo o contexto geral do cenário do DNA forense e a utilização do DNA Mitocondrial como instrumento para a identificação humana, em especial, a partir de restos mortais degradados. A justificativa para a realização deste trabalho é descrita através das motivações e, em seguida, são apontados os objetivos a serem atingidos com o ultimar das atividades estabelecidas neste projeto de mestrado.

O segundo capítulo descreve uma introdução ao DNA Mitocondrial focada, basicamente, em suas características de localização e estrutura dentro da célula, informações estas que são utilizadas para o seu estudo em casos forenses. Também são apresentadas, detalhadamente, as técnicas ferramentais utilizadas nas suas análises e as práticas de manipulação dos seus dados.

O terceiro capítulo é especificamente voltado para a dicussão das diversas técnicas utilizadas na comparação de sequências biológicas. Este capítulo é uma extensão do segundo capítulo, na apresentação das diversas problemáticas encontradas nas análises do DNA Mitocondrial, onde é descrito, especificamente, a carência de ferramentas para o alinhamento de suas sequências nucleotídeas, com exemplos de uso.

O quarto capítulo descreve uma solução computacional fundamentada de acordo com a concepção das problemáticas apresentadas nos capítulos anteriores. Neste capítulo é descrito, em detalhes, a modelagem do sistema computacional e de seus módulos que, em conjunto, reunem as ferramentas essencias para o estudo de casos forenses, através da análise do DNA Mitocondrial humano.

No capítulo cinco são descritos os experimentos e os resultados obtidos, como teste, para comprovar a utilidade do ambiente computacional, proposto no capítulo quatro.

Por último, o capítulo seis apresenta as conclusões gerais deste trabalho e as diversas indicações para trabalhos futuros que visam dar continuidade no desenvolvimento da solução computacional, elaborada pelo presente estudo.

2 *Análises do DNA Mitocondrial*

"O DNA é o mensageiro, que ilumina (nossa conexão com o passado), transmitido de geração para geração, carregado, literalmente, nos corpos de (nossos) antepassados. Cada mensagem traça uma jornada através do espaço e do tempo, uma jornada realizada pelas longas linhas que nascem através das mães antecessoras."

Bryan Sykes, *The Seven Daughters of Eve*

Este capítulo descreve as características do DNA Mitocondrial, os passos que envolvem a obtenção de resultados em estudos de casos forenses, e também sobre as importantes temáticas envolvidas na interpretação de seus resultados ...

2.1 Localização, estrutura e características

O genoma humano está localizado dentro do núcleo de cada célula. No entanto, existe um pequeno genoma circular encontrado dentro da mitocôndria[1]. Este genoma, o DNA Mitocondrial (DNAmt), molécula de DNA extra-nuclear presente apenas nas mitocôndrias, pode variar bastante em número dentro da célula. Existem, em média, 4 a 5 cópias de moléculas de DNAmt por mitocôndria, podendo até variar entre um número de 1 a 15 cópias (SATOH; KUROIWA, 1991). Devido a célula poder conter centenas de mitocôndrias, estimado na média de 500 cópias, matematicamente, poderão existir milhares de moléculas de DNAmt em cada célula (SATOH; KUROIWA, 1991)(figura 1).

Por possuir o maior número de moléculas de DNA por célula, o DNAmt obtém um índice de maior sucesso em relação aos marcadores do DNA Nuclear STR[2] na tipagem de amostras biológicas que possam ter sido danificadas com o calor ou com a umidade (Tabela 1). Também vale salientar que a soma de todas as cópias de moléculas de DNAmt geram menos de 1% do conteúdo total de DNA Nuclear (DNAn) que uma célula pode possuir, assumindo que existam 1000 cópias de DNAmt no tamanho de 16569 pares de base (pb) e duas cópias de DNAn no tamanho de 3.2 bilhões de pares de base.

O DNAmt tem aproximadamente 16569pb e possui 37 genes que codificam proteínas utilizadas na produção de energia da célula. O número total de nucleotídeos no genoma mitocondrial pode variar devido a pequena taxa de mutações resultantes de inserções ou deleções. Por exemplo, existe uma repetição de binucleotídeos entre as posições 514 a 524 que na maioria dos indivíduos é de ACACACACAC ou $(AC)_5$, mas esse número pode variar entre $(AC)_3$ a $(AC)_7$. Dos 37 genes transcritos no DNAmt, encontrados na região codificante, 13 incluem proteínas, dois RNAs ribossomais (rRNA) e 22 RNAs transportadores (tRDNA) (figura 2).

A variabilidade de nucleotídeos e polimorfismos entre indivíduos no *displacement loop*[3] é mais abundante do que nas regiões codificantes, pois suas restrições são menores devido a não codificação de genes. Em outras palavras, podem ocorrer diferenças na região D-loop, pois nela não se codifica qualquer substância necessária para a funcionalidade da célula. A maior parte da atenção em estudos de DNA forense envolve o uso de duas regiões hipervariáveis, dentre a região controle, comumente referidas como HV1 e HV2.

[1]Organela celular, produtora de energia, que reside dentro do citoplasma.

[2]STR (*Short Tandem Repeat*), pequenos fragmentos de sequências de DNA (2 à 6 bases) que se repetem. É, nos dias de hoje, o marcador mais utilizado no estudo de casos forenses e testes de paternidade.

[3]Posições não condificantes da região controle, também conhecida por D-loop.

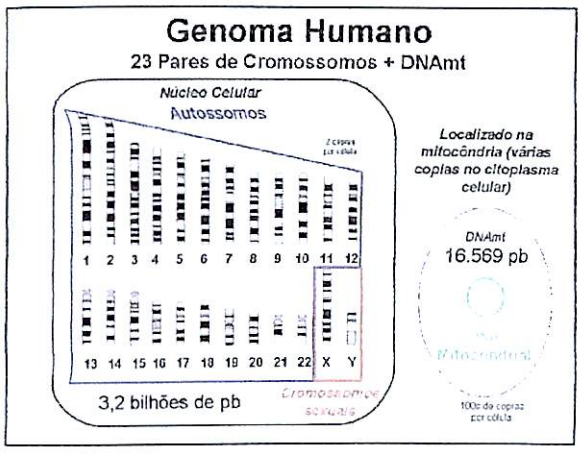


Figura 1: Ilustração dos dois tipos de genoma humano presentes na célula. Fonte: (BUTLER, 2005).

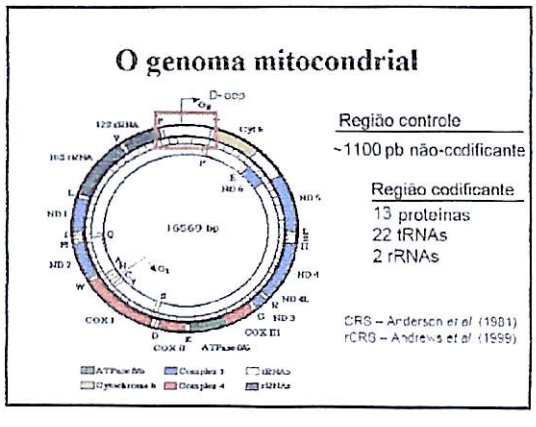


Figura 2: Esquema do genoma circular do DNA Mitocondrial. No topo do genoma, demarcado com o retângulo vermelho, encontra-se a região hipervariável onde os cientistas forenses analisam o DNA para fins de identificação humana. Fonte: (BRANDON et al., 2004).

Características	DNA Nuclear (DNAn)	DNA Mitocondrial (DNAmt)
Tamanho do genoma	~ 3.2 Bilhões pb	~ 16569 pb
Cópias por célula	2 (1 alelo de cada parente)	Pode ser > 1000
% total do conteúdo de DNA por célula	99.75%	0.25%
Estrutura	Linear	Circular
Herança	Pai e mãe	Materna
Recombinação	Sim	Não
Exclusividade	Único por indivíduo (exceto irmão gêmeos)	Não único por indivíduo (mesmo que dos parentes maternos)
Taxa de mutação	Pequena	No mínimo 5-10 vezes mais que o DNAn
Sequência de Referência	Descrito em 2001 pelo Projeto do Genoma Humano	Descrito em 1981 por Anderson e Colaboradores

Tabela 1: Comparações entre o DNA Nuclear e o DNA Mitocondrial.

2.2 Heteroplasmia

A heteroplasmia é a presença de mais de um tipo de DNAmt em um indivíduo, existindo assim a possibilidade de ocorrer duas ou mais populações de DNAmt dentre as células de um indivíduo, como também, dentro de somente uma célula ou dentro de uma mitocôndria (MELTON, 2004).

É improvável que as milhares de moléculas do DNAmt sejam completamente idênticas, dado que as regiões do genoma mitocondrial podem variar de 6 a 17 vezes a mais do que uma única cópia de um gene nuclear. Assim, para que uma transmissão de mutação possa ser detectada no DNAmt, essa mutação deverá se espalhar a uma freqüência considerável dentre as células e as moléculas de mitocôndria. A heteroplasmia poderá se manifestar a partir das seguintes formas (CARRACEDO et al., 2000):

- Um indivíduo poderá ter mais de um tipo de DNAmt em um único tecido;
- Um indivíduo poderá exibir um único tipo de DNAmt em um tecido e um outro tipo de DNAmt em um outro tecido;
- Indivíduos poderão ser heteroplásmicos em um único tecido e homoplásmico num outro tecido.

Dado o fato da ocorrência de heteroplasmia, deve-se então recorrer as recomendações de interpretação para resolver as diferenças entre as amostras de evidência e referência.

Os dois tipos de heteroplasmia validados pela literatura são a heteroplasmia de sequência e a de comprimento (MELTON, 2004). A heteroplasmia de comprimento geralmente ocorre nas regiões do poly-C, em HVI, nas posições 16184 - 16193, e também em HVII, nas posições 303 - 310 (figura 3) (STEWART et al., 2001). A heteroplasmia de sequência é tipicamente observada pela presença de dois ou mais nucleotídeos ocupando o mesmo sítio em diferentes moléculas de DNAmt, podendo ser vista em eletroferogramas onde existam picos sobrepostos (figura 4).

A heteroplasmia poderá também ocorrer em mais de um sítio. Esta condição é conhecida por "triplasmia" e tem sido observada na literatura (TULLY et al., 2000). Mas, a sua ocorrência é raríssima em relação a heteroplasmia de um só sítio.

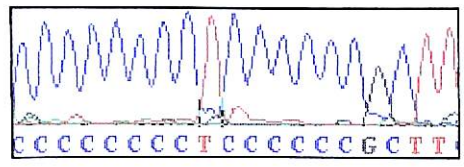


Figura 3: Ilustração da ocorrência de uma heteroplasmia de comprimento, com a observação de um T em meio ao poly-C, em HVII.

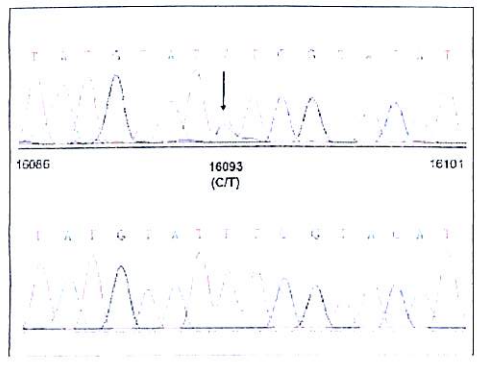


Figura 4: Observação de heteroplasmia na sequência (a) no sítio 16093, possuindo C e T como base, comparando-se à sequência (b) que, no mesmo sítio, possui apenas um T. Fonte: (BUTLER, 2005).

Visto a raridade de encontrar mais de uma ocorrência de heteroplasmia dentre os 610 nucleotídeos sequenciados para HVI e HVII, o relato de Grybowski sobre a ocorrência de seis sítios heteroplásmicos, em um indivíduo, levantou suspeitas sobre a sua estratégia utilizada no sequenciamento. O estudo de Grybowski foi criticado por conter, possivelmente, contaminação devido ao excessivo número de ciclos utilizados na amplificação das amostras(BRANDSTäTTER; PARSONS, 2003). Uma nova análise realizada nestas mesmas amostras, desta vez utilizando-se uma abordagem direta de PCR, obteve como resultado a redução no número de sítios heteroplásmicos.

Um dos maiores problemas em amostras heteroplásmicas é que a frequência das bases podem não permanecer as mesmas dentre os outros tipos de tecidos, tais como o sangue e cabelo ou entre vários cabelos. Alguns protocolos do DNAmt recomendam sequenciar vários cabelos de um indivíduo no esforço de se confirmar a heteroplasmia.

Os *hotspots* de heteroplasmia incluem as seguintes posições em HVI: 16093, 16129, 16153, 16189, 16192, 16293, 16309 e 16337; e para HVII, nas posições: 72, 152, 189, 207 e 279 (TULLY et al., 2000; BRANDSTäTTER; PARSONS, 2003).

Enquanto que o acontecimento de uma heteroplasmia pode, as vezes, dificultar a interpretação final do resultado do DNAmt, a sua presença em sítios idênticos pode aprimorar ainda mais a probabilidade de um *match*, tal como é descrito no estudo do caso da família Romanov (seção 2.3.2).

2.3 Herança materna

Para fins forenses e na identificação de pessoas desaparecidas, como também, em qualquer outra área de pesquisa, o DNAmt humano é considerado estritamente herdado da mãe para seus filhos.

Durante a fecundação, somente o núcleo do espermatozóide consegue penetrar e unir-se diretamente ao núcleo do óvulo. Desta forma, a mitocondria, juntamente com suas moléculas de DNAmt, é passada diretamente para o prole, independente de qualquer influência paterna.

Nos óvulos, tem se observado uma quantidade de 100 mil moléculas de DNAmt, o que acarreta na diluição extrema da passagem de qualquer molécula de DNAmt paterno para o zigoto (CHEN et al., 1995). Portanto, com a exceção de mutações, a mãe passa adiante o seu DNAmt para seus filhos e consequentemente tanto esses descendentes, como os parentes maternos, possuem ou possuirão seus DNAmt idênticos.

Na figura 5 é dado um exemplo de uma árvore genealógica de famílias para demonstrar a herança padrão do DNAmt. Neste exemplo, os indivíduos 1, 5, 7 e 12 possuem exclusivamente um único tipo de DNAmt. Note também que o indivíduo 16 irá possuir o mesmo DNAmt que os outros sete indivíduos, representados pelo número 2, 3, 6, 8, 11, 13 e 15.

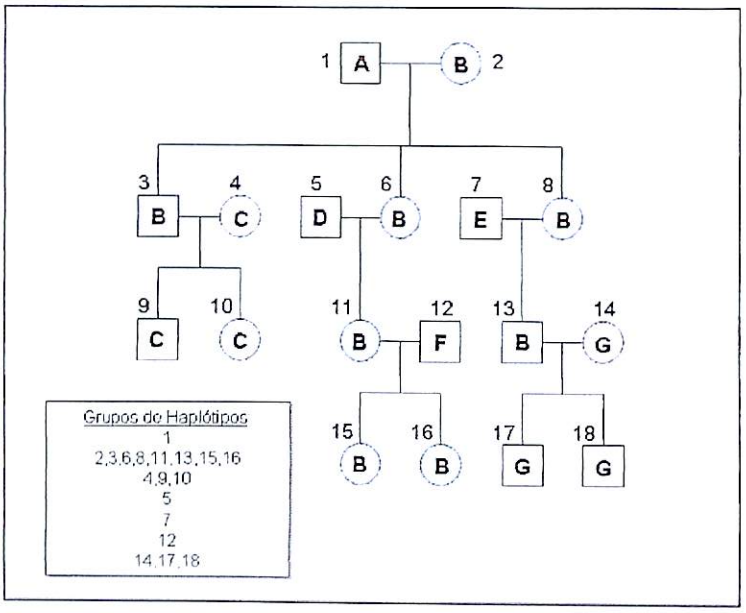


Figura 5: Ilustração da herança materna do DNAmt para 18 indivíduos em uma árvore genealógica hipotética. Os quadrados representam os homens e os círculos as mulheres. Cada tipo único de DNAmt é representado por uma letra alfabética. Fonte: (BUTLER, 2005).

Esta característica pode servir de grande ajuda na resolução de pessoas desaparecidas ou na investigação de desastres em grande massa, por exemplo: a identificação de restos mortais no ataque terrorista ao World Trade Center, no dia 11 de setembro de 2001 em Nova York, EUA. Mas, ao mesmo tempo, esta característica poderá reduzir a significância de um *match*[1] em casos forenses, visto que até mesmo parentes muito distantes podem possuir o mesmo tipo de DNAmt, o que também acarreta no aumento do número de amostras que poderão servir como referência para confirmar a identidade de uma pessoa desaparecida.

Evidências a partir do DNAmt tem ajudado muito na tarefa de associar e/ou ligar famílias, por exemplo: o famoso caso da identificação do desconhecido soldado da guerra do vietnam, como também na resolução de quebra cabeças históricos, tais como, o caso da família Romanov. Os dois casos são descritos nas duas próximas seções.

[1]Palavra de origem inglêsa que tem como significado, para o nosso contexto, a observação de uma igualdade ou semelhança muito próxima na comparação entre duas sequências.

2.3.1 Identificação dos restos mortais da tumba de soldados da guerra do Vietnam

No dia 30 de Junho de 1998, o secretário de defesa americano, William Cohen, anunciou ao mundo que a tecnologia do DNA foi utilizada para identificar um soldado da guerra do Vietnam, da tumba de soldados não identificados localizado no cemitério national de Arlington. Os restos mortais do 1º tenente da força aérea dos EUA, Michael J. Blassie, foram identificados através do uso de seu DNAmt. Uma comparação exata dos 610 nucleotídeos da polimórfica região controle do seu DNAmt foi obtida através dos de sua mãe. Ao mesmo tempo, oito dos demais possíveis soldados foram excluídos devido a um não *match* entre outras referências de famílias a procura de seus parentes desaparecidos na guerra.

O tenente Blassie chegou ao Vietnam em Janeiro de 1972. Seu jato de caça A-37B foi derrubado no dia 11 de maio, de 1972, num região próxima a fronteira da cidade de Cambodgia. A região não pôde ser revistada e seus restos mortais não puderam ser recuperados até cinco meses depois. Por este tempo, foram encontradas apenas algumas ossadas e alguns itens pessoais, incluindo o seu cartão de identificação. Os seus restos mortais foram enviados para o laboratório central de identificação do exército no Hawaii, onde lá permaneceu por oito anos, quando uma revisão militar alterou sua designação para "desconhecido" e seu cartão de identificação, encontrado junto aos restos mortais, havia desaparecido.

Em meados de junho de 1998, conseguiu-se extrair informações sobre a sequência do DNAmt através da ossada de um pélvis da tumba de soldados não identificados, cujo material foi analisado no "laboratório de identificação através do DNA das forças armadas" (AFDIL). Supostos parentes maternos dos oito soldados não identificados, cujos restos mortais foram encontrados naquela região, foram avaliados como amostras para referência. As posições 16021 a 16365 (HVI) e as posições 73 a 340 (HVII), da polimórfica região controle, foram submetidas a análises. Dentre as comparações entre as amostras de referências e os restos mortais, apenas um *match* perfeito foi observado: o que ligou a mãe de blassie ao resto mortal de seu filho (Michael Blassie). Devido a essa identificação positiva, a família Blassie foi autorizada a enterrar os restos mortais do 1º tenente no cemitério nacional de Jefferson Barracks (HOLLAND; PARSONS, 1999).

2.3.2 Identificação dos restos mortais da família Romanov

Em 1918, após a revolução soviética, o czar[5] russo e sua família foram mantidos como prisioneiros do novo regime. Acredita-se que o czar russo Nicolás II, sua mulher, a czarina Alexandra, suas 4 filhas e 1 filho, três empregados e o médico da família foram fuzilados pelos bolcheviques. Parece que algumas das vítimas não morreram imediatamente. Como consequência, os corpos foram golpeados até a comprovação do falecimento de todos. Acredita-se também que, depois de esquartejados, os corpos foram lançados em uma fossa a cerca de 20 Km do local onde a família foi mantida prisioneira. Supostamente, despejou-se ácido sulfúrico sobre os cadáveres para evitar uma posteior identificação e a fossa foi aterrada.

Em 1991, esta fossa foi descoberta e encontrados restos ósseos humanos que correspondiam a 9 cadáveres com traços de tortura e marcas de bala (GILL et al., 1994). Especialistas forenses russos realizaram estudo de reconstrução facial computadorizada, comparação de arcadas dentárias e estimativa de idade e sexo dos indivíduos. A presença de peças dentárias em ouro indicava que alguns dos cadáveres poderiam ter pertencido à aristocracia. Por fim, os estudos constataram que os restos mortais poderiam corresponder ao czar, czarina e a 3 dos 5 filhos. Concluiu-se que faltavam os restos do filho Alexei e de Anastasia, uma das filhas.

Em 1992, um laboratório inglês iniciou estudos através da análise por DNA, visando a identificação fidedigna das partes. Este laboratório adotou dois procedimentos: a análise do STR, para comprovar se os restos mortais se tratavam de um grupo familiar, e análise de DNA Mitocondrial, para determinar relação de parentesco com os descendentes da família Romanov por via materna.

O sexo dos indivíduos foi determinado através da amplificação do gene amelogenina. Após a análise dos 9 esqueletos, as tipagens por DNA confirmaram as conclusões do exame físico, constatando-se que os restos mortais eram correspondentes a 4 homens e 5 mulheres. Mesmo diante de dificuldades decorrentes de artefatos gerados durante as reações de PCR[6], foi concluído que 5 restos ósseos pertenciam a um mesmo grupo familiar. Caso se tratasse da família Romanov, os dados indicavam que uma das princesas e o príncipe não se encontravam na tumba. O fato corroborou a hipótese de que dois corpos teriam sido enterrados ou incinerados separadamente, ou que Anastasia e Alexei teriam sobrevivido.

[5] Título que se dava ao imperador na Rússia, e aos antigos soberanos sérvios e búlgaros.

[6] *Polymerase Chain Reaction*. Método de amplificação (na criação de múltiplas cópias) do DNA, sem o uso de um organismo vivo.

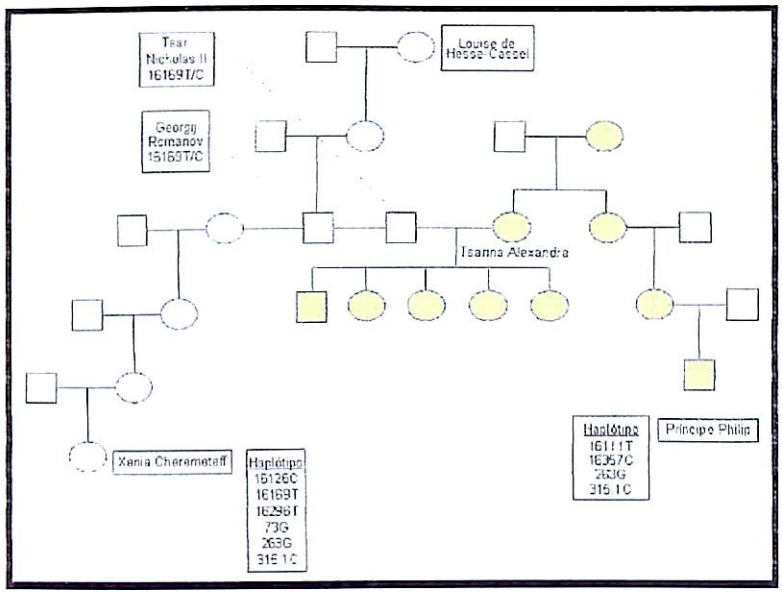


Figura 6: Linhagem da família Romanov. Os indivíduos representados pela cor azul são parentes maternos de Tsar II e os em verde são parentes maternos de Tsarina. Os parentes maternos vivos Príncipe Philip (de Tsarina) e Xenia (de Tsar II) serviram como amostra de referencia, com os polimorfismos de seus haplótipos relativos as posições da rCRS. Fonte: (BUTLER, 2005).

Com respeito ao DNAmt, os 9 esqueletos tiveram sequenciadas as regiões hipervariáveis HVI e HVII. Sequências idênticas foram obtidas para a suposta czarina e os três possíveis filhos. Como referência, foi utilizada a sequência de DNAmt produzida a partir de amostra de sangue do Príncipe Philip, Duque de Edimburgo, sobrinho-neto da czarina. Sua sequência era idêntica àquela da czarina e de seus filhos (figura 6).

Em relação ao czar, a sequência proveniente dos restos mortais foi inicialmente comparada com a de uma neta, de sua irmã, ainda viva. As duas sequências eram idênticas, exceto quanto a uma heteroplasmia na posição 16169 no DNA extraído a partir do esqueleto atribuído ao czar, que indicava as bases citosina e timina. Na sequência da amostra de referência, esta posição era ocupada somente por timina. Os dados apontavam ser a análise inconclusiva quando decidiu-se exumar o corpo do irmão do czar, George Romanov, morto em 1899. Neste, verificou-se no DNAmt, exatamente a mesma sequência e heteroplasmia em relação ao czar, finalizando, então, a identificação (DEBENHAM, 1996).

A partir desde caso, verifica-se que uma heteroplasmia pode reforçar o poder de discriminação da técnica. Como não era possível analisar as sequências da mãe e do avó czar, não se sabe se elas também apresentavam esta heteroplasmia. Provavelmente, o DNAmt contendo timina na posição 16169 segregou para homoplasmia nas gerações seguintes.

O caso Romanov ainda esteve no cenário científico até 1995, quando a sequência de DNAmt de uma mulher chamada Anna Manahan, suposta princesa Anastasia, que teria, então, escapado da chacina, não coincidiu com o perfil já verificado para a czarina (STONEKING et al., 1995).

O caso da família Romanov é um dos mais famosos exemplos da aplicação do DNAmt na identificação humana. Enquanto o emprego da análise de DNAn STR foi útil para provar a relação de parentesco entre os esqueletos, o DNAmt foi utilizado para identificação através da comparação com amostras de indivíduos separados por várias gerações.

2.4 Sequência de referência

O primeiro DNAmt humano foi sequenciado em 1981 no laboratório de Frederick Sanger na Universidade de Cambridge, Inglaterra (ANDERSON et al., 1981). Por muitos anos, a sequência original de Anderson foi utilizada como a sequência de referência para ser comparada com outras novas sequências. A sequência de Anderson é também conhecida como a *Cambridge Reference Sequence* (CRS). Tipicamente, os laboratórios de DNA exibem como resultado as variações resultantes da comparação entre a *L-strand*[7] da CRS. Assim, por exemplo, a observação de um nucleotídeo C na posição 16126 em uma sequência de amostra, na qual é observado um T na sequência de Anderson, é representado por 16126C. Caso não seja verificado mais nenhuma outra variação, fica subentendido que as demais bases entre as duas sequências são idênticas, portanto, não será necessário qualquer tipo de representação das mesmas.

A CRS foi novamente sequenciada em 1999, através da placenta original, mesmo material utilizado por Anderson e colaboradores do primeiro sequenciamento na época (ANDREWS et al., 1999). Devido as melhorias conquistadas nas últimas duas décadas na área da tecnologia do sequenciamento de DNA, existiu um consentimento em revisar e corrigir os erros da CRS original para tornar seu uso mais robusto como sequência de referência no futuro. Um dos erros representou a perda de uma única citosina na CRS revisada ou *revised* CRS (rCRS) na posição 3106-3107, tornando-a uma base menor com 16568pb em relação a CRS original com 16569pb (ANDERSON et al., 1981). Entretanto, é utilizado uma deleção na posição 3107 que serve de apoio para manter a numeração histórica (ANDREWS et al., 1999). Felizmente, nenhum erro foi observado nas duas regiões hipervariáveis que são intensamente utilizadas em aplicações forenses, cuja sua extensão abrange as posições 16024-16365 e 73-340. A rCRS com 16568pb está disponível no site do MITOMAP (BRANDON et al., 2004).

É importante notar que a rCRS não é a única utilizada como sequência de referência. Por exemplo, o genoma mitocondrial utilizado como referência pelo *National Center for Biotechnology Information* é o Genbank AF347015[8], sequenciado por Ingman (INGMAN et al., 2000). Seu tamanho é de 16571pb e é derivada de um indivíduo Africano (Yoruba). Portanto, é indispensável destacar a sequência de referência que esteja sendo utilizada no estudo (BUTLER, 2005).

[7] O genoma mitocondrial é composto por duas fitas, a fita leve (*light strand* ou L-strand) e a fita pesada (*heavy strand* ou H-strand). A fita pesada é representada pelo círculo externo contendo um número maior de Guanina e Citosina do que a fita leve, representada pelo círculo interno (figura 2).

[8] Número de referência para busca no site. URL - `http://www.ncbi.nlm.nih.gov/Genbank/`

2.5 Marcadores forense

A região de maior variação do DNAmt, entre indivíduos humanos na população, é encontrada na região controle ou D-loop, como é descrito na seção 2.1. As duas regiões do D-loop conhecidas por região hipervariável I (HVI, HV1 ou HVS-I) e a região hipervariável II (HVII, HV2 ou HVS-II) são normalmente examinadas pela amplificação do PCR, e em seguida, são realizados análises das suas seqüências. Aproximadamente 610pb são avaliados através da região controle, 342pb da HVI e 269pb da HVII (figura 7).

Para cada seqüência de amostra, suas posições de 16024 a 16365 em HVI e 73 a 340 em HVII são determinadas e comparadas com a rCRS. As diferenças entre a comparação são anotadas e exibidas indicando a posição e o nucleotídeo da base alterada.

Em alguns casos, uma terceira região hipervariável (HVIII), com 137pb e compreendendo as posições 438 a 574, é examinada. Sitios adicionais de polimorfismos, dentre HVIII, poderão ser de grande valia na resolução de casos forenses, onde as regiões HVI e HVII de uma amostra estejam indistinguíveis.

Um número de diferentes tipos de PCR e primers[9] são utilizados no sequenciamento para gerar dados de seqüências do DNAmt nas regiões HVI e HVII.

A região controle do DNAmt tem sido estimada a variar em torno de 1 a 2%, ou seja, perto de 7 a 14 nucleotídeos dentre os 610pb na comparação entre indivíduos sem nenhuma ligação de parentesco (BUDOWLE et al., 1999). Esta variação abrange toda região HVI e HVII, cuja medição é determinada com a análise do sequenciamento.

No entanto, existem *hotspots*[10] ou sítios hipervariáveis e regiões onde a maioria dessas mutações estão próximas (STONEKING et al., 1991). A parte disso, já foram desenvolvidos outros métodos para que se possa visualizar rapidamente variações do DNAmt e excluir amostras que não coincidirem (BUTLER, 2005).

[9]São filamentos de nucleotídeo que servem como ponto de partida para a replicação do DNA.

[10]Na genética, o termo *hotspot* significa: posições no DNA onde ocorrem mutações incomuns com um certo grau de frequência.

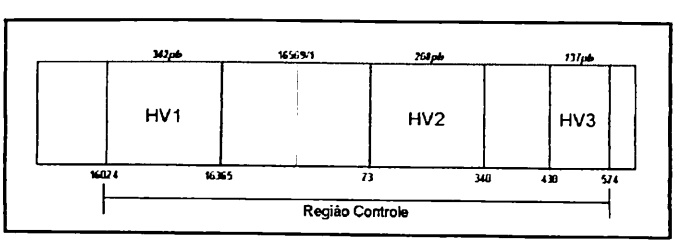


Figura 7: Ilustração das 3 regiões hipervariáveis do D-loop para utilização em investigações forenses.

2.6 Sequenciamento em casos forenses

Os passos envolvidos para se chegar na comparação de sequências de DNAmt estão ilustrados na figura 8. A extração do DNAmt deverá ser realizada em laboratórios com rigoroso controle de higiene, devido ao DNAmt ser mais sensível à contaminação do que o DNAn, uma vez que o seu número de cópias por célula é bem maior. Consequentemente, é preferível que se analise as amostras de referência depois que as amostras de evidência tenham sido processadas, para evitar qualquer pontencialidade de contaminação.

As amostras de DNA em investigação geralmente encontram-se bastante degradadas, o que pode dificultar a sua total legibilidade na etapa de leitura do sequenciamento. A amostra de referência de uma vítima, um suspeito e ou de um parente materno, é tipicamente disponível como manchas de sangue ou swabe bucal, contendo assim uma melhor qualidade de DNA (BUTLER, 2005).

O sequenciamento do DNAmt é realizado em ambas direções, *forward* (F) e *reverse* (R), para que as fitas complementares possam ser comparadas entre si, com o propósito de controle de qualidade. Se não for possível obter os dois filamentos de sequência, por exemplo, o segmento do poly-C[11], então o mesmo filamento poderá ser sequenciado duas vezes em reações separadas.

[11] Significa um conjunto contínuo de várias citosinas na sequência.
Exemplo: ATGCT*CCCCCCCC*GGTC

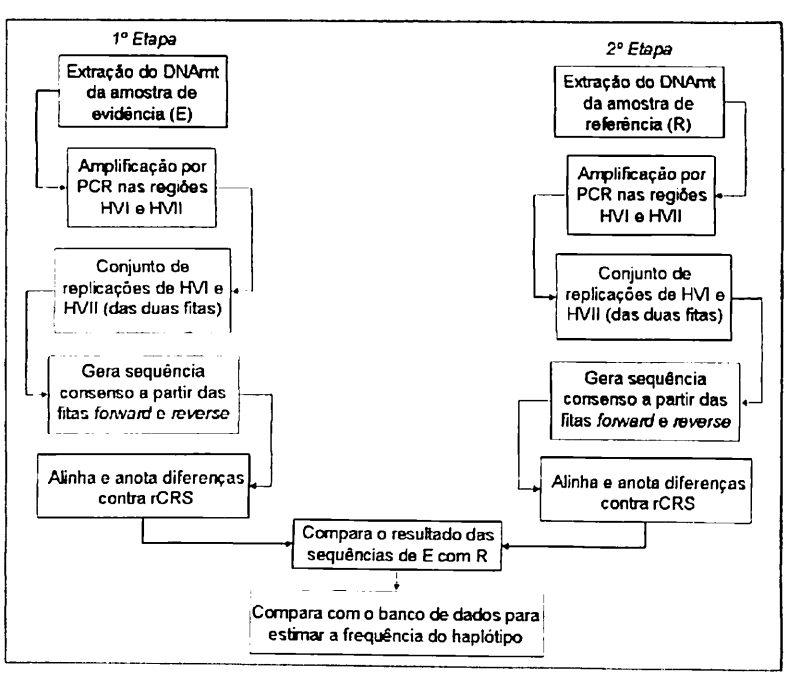


Figura 8: Processo para avaliação do DNAmt.

O processo de sequenciamento nem sempre tem como resultado dados limpos, claros e sem ambiguidade para cada base. Algumas regiões como o poly-C são desafiadoras de se decifrar (STEWART et al., 2001) e poderão até não ser incluídas na interpretação dos resultados finais.

Os sequenciamentos realizados por meios químicos e ou por instrumentos têm sido aperfeiçoados ao longo dos anos, o que proporcionou uma melhora na aparição dos picos, apresentando mais sensibilidade e pouco ruído. Mesmo assim, ainda é necessário uma revisão manual de cada nucleotídeo com o objetivo de editar as bases, caso o algoritmo de determinação de base tenha cometido algum erro (figura 9).

Ainda não existe nenhum software que possa, de forma robusta, avaliar sequências e dados do DNAmt de forma automatizada com confiança, sem nenhuma intervenção humana (BUTLER, 2005).

O processo de edição de sequência é realizado com a ajuda do alinhamento das duas sequências, *forward* e *reverse*, de uma amostra em questão. Deste alinhamento é gerado a sequência consenso (figura 10). Softwares tais como o SeqScape (Applied Biosystems[12]) alinham as sequências F e R, posicionando-as lado a lado e possibilitam a visualização de seus eletroferogramas.

Todavia, é aconselhável que dois analístas forenses examinem, interpretem e editem a mesma sequência para depois comparar o resultado das duas análises como uma última medida de qualidade de segurança (ISENBERG, 2004).

[12]URL - `http://www.appliedbiosystems.com/`

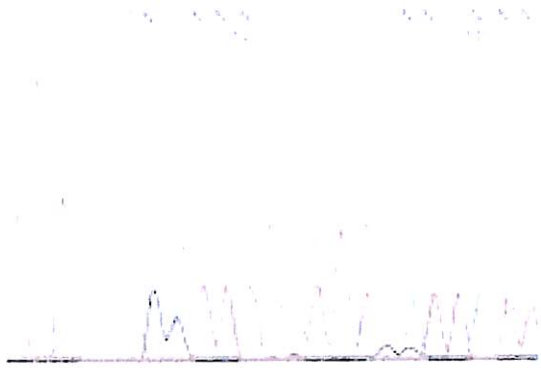


Figura 9: Exemplo ilustrativo de um eletroferograma. Os picos gerados do sequenciamento correspondem ao valor de qualidade de cada nucleotídeo na sua posiçao, representados por suas respectivas cores. Quanto maior o pico, maior o valor de qualidade da base. As bases com N significam que o sequênciador não conseguiu ler ou determinar um único tipo de nucleotídeo (A, G, T ou C) presente naquela posição.

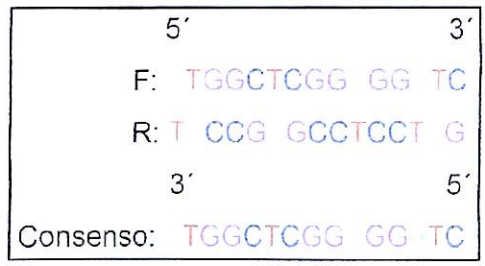


Figura 10: Ilustração do alinhamento de F e R para gerar a sequência consenso. Note que apenas as bases da F são mantidas tal como a rCRS quando foi sequênciada, para que ambas sejam alinhadas e comparadas entre si.

2.7 Exibindo diferenças em relação à rCRS

Após o processo de alinhamento exibe-se apenas, como resultado final das análises, as diferenças entre a comparação de uma sequência de amostra com a rCRS.

Quando diferenças são observadas, anota-se então a posição do nucleotídeo seguido de sua base. Por exemplo, na figura 11 as diferenças são obeservadas nas posições 16095 e 16131, e depois são relatadas no seu formato de dado 16095C e 16131A. Neste formato, assume-se que todos os outros nucleotídeos são idênticos aos da rCRS.

Bases ambíguas que não possam ser determinadas são usualmente anotadas por um N. Já em posições de ambiguidades confirmadas (heteroplasmia) os códigos da *International Union of Pure and Applied Chemistry* (IUPAC) poderão ser utilizados, tais como na Tabela 2 (SWGDAM, 2003).

A inserção de um nucleotídeo em uma sequência de DNAmt em relação a rCRS é relatada por anotar primeiro o sítio imediato a 5´ da posição de inserção em comparação com a rCRS, seguido de um ponto '.' e um 1 (para a primeira inserção) e um 2 (caso exista uma segunda inserção), e assim em diante, anotando no fim o nucleotídeo que foi inserido (BUDOWLE et al., 2003). Por exemplo, 315.1C é uma observação comum onde seis Cs são observados, acompanhado de um T na posição 310 numa sequência de amostra. Já a rCRS contém apenas cinco Cs nas posições 311-315 (ANDREWS et al., 1999). Portanto, a anotação 315.1C descreve a presença de cinco citosinas nas posições 311-315 na rCRS, e um extra C numa sequência de amostra, como inserção '.1C', antecedente a posição 316 (figura 12).

Deleções são anotadas pelo número da posição onde se observou a deleção relativo a rCRS e seguido por um hífen '-' ou por um D, d ou del, por exemplo: 309D, 309d, 309- ou del (figura 13).

A combinação das diversas formas para anotar inserções e deleções poderão gerar multiplas possibilidades na hora de relatar o resultado das diferenças em relação a sequência de referência. Devido a isto, foram criadas recomendações para garantir a consistência no tratamento do tamanho de variantes, no qual veremos com detalhes na próxima seção.

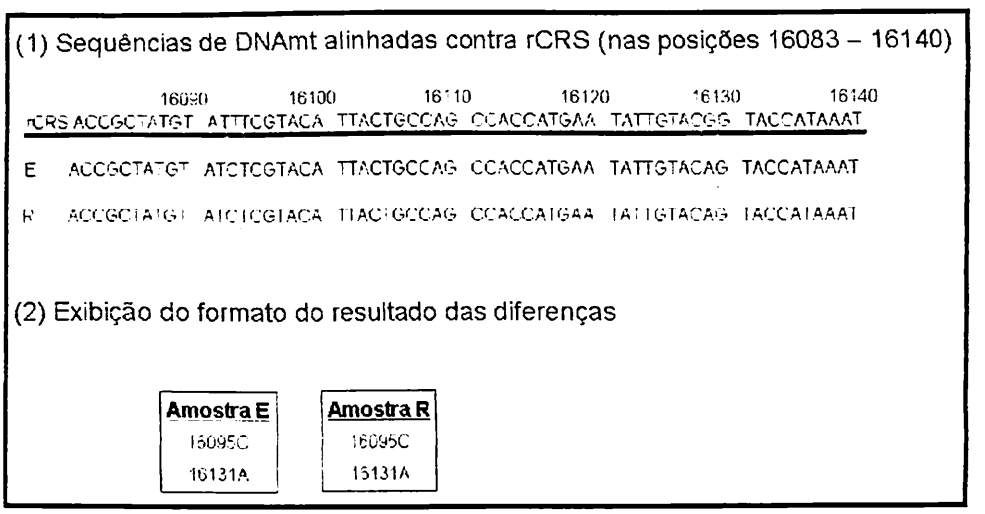


Figura 11: Comparação das sequencias de amostra (E) e (R) com a rCRS, relatando suas diferenças no seu devido formato de dado.

Código	Ambiguidade
R	A ou G
Y	C ou T
K	G ou T
M	A ou C
B	C, G ou T
D	A, G ou T
H	A, C ou T
V	A, C ou G
S	G ou C
W	A ou T

Tabela 2: Códigos da IUPAC para denominação da base de sítios que apresentarem mais de um nucleotídeo (heteroplasmia).

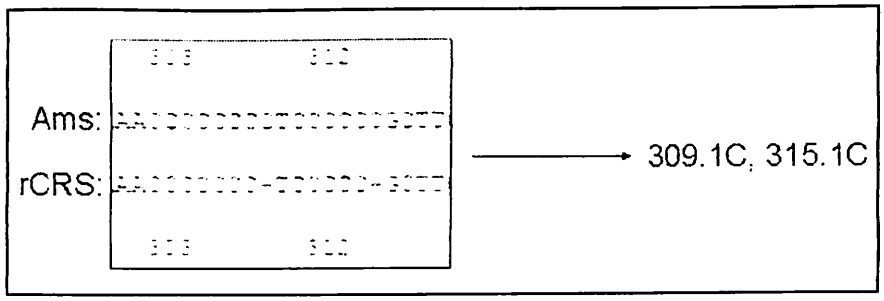


Figura 12: Anotação do resultado das inserções.

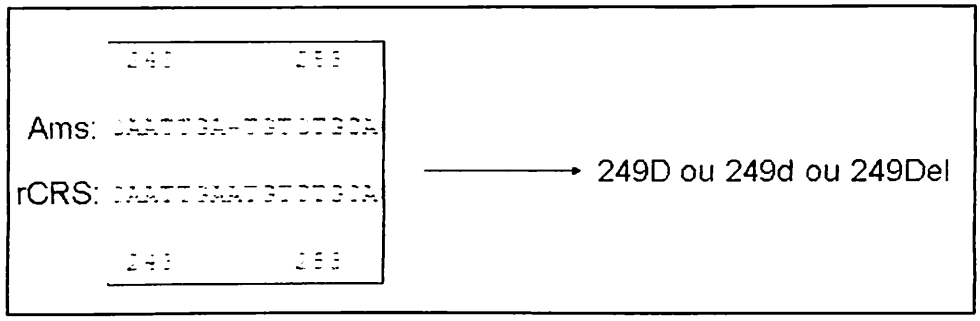


Figura 13: Anotação do resultado das deleções. A nomenclatura de anotação nos círculos vermelhos não são recomendadas, porém a sua utilização ainda poderá ser vista na literatura.

2.8 Interpretando resultados

Após ter completado a etapa de sequenciamento e análise das sequências, os resultados das amostras (E) e (R) são comparados e avaliados (figura 11). A comparação entre as duas sequências em questão irão resultar em um *match* perfeito ou não, muito embora essa interpretação nem sempre seja tão elementar. Os resultados estão organizados em três tipos de categorias: exclusão, inconclusivo ou não-excludente.

O *Scientific Working Group on DNA Analysis Methods* (SWGDAM) lista as seguintes recomendações na etapa de interpretação dos resultados:

- *Exclusão* ▷ Caso seja verificado a presença de dois ou mais polimorfismo na comparação entre o haplótipo de duas amostras em questão. O grau de parentesco materno ou a ligação de um indivíduo a uma amostra, neste caso, poderá ser de fato um resultado nulo.

- *Inconclusivo* ▷ Caso exista apenas uma única diferença, entre uma amostra de evidência em relação a uma amostra de referência, o resultado será encarado como inconclusivo, podendo este ter ou não um grau de parentesco materno.

- *Não-Exclusão* ▷ Se o haplótipo de cada amostra em questão apresentar o mesmo número de polimorfismo, como também, por exemplo: o mesmo tipo de base, um comprimento de variação comum no poly-C (heteroplasmia) em HV2; poderá-se concluir que existe um forte grau de parentesco materno entre as duas amostras em questão. Isto fortalece a acusação delas serem referentes a uma mesma pessoa ou de uma mesma linhagem materna.

Um exemplo sobre a interpretação de uma não-exclusão seria: uma sequência que contenha heteroplasmia num determinado sítio em relação a uma outra sequência em comparação e as cujas duas compartilham do mesmo perfil de haplótipo (figura 1). Neste caso, elas não poderão ser excluídas de possuir um grau de parentesco materno em comum (ver seção 2.3.2). Diversos exemplos poderão ser vistos com respeito às normas de interpretação (tabela 3), recomendadas pelo SWGDAM.

Resultado do alinhamento	Observações	Interpretação
(E) TATTGTACGG (R) TATTGTACGG	As duas sequências estão concordantes com o mesmo tipo de base em cada posição	Não Exclusão
(E) TATTGCACAG (R) TATTGTACGG	As sequências variam em duas posições	Exclusão
(E) TATTNTACGG (R) TATTGTACGG	Observado apenas uma única base não especificada (heteroplasmia), com todas as outras em comum	Não Exclusão
(E) TATTNTACGG (R) TATTGTACNG	Observado apenas uma ambiguidade em ambas as sequências em posições diferentes	Não-Exclusão
(E) TATTGTACA/GG (R) TATTGTAC G G	Detecção de heteroplasmia na sequência (E), com a sequência (R) apontando uma base de nucleotídeo em comum, ambas na mesma posição	Não-Exclusão
(E) TATTGTACA/GG (R) TATTGTACA/GG	Ambas as sequências compartilham de uma mesma heteroplasmia na mesma posição	Não-Exclusão
(E) TATTGCACGG (R) TATTGTACGG	Ocorrência de apenas uma única diferença entre as sequências	Inconclusivo

Tabela 3: Interpretação de resultados da análise entre duas sequências de DNAmt.

A única razão pela qual a diferença de uma única base é considerada como resultado inconclusivo é devido ao fato de já ter sido observado na literatura a ocorrência de mutação no genoma mitocondrial passado de mãe para filho (PARSONS et al., 1997). Por exemplo, se uma amostra de evidência de um sujeito (a) é analisada e comparada com a de um parente materno (b), como sendo a amostra de referência no teste, é possível que aconteça a diferença de uma única base (quando se espera o resultado de um *match* perfeito), mesmo que as duas amostras em questão sejam de mãe e filho. Neste caso, outras amostras são coletadas de ambas as partes para serem averiguadas com mais testes que possam vir a esclarecer esta incógnita a um resultado conclusivo.

2.9 Estimando o peso da evidência

Quando se obtem uma 'não-exclusão' como resultado de uma interpretação entre uma amostra de evidência e uma de referência, o próximo passo será estimar (estatísticamente) o grau de significância deste resultado de um *match* perfeito. A prática atual em communicar a raridade de um perfil de DNAmt, dentre um número de outros perfis escolhidos ao acaso (sem nenhuma ligação de parentesco entre si), envolve a contagem do número de vezes que este haplótipo em particular (perfil de DNAmt) é observado no banco de dados (WILSON et al., 1993; BUDOWLE et al., 1999). Esta aproximação é commumente denominada de "o método da contagem" na qual se baseia, totalmente, na busca do número de amostras presentes no banco de dados. Consequentemente, quanto maior for o número de amostras sem nenhuma ligação de parentesco (escolhidas ao acaso) no banco de dados, melhor será o valor do dado estatístico da estimativa na busca pela frequência daquele haplótipo em questão.

A frequência da maioria dos perfis de DNAmt na população é em torno de 60% desconhecida, no presente momento, pois eles ocorrem apenas uma só vez dentro do banco de dados (ISENBERG, 2004). Baseado em informações populacionais, o intervalo de confiança poderá ser utilizado para estimar os limites inferior e superior no cálculo de frequência (HOLLAND; PARSONS, 1999; TULLY et al., 2001).

Em casos onde um perfil de DNAmt em questão é observado em um número X vezes no banco de dados, composto por um número de N perfis, a frequência p deste perfil em questão poderá ser calculada utilizando-se a seguinte equação:

$$p = \frac{X}{N} \tag{2.1}$$

Um intervalo de confiança de 95% poderá ser utilizado como estimativa da frequência deste perfil, utilizando:

$$p \pm 1.96\sqrt{\frac{(p)(1-p)}{N}} \tag{2.2}$$

Em casos em que o perfil de DNAmt não seja observado no banco de dados, o intervalo de confiança de 95% é utilizado:

$$1 - \alpha^{\frac{1}{N}} \tag{2.3}$$

Onde α é o coeficiente de confiança (0.05 para o intervalo de confiança de 95%) e N é o tamanho do banco de dados.

Por exemplo, o perfil de DNAmt (haplótipo) do tipo: 16129A, 263G, 309D, 315.1C é observado duas vezes no banco de Africanos-Americanos dentre 1148 perfis, duas vezes em 1655 perfis Caucasianos e nenhuma vez em 686 perfis Hispânicos, quando comparado com perfis do *mtDNA Population Database* (MONSON et al., 2002). A seguir, utilizando as equações acima, calcularemos a raridade deste perfil nos seguintes exemplos:

1. Para Africano-Americano:

$$p = \frac{2}{1148} \pm 1.96\sqrt{\frac{(\frac{2}{1148})(1-\frac{2}{1148})}{1148}} = 0.0017 \pm 0.002 = 0.4\%$$

2. Para Caucasianos:

$$p = \frac{2}{1655} \pm 1.96\sqrt{\frac{(\frac{2}{1655})(1-\frac{2}{1655})}{1655}} = 0.0012 \pm 0.0017 = 0.29\%$$

3. Para Hispanicos:

$$1 - (0.05)^{\frac{1}{686}} = 1 - 0.9956 = 0.0044 = 0.44\%$$

Estes cálculos demonstram que o peso da evidência poderá ser um tanto similar quer seja encontrado ou não, um *match* do perfil no banco de dados. No primeiro exemplo, nós calculamos, com 95% de certeza, que a sua frequência varia entre os limites 0 e 0.004 (0.0017 ± 0.002). Mesmo assim, o maior valor que sua frequência poderá atingir é de 0.004%, ou seja, a probabilidade de encontrarmos este perfil na população é praticamente zero (0.4%). Com um perfil apresentando tal frequência, nós podemos excluir também, com 95% de certeza, 99.6% da população como possíveis fontes deste perfil DNAmt. Já no exemplo 3, amostras que nunca tenham sido observadas no banco de dados (Hispânico), poderá excluir 99.56% da população como possíveis fontes daquele perfil, com N igual a 686.

É importante ressaltar que o DNAmt nunca terá o poder de discriminação que um marcador STR de autossomos[13], visto que sua herança é uniparental. (HOLLAND; PARSONS, 1999; TULLY et al., 2001; ISENBERG, 2004)

[13] Qualquer região no cromossomo que não esteja ligada ao sexo.

2.10 Banco de dados populacional

Os bancos de dados populacionais atuam como uma ferramenta importantíssima e de uso indispensável para estimar a frequência dos haplótipos de DNAmt que são tratados em casos forenses, ou seja, quando se obtém um *match* na comparação entre o perfil DNAmt de um suspeito e o de uma amostra de referência.

Um grande esforço, de diversos laboratórios, já foi gasto para reunir milhares de informações sobre perfis de DNAmt de pessoas não relacionadas maternalmente e de diferentes grupos populacionais em todo mundo. O mais importante e também o mais difícil é gerar, construir e manter informações de alta qualidade para serem alimentadas aos bancos de dados, no intuito de poder estimar, com segurança, a frequência de um *match* ao acaso.

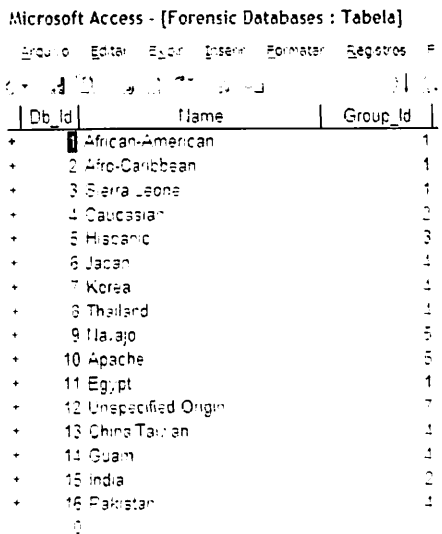
Uma das maiores fraquezas na análise do DNAmt é que alguns haplótipos são um tanto comum, mesmo nos diversos grupos de populações. Por exemplo, dentre os 1655 perfis DNAmt de Causacianos no banco de dados populacional do FBI[14], existem 15 perfis idênticos que dividem o mesmo haplótipo {263G, 315.1C} e outros 153 perfis que diferem em apenas um único polimorfismo. Assim, os 168 perfis de 1655 (10.2%) no banco de Caucasianos não podem ser excluídos caso seja observada uma amostra com este mesmo perfil de DNAmt.

Segundo Coble, já existem estudos para conseguir informações adicionais de sítios polimórficos em outras regiões do genoma mitocondrial, com o devido propósito de facilitar a resolução destes tipos comuns de perfil DNAmt (COBLE et al., 2004).

Os resultados da tipagem de amostras desconhecidas de DNAmt só serão valiosos se estas amostras forem avaliadas quando comparadas com uma amostra de referência e um banco de dados populacional. Já existem bancos de dados com mais de mil indivíduos sem nenhuma ligação de parentesco de múltiplos grupos populacionais (ATTIMONELL et al., 2000; WITTIGA et al., 2000; RÖHL et al., 2001; MONSON et al., 2002).

O maior banco de dados já anotado até então contem 14138 sequências de indivíduos das regiões HVI e HVII. Esta informação foi coletada de 103 publicações desde Janeiro de 2000, treze base de dados publicadas em 2000 e 2001 e de duas bases de dados não publicadas. Das 116 publicações, 90 delas requereram algum tipo de mudança e correção de erros ou ajustes na nomenclatura (RÖHL et al., 2001).

[14] *Federal Bureau of Investigation:* Orgão federal dos Estados Unidos que faz o papel de polícia federal daquele país. Sua sede fica em washington, DC.


Microsoft Access - [Forensic Databases : Tabela]

Arquivo Editar Exibir Inserir Formatar Registros

Db_Id	fName	Group_Id
1	African-American	1
2	Afro-Caribbean	1
3	Sierra Leone	1
4	Caucasian	2
5	Hispanic	3
6	Japan	4
7	Korea	4
8	Thailand	4
9	Navajo	5
10	Apache	5
11	Egypt	1
12	Unspecified Origin	7
13	China Taiwan	4
14	Guam	4
15	India	2
16	Pakistan	4

Figura 14: Ilustração de uma tabela do banco de dados de perfis de DNAmt do CODIS^{mt}.

Isto demonstra que gerar dados precisos de sequências de DNAmt para serem utilizados em bancos de dados forenses é uma tarefa complexa. A estimativa da frequência populacional para uma eventual "defesa legalizidada" de um tipo de DNAmt só poderá ser concretizada com a utilização de dados de alta qualidade e confiança que, consequentemente, ainda estão por ser produzidos.

O FBI já compilou o *mtDNA Population Database*, também conhecido como o CODIS^{mt} (MONSON et al., 2002), com o propósito de poder estimar frequências de perfis de DNAmt para uma defesa legalizada.

O banco de dados do CODIS^{mt} possui dados forenses e dados publicados pela literatura (MILLER; BROWN; BUDOWLE, 2003). A modelagem do seu banco de dados tem o intuito de separar dados obtidos de laboratórios, validados pela utilização de protocolos forenses, daqueles laboratórios de pesquisas acadêmicas cuja qualidade dos dados não foram revisados segundo os protocolos publicados na literatura (figura 16).

O banco de dados forense do CODIS^{mt} contem 4839 perfis de DNAmt de 16 diferentes tipos de populações (figura 14). Seus perfis foram sequenciados e seus eletroferogramas revisados nas posições 16024-16365 para HVI e nas posições 73-430 para HVII.

De modo a classificar os perfis de DNAmt, o CODIS^{mt} utiliza um padrão de 14 caracteres alfanuméricos, com 3 blocos separados por um '.', designados a identificar cada perfil de DNAmt no banco, por exemplo: 'BRA.CAU.000063'. O primeiro bloco de três caracteres representa o país de origem, o segundo block representa o grupo étnico e o último bloco, com 6 caracteres, representa a numeração sequencial.

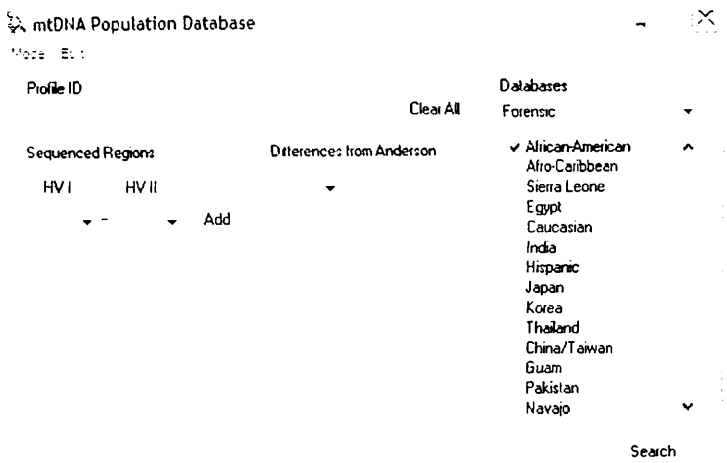


Figura 15: Tela principal do MitoSearch para a entrada dos polimorfismos a serem examinados na sua base de dados.

Este banco foi lançado em Abril de 2002 no formato de banco de dados do Microsoft Access e o seu download poderá ser feito pelo site do FBI na Internet, em conjunto com uma ferramenta de análise denominada MitoSearch (MONSON et al., 2002).

O MitoSearch pode comparar o haplótipo de perfis de DNAmt com os dados das populações da figura 14, na qual sua entrada de dados exige apenas os polimorfismos em relação a rCRS (figura 15). Como resultado, o obtém-se número de vezes que um perfil aparece dentre cada um dos grupos populacionais. Por exemplo, o perfil de DNAmt {16129A, 263G, 309D, 315.1C} ocorre duas vezes em 1148 perfis de Afro-Americanos, duas vezes em 1655 perfis de Caucasianos e nenhuma vez em 686 perfis Hispânicos.

Em contrapartida, a comunidade Européia de sequenciamento do DNAmt forense vem desenvolvendo um novo banco de dados populacional de alta qualidade para fins de aplicações de testes forenses e de identificação humana. A *European DNA Profiling Group* lançou o projeto *mitochondrial DNA population database* (EMPOP) para construir este novo banco. O sistema já está em fase de testes e poderá ser acessado *online*[15] para avaliação (com um pedido de autorização).

[15]URL - `http://www.empop.org`

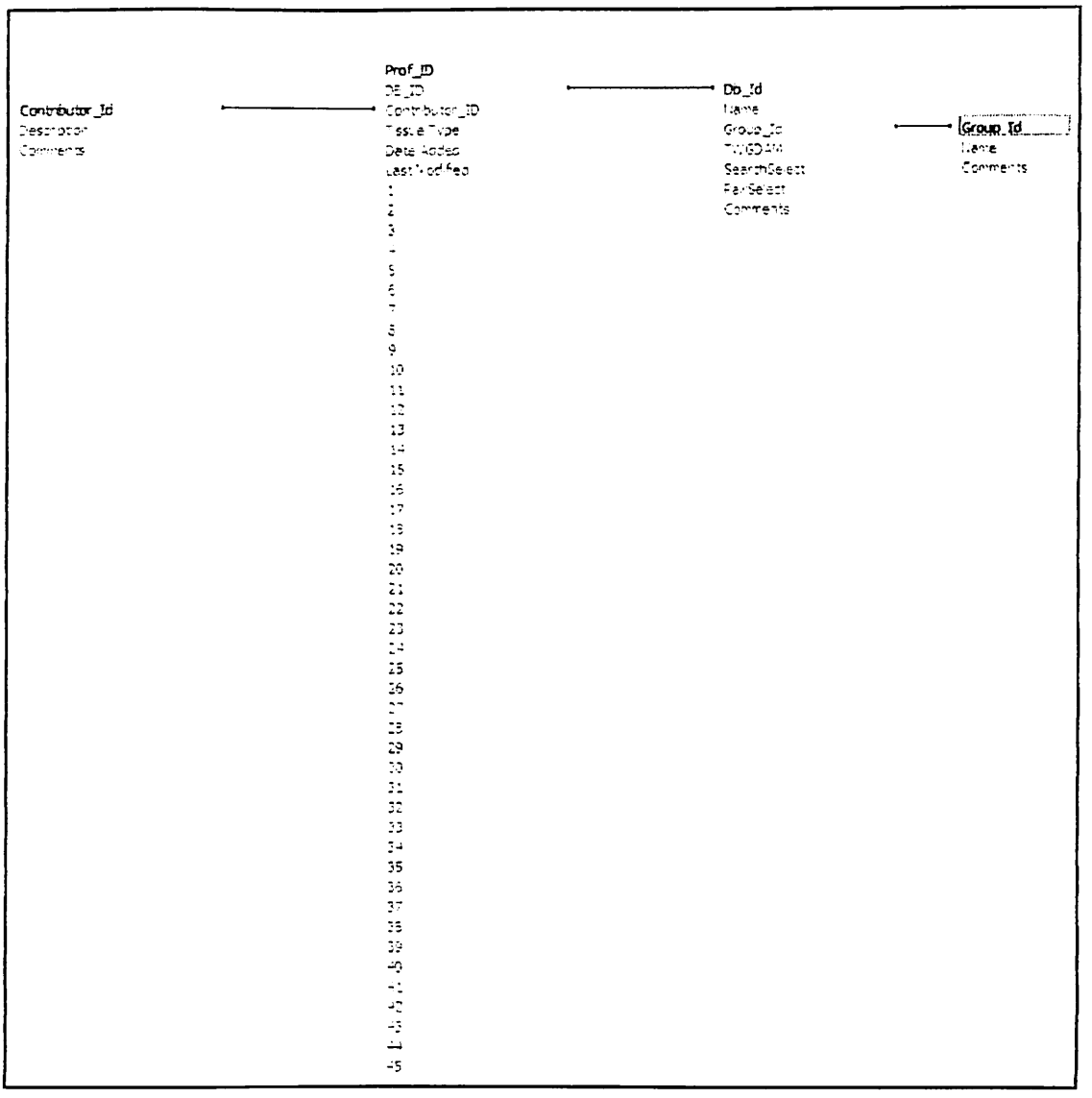


Figura 16: Ilustração do relacionamento entre as tabelas do banco de dados populacional do CODIS^{mt} (FBI). A tabela que armazena os perfis de DNAmt é a *"Forensic Profiles"*. Observe que esta tabela foi modelada de forma que cada polimorfismo, do haplótipo do perfil, é representado como sendo um atributo da mesma (pelos números de 1 a 45).

2.11 Definindo haplogrupos

Durante o progresso da tipagem do DNAmt, semelhanças entre os polimorfismos dos haplótipos das diversas amostras de diferentes regiões do globo foram observadas. Com isto, os haplótipos que apresentassem um certo grau de semelhança entre seus polimorfismos foram agrupados em haplogrupos (WALLACE; BROWN; LOTT, 1999; RUIZ-PESINI et al., 2004). Estes haplogrupos foram originalmente definidos em meados dos anos 80 e 90, agrupando-se as amostras que compartilhassem do mesmo, ou similiar, padrão de polimorfismos, de acordo com cada grupo populacional e sua origem geográfica (Tabela 4).

Os haplogrupos estão correlacionados aos polimorfismos das regiões HVI e HVII e também estão de acordo com outras variações do genoma mitocondrial que podem vir a ocorrer (SNP)[16]. Os haplogrupos denominados por letras alfanúmericas tais como A, B, C, D, E, F, G e M são associados a Asiáticos, enquanto que a maioria dos nativos Americanos se encaixam nos haplogrupos A, B, C e D. Os haplogrupos L1, L2 e L3 são de origem Africana e os haplogrupos H, I, J, K, T, U, V, W e X estão associados a populações da Europa (figura 17) (ALLARD et al., 2002, 2004, 2005; ALVES-SILVA et al., 2000).

As informações de haplogrupo autentica e liga os perfis de DNAmt a um determinado grupo populacional. Esta informação é importante no contexto forense pois serve como ajuda no controle de qualidade das seqüências dos perfis armazenados em bancos de dados populacionais (ACHILLI et al., 2004; BUDOWLE et al., 2003; YAO; BRAVI; BANDELT, 2004).

[16] A ocorrência de uma variação no DNA que envolve a mudança de apenas um único nucleotídeo.

Haplogrupo (População)	Região Codificante Polymorfismos	Região Controle Polymorfismos
A (Asiático)	663	16233T, 16290T, 16319A, 235G
B (Asiático)	Deleções de 9pb, 16159C	16217C, 16189C
J (Caucasiano)	4216C, 12612G, 13708C	16069T, 16126C, 259T
L1 (Africano)	2758A, 3594T, 10810C	16187T, 16189C, 16223T, 16278T, 16311C

Tabela 4: Exemplo de 4 haplogrupos demonstrando o seu padrão de polimorfismos das regiões codificantes e da região controle, como também a sua origem geográfica.

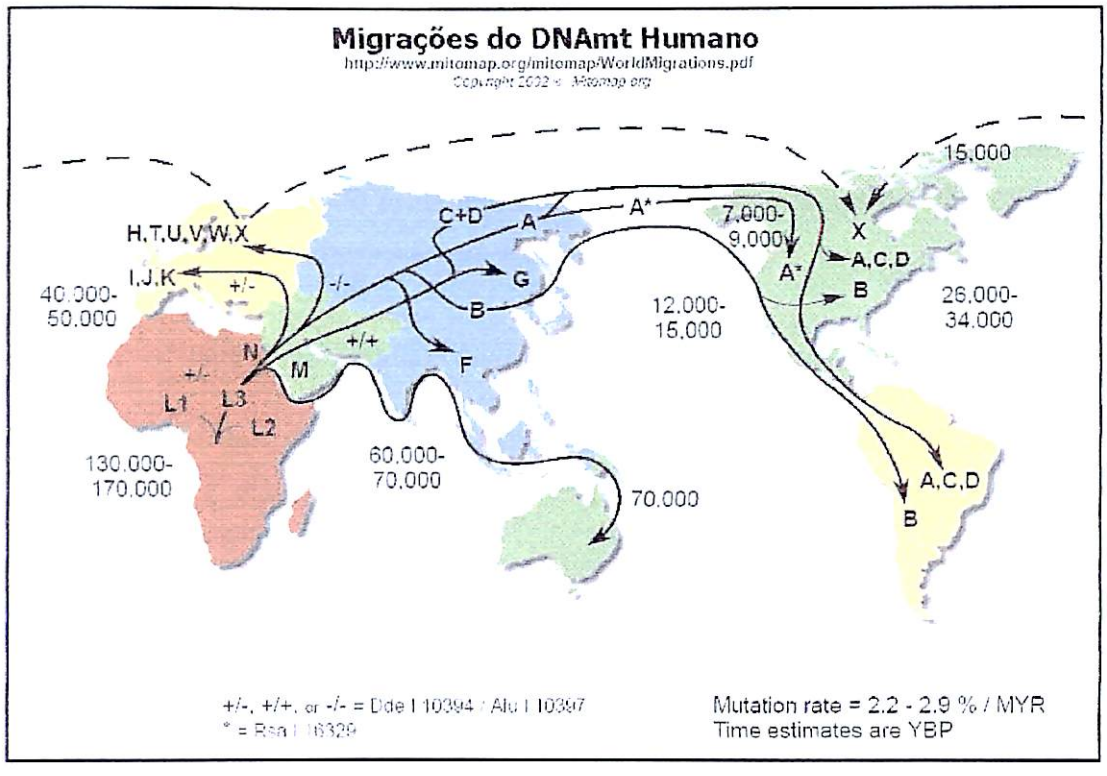


Figura 17: Distribuição dos haplogrupos nas diversas regiões do mundo.

2.12 Normas de alinhamento

Ambiguidades com respeito a nomenclatura dos resultados do DNAmt poderão resultar em duas diferentes análises com relação a uma mesma amostra. Banco de dados populacionais poderão ter multiplas entradas diferentes para um mesmo haplótipo de DNAmt, impedindo assim uma estimativa precisa da frequência de um perfil em particular. Portanto, a padronização na designação das sequências de DNAmt é de extrema importância para que se possa gerar e compartilhar dados entre os laboratórios (CAETANO et al., 2005a, 2005b).

O alinhamento entre as sequências de amostra e a rCRS tem como grande desafio a determinação de polimorfismos. O modo de tratamento de inserções e deleções (*gaps*) poderão variar entre os laboratórios, o que causará diferentes codificações para uma mesma amostra. O laboratório do FBI elaborou um número de recomendações para possibilitar a consistência no tratamento do tamanho de variantes em sequências de DNAmt (WILSON et al., 2002). Três recomendações foram feitas:

1. Relatar os perfis com a menor quantidade de polimorfismos em relação a rCRS.
2. Caso exista mais de uma forma para relatar perfis com a menor quantidade de polimorfismos em relação a rCRS, tais polimorfismos deverão ser priorizados da seguinte maneira:
   a) inserções deleções (indels).
   b) transição.
   c) transversão.
3. Inserções e deleções deverão ser postos na 3´ em relação a *light strand*.

As inserções e deleções deverão ser combinadas em situações que mantenham o mesmo número de polimorfismos em relação a rCRS. Estas recomendações estão organizadas em hierarquia, portanto, a recomendação 1 deverá tomar precedência sobre a recomendação 2 e 3. Demonstraremos alguns exemplos na seção 3.2.1 que necessitam de consistência no tratamento do tamanho de variantes na análise de sequências de DNAmt, e como é realizado a anotação do seu resultado.

2.13 Parâmetros de controle de qualidade

É evidente a preocupação dos responsáveis pelos bancos de dados forenses com a qualidade das sequências de DNAmt armazenadas e a importância destes bancos no processo de estimar a freqüência de um *match* de uma amostra de evidência ligada a um suspeito, para ser utilizada como prova em um julgamento legal nos tribunais de justiça.

É demonstrado que as sequências de DNAmt publicadas estão predispostas a conter erros, principalmente devido a má interpretação dos dados brutos de sequência (eletroferograma) e ou a introdução de erros no processo de transcrição dos dados e dos resultados (RÖHL et al., 2001; BANDELT et al., 2001, 2002; FORSTER, 2003; DENNIS, 2003).

Em 1999, na conferência internacional da *International Society for Forensic Genetics* (ISFG), o *European DNA Profiling* (EDNAP) apresentou a proposta de desenvolver uma aplicação *Web* para a criação de um banco de dados *online* de DNAmt (EMPOP) (PARSON et al., 2001b) que armazene dados populacionais de alta qualidade. Os dados que são importados para o banco de dados do EMPOP são fornecidos somente pelos laboratórios internacionais de DNA forense que participarem dos exercícios colaborativos (*collaborative exercises* - CE) da EDNAP (PARSON et al., 2004a). Os princípios básicos do CE foram modelados em sistemas existentes de controle de qualidade forense (CARRACEDO et al., 1998, 2001; RAND; SCHüRENKAMP; BRINKMANN, 2002) na tentativa de avaliar as seguintes áreas:

1. A geração de resultados coerentes de sequências de DNAmt entre diferentes laboratórios, utilizando suas técnicas individuais e instrumentais.

2. Os requisitos necessários para estabelecer, com sucesso, um processo padronizado de análise das sequências de DNAmt para uniformizar a nomenclatura e a interpretação dos dados.

3. Introduzir o uso de um sistema computacional para garantir a transferência e o armazenamento dos dados, com segurança.

De forma conceitual, o método mais utilizado na detecção de erros é bastante simples: o haplótipo de DNAmt precisa se encaixar dentro de uma parte específica da filogenia que é caracterizada por polimorfismos ou mutações específicas. Em situações onde não se consegue encaixar o haplótipo dentro de um padrão filogenético, é esperado então que este perfil tenha tido a sua natureza biológica modificada ou que ele seja apenas desconhecido.

Sítio	100	101	102	103
rCRS	T	G	A	T
1º Amostra	▪	▪	▪	A
2º Amostra	▪	C	▪	▪
3º Amostra	▪	C	▪	▪

Figura 18: Exemplo hipotético da troca de coluna na preparação de uma tabela de dados, demonstrando os polimorfismos das amostras em relação a rCRS. Os pontos representam a igualdade das bases. A listra pontilhada significa a coluna que foi trocada e as setas representam as devidas posições das colunas.

Este processo é realizado utilizando-se uma ferramenta de análise filogenética. Através dela, as similaridades e as diferenças entre múltiplas sequências relacionadas (isto é, de uma mesma região) poderão ser comparadas sistematicamente. Nela, o haplótipo é comparado com vários outros perfis na tentativa de verificar se o haplótipo difere extremamente dos outros. Um resultado com diferenças extremas ou incomuns poderá ser o indício de que a sequência foi contaminada ou anotada de maneira indevida. Por exemplo, o laboratório poderá trocar a sequência de HVI da amostra (a) e uni-la com a HVII da amostra (b), criando assim uma recombinação artificial ou uma composição de sequência por acidente. Portanto, o uso da análise filogenética pode ajudar na verificação de erros e na qualidade das sequências (BANDELT et al., 2001; BUDOWLE; POLANSKEY; ALLARD, 2004). Segundo (SALAS et al., 2005), existem cinco classes de erros que podem afetar a sequência, são elas:

- Classe 1 (*base shift*) : Um ou mais sítios são mau classificados no alinhamento, por exemplo, 316.1C ao invés de 315.1C; Lê-se a posição errada do sítio no alinhamento, por exemplo, o alinhamento aponta 73G e o resultado anotado é 74G; Troca-se a coluna durante a preparação de uma tabela de dados (figura 18).

- Classe 2 (*reference bias*) : Anota-se a base da rCRS ao invés do polimorfismo da amostra. Por exemplo, rCRS: 73A e amostra: 73G, anotando 73A ao invés de 73G.

- Classe 3 (mutação fantasma) : São polimorfismos incomuns que podem aparecer na sequência devido ao uso incorreto do sequenciador, da contaminação no manuseio da amostra (produtos bioquímico). É considerado o estado de degradação da amostra.

- Classe 1 (*base misscoring*) : Anotar errado, numa tabela de dados, a letra do nucleotídeo, trocar um nucleotídeo por um ponto '.' e/ou confundir uma transição a uma transversão e vice-versa, como ilustra a figura 18.

- Classe 5 (recombinação artificial) : Combinar HVI e HVII de diferentes amostras ao determinar o haplótipo de um perfil.

Uma inspeção geral realizada nos dados Afro-Americanos do banco de dados de DNAmt do SWGDAM (CODIS - FBI) apontou uma série de deficiências em relação a sua qualidade. Dentre os 1148 perfis, foi detectado cinco erros de recombinação artificial que podem ter sido ocasionados devido a uma mistura de amostras e/ou de materiais em laboratório, ou até mesmo, durante a transcrição dos dados (BANDELT; SALAS; BRAVI, 2004).

Em Bandelt, Salas e Lutz-Bonengel (2001), é descrito um dos erros mais explícitos encontrados no banco do FBI referente a combinação híbrida de dados no haplótipo do perfil USA.AFR.000942 que combina o primeiro segmento HVI, referido ao haplogrupo Africano L1b, com o segundo segmento HVII, referido a um haplogrupo de Nativo Americanos, denominado C1.

Uma série de outros erros como os da classe 1, 2, 3 e 4 já foram detectados no banco do FBI e acredita-se que ainda existam erros a serem descobertos e corrigidos. Estes erros vêm sendo detectados desde o início de 2001 até os dias de hoje nos diversos bancos de dados de DNAmt publicados na literatura. A oposição do instituto nacional de justiça dos E.U.A ao CODIS, devido a sua extensa avaliação sobre a qualidade e confiança dos seus dados, inibiu o FBI na geração de um novo e confiável banco de dados de DNAmt nos E.U.A.

A abordagem filogenética é ainda considerada o ponto de partida para uma reanálise sistemática dos dados de DNAmt, mesmo levando em consideração que esta ferramenta detecta apenas, em média, 50% dos erros (SALAS et al., 2005). Inúmeros erros que são difíceis ou quase impossíveis de se observar através da análise filogenética poderão permanecer no banco.

Desta forma, alguns dos haplogrupos oeste eurasianos, tal como o haplogrupo H, possuem quase nenhum sítio para o diagnóstico de HVI e HVII que facilitaria a detecção de recombinação artificial através do seu estudo filogenético (SALAS et al., 2005). O Mitomap *database* (BRANDON et al., 2004) também poderá servir de ajuda na verificação de erros através da sua base de dados de mutações variantes, da região controle, validadas pela literatura.

Neste capítulo, abordamos as principais características do DNAmt e suas problemáticas encontradas pela commnidade forense para produzir e gerar dados de perfis genéticos de qualidade para serem utilizados em estudos de casos forenses.

No próximo capítulo discutiremos a problemática de se alinhar sequências de DNAmt com exemplos de testes, realizados em ferramentas de análises, para alinhar sequências de nucleotídeos como, por exemplo, a ferramenta comercial SeqScape, bastante utilizada em laboratórios de DNA forense.

3 *Alinhando Sequências do DNA Mitocondrial Forense*

"Biology easily has 500 years of exciting problems to work on"
Donald E. Knuth

Este capítulo introduz o conceito do alinhamento de sequências, em especial o alinhamento global entre duas sequências, apresentando um dos métodos mais utilizados na comparação de sequências de DNA através da programação dinâmica. Por fim, é demonstrado a problemática encontrada no alinhamento de sequências de DNA Mitocondrial no estudo de casos forenses. A seção 3.1 e suas subseções estão fortemente baseadas nos capítulos 2 e 3 do livro Setubal e Meidanis (1997), no capítulo 6 do livro Pevzner (2001) e nas dissertações de mestrado Ticona (2003), Montera (2004) ...

3.1 Comparação de sequências biológicas

A comparação de sequências pode ser considerada a operação primitiva mais importante na biologia computacional. Basicamente, esta operação consiste em encontrar o maior número de semelhanças ou o menor número de diferenças entre duas ou mais sequências. Existe uma variedade de problemas, com diversas formulações, que podem requerer a utilização de diferentes estruturas de dados e algoritmos para obter uma solução eficiente ou até mesmo uma que seja apenas razoável.

Uma das funcionalidades na comparação de sequências é a determinação de um ancestral comum a partir de um conjunto de sequências. Por exemplo, ao comparar sequências de seres humanos com a de alguns primatas, é possível descobrir a sequência hipotética de uma espécie ancestral de ambos extinta. Se duas sequências possuem um ancestral comum, elas são ditas homólogas. É muito provável que duas sequências muito parecidas sejam homólogas, e as vezes, o mesmo pode ser dito em relação a duas outras sequências que não sejam muito parecidas. Ainda mais, é possível que duas sequências sejam parecidas embora não sejam homólogas.

Os primeiros algoritmos utilizados na comparação de sequências biológicas apareceram na década de 70 através da técnica de programação dinâmica (NEEDLEMAN; WUNSCH, 1970; SANKOFF, 1975; SMITH; WATERMAN, 1981). Posteriormente, surgiram modelos heurísticos e probabilísticos. Técnicas baseadas no aprendizado de máquina como redes neurais, algoritmos evolutivos, algoritmos genéticos e *simulated annealing* também foram desenvolvidas (LECOMPTE et al., 2001).

O alinhamento de sequências, o qual é formalmente denominado a comparação de sequências biológicas, tem um papel central na era pós-genômica. Lecompte e colaboradores (LECOMPTE et al., 2001) apontam algumas informações obtidas a partir do resultado de sua análise:

- Determinação da função biológica mediante a homologia entre sequências
- Busca de padrões conservados ao longo da evolução que podem servir como pistas para encontrar estruturas conservadas, sinais de localização ou resíduos funcionais que podem descrever uma família ou subfamílias de proteínas.
- Estudos evolutivos para definir relações filogenéticas entre sequências.
- Organização dos domínios dentro de uma família protéica.

- Construção da estrutura molecular a partir do alinhamento de nucleotídeos com uma proteína de estrutura conhecida e, consequentemente, determinar a sua função biológica.

Em se tratando de comparar sequências de DNAmt para estudos de casos forenses, é levado em conta, neste trabalho, a seguinte situação:

- Teremos duas sequências compostas pelo mesmo alfabeto (A, G, T, C e/ou N), ambas com quase o mesmo tamanho e possuindo algumas centenas de pares de base. Sabemos que as duas sequências são muito similares, apresentando apenas algumas diferenças isoladas, tais como: inserções, deleções e substituições de caracteres. Com isto, nosso objetivo será o de encontrar o menor número de diferenças entre duas sequências através do alinhamento.

A próxima seção apresenta os principais métodos utilizados na comparação de sequências biológicas.

3.1.1 Alinhamento de sequências

Esta seção descreve os métodos mais utilizados na comparação de duas sequências. Especificamente, o objetivo é encontrar o melhor alinhamento entre elas.

Na prática, poderão acontecer diversas versões desta problemática na qual dependerá se estamos interessados em alinhar as sequências por inteira ou apenas suas subsequências. A primeira definição leva ao conceito de uma comparação global e a segunda ao de uma comparação local. Também poderá acontecer uma terceira prática de comparação, cujo objetivo é alinhar os prefixos e sufixos entre as duas sequências. Denominamos esta terceira definição por comparação semiglobal e explicaremos cada uma delas na seção 3.1.3.

Alinhar sequências é colocar uma sequência sobre a outra, de forma que a correspondência entre elas fique clara. Por exemplo, considere estas duas sequências de DNA onde t é a sequência que desejamos comparar, a sequência de amostra, e s a sequência de referência:

$$t = \text{GACGGATTAG}$$
$$s = \text{GATCGGAATAG}$$

Ambas as sequências são muito parecidas e diferem em tamanho por apenas uma posição. Ao alinharmos estas duas sequências suas diferenças ficam ainda mais nítidas, como podemos ver:

$$t = \text{G A - C G G A T T A G}$$
$$s = \text{G A T C G G A A T A G}$$

Após o alinhamento de s e t verificamos que as únicas diferenças são apenas a deleção de um T na sequência de amostra e uma substituição, ou mudança de base, de A para T, na quarta coluna, lendo-se da direita para a esquerda. Foi inserido um *gap* na sequência de amostra para equacionar perfeitamente as bases entre as duas sequências, tanto antes como depois do *gap*.

Com o exemplo do alinhamento acima, podemos validar a variação no tamanho entre as duas seqüências pela inserção do *gap*. Desta forma, o alinhamento é definido como sendo a inserção arbitrária de *gaps* ao longo das posições das seqüências para que elas se tornem do mesmo tamanho. Ao termos as duas seqüências em tamanhos iguais, elas poderão então ser posicionadas, uma sobre a outra, criando assim uma correspondência entre seus caracteres e *gaps*. A inserição do *gap* pode ser posicionado tanto no começo, no meio ou no fim das seqüências de forma a não ter *gaps* alinhados.

3.1.2 Esquema de pontuação

Dado o alinhamento entre duas seqüências, o próximo passo a ser tomado é verificar o grau de similaridade deste alinhamento através de um sistema de pontuação.

Cada coluna do alinhamento irá receber um certo valor e o valor total do alinhamento será a soma de todos os valores designados às colunas, de acordo com o seguinte esquema de pontuação:

- *Caracteres ou pares de base iguais* ▷ Indica a ocorrência de um "casamento" ou um *match* entre o caractere de cada seqüência, na mesma posição ou coluna: valor +1.
- *Caracteres ou pares de base distintos* ▷ Indica a ocorrência de uma substituição de bases ou um *mismatch* entre o caractere de cada seqüência, na mesma posição ou coluna: valor -1.
- *Um gap* ▷ Indica a ocorrência de uma deleção quando o *gap* estiver presente na seqüência t; uma inserção, caso o *gap* esteja presente na seqüência s: valor -2.

Dentre as possibilidades de alinhamento entre duas seqüências, o melhor deles é aquele que apresenta o *valor máximo* pontuado, ou seja, aquele com o maior grau de similaridade entre as bases.

Considere as seqüências s = {ATCCGAT} e t = {ACGAAGT}, as figuras 19, 20 e 21 ilustram três exemplos de possíveis alinhamentos entre elas. Utilizando o esquema de pontuação acima, o melhor alinhamento é o ilustrado na figura 21 cujo *valor máximo* é -2.

$$\begin{array}{cccccccc} A & T & C & C & G & A & T & - \\ & | & & | & | & & | & | \\ A & - & C & G & A & A & G & T \end{array}$$

Figura 19: Alinhamento entre s e t com *valor máximo* de: $3 \cdot 1 + 3 \cdot -1 + 2 \cdot -2 = -4$.

$$\begin{array}{cccccccc} A & T & C & C & G & A & - & - & T \\ & | & | & & & & | & | & \\ A & - & - & C & G & A & A & G & T \end{array}$$

Figura 20: Alinhamento entre s e t com *valor máximo* de: $5 \cdot 1 + 0 \cdot -1 + 4 \cdot -2 = -3$.

$$\begin{array}{ccccccccc} A & T & C & C & G & A & - & T \\ & | & & | & | & & | & \\ A & - & C & G & A & A & G & T \end{array}$$

Figura 21: Alinhamento entre s e t com *valor máximo* de: $4 \cdot 1 + 2 \cdot -1 + 2 \cdot -2 = -2$.

3.1.3 Tipos de alinhamento de sequências

Nesta seção, descreveremos, brevemente, os três tipos de alinhamentos mais utilizados na literatura. Eles são: o alinhamento global, local e semiglobal.

Alinhamento Global

O objetivo do alinhamento global é alinhar todas as posições da sequência s com todas as posições da sequência t. Por exemplo:

- Dado as sequências s = {GAAGGATTAG} e t = {GATCGGAAG};
- Temos o seguinte alinhamento global:

$$\begin{array}{cccccccccc} G & A & T & C & G & G & A & - & - & A & G \\ & | & | & & & & & & | & | & \\ G & A & A & - & G & G & A & T & T & A & G \end{array}$$

Com o valor máximo de: $7 \cdot 1 + 1 \cdot -1 + 3 \cdot -2 = 0$.

Note que todas as posições de ambas as sequências foram alinhadas com sucesso.

Alinhamento Local

O objetivo do alinhamento local é alinhar uma subcadeia de s com uma subcadeia de t. A pontuação é calculada apenas sobre a região onde as subcadeias estão alinhadas. Por exemplo:

- Dado as sequências s = {AAGACGG} e t = {GATCGAAG};
- Temos o seguinte alinhamento local:

$$\begin{array}{cccccccc} G & A & T & C & G & A & A & G \\ & & & & & & & \\ & & & & A & A & G & A & C & G \end{array}$$

Com o valor máximo de: $3 \cdot 1 + 0 \cdot -1 + 0 \cdot -2 = 3$.

Este tipo de alinhamento é utilizado ao comparar grandes sequências de DNA ou quando duas proteínas compartilham um domínio comum. É também utilizado para comparar sequências com menor grau de similaridade.

Alinhamento Semiglobal

O objetivo do alinhamento semiglobal tem o mesmo princípio do alinhamento global: comparar duas sequências por inteira. Mas, no alinhamento semiglobal, todos os *gaps* encontrados na extremidade da sequência são ignorados no cálculo da pontuação.

Os *gaps* da extremidade são aqueles que estão localizados antes do primeiro e/ou depois do último caractere na sequência. Por exemplo:

- Dado as sequências s = {GACGACTTTCCATT} e t = {GATCGTCC};
- Temos o seguinte alinhamento semiglobal:

$$\begin{array}{cccccccccccccc} G & A & C & G & A & - & C & T & T & T & C & C & A & T & T \\ | & | & | & & & & | & & | & & | & & & | & | & | \\ - & - & - & G & A & T & C & G & T & - & C & C & - & - & - \end{array}$$

Com o valor máximo de: $6 \cdot 1 + 1 \cdot -1 + 2 \cdot -2 = 1$.

Este tipo de alinhamento é geralmente utilizado quando se compara sequências de tamanhos divergentes. Sendo também bem empregado na localização de genes similares.

3.1.4 O algoritmo de alinhamento global

Uma das abordagens para se computar as similaridades entre duas sequências seria gerar todos os alinhamentos possíveis entre elas, depois, analisar cada um dos resultados e, em seguida, escolher somente um deles como sendo o melhor alinhamento.

A seguir, apresentaremos o algoritmo utilizado neste trabalho baseado na técnica de programação dinâmica. Esta técnica consiste em resolver a instância de um problema aproveitando as soluções já computadas e resolvidas de outras instâncias menores do mesmo problema. Por exemplo: dado duas sequências s e t, ao invés de determinar suas similaridades, considerando o tamanho de suas sequências por inteiro, são determinadas as similaridades entre prefixos arbitrários das sequências, utilizando a resolução de pequenos prefixos para resolver prefixos maiores. Neste trabalho, cada base nucléica da sequência representa um prefixo.

Considere m o tamanho da sequência s, e n o tamanho da sequência t. Irão existir $m + 1$ possíveis prefixos de s e $n + 1$ possíveis prefixos de t. Assim, poderemos arrumar nossos cálculos como uma matriz de $(m + 1) \times (n + 1)$ onde a entrada (i, j) contêm as similaridades entre $s[1..i]$ e $t[1..j]$.

Exemplificando, a figura 22 ilustra a matriz correspondente a s = AAAC e t = AGC. Colocamos s ao longo da margem vertical à esquerda e t ao longo da margem horizontal acima para facilitar a visualização dos prefixos. Notem que a primeira coluna e a primeira linha são inicializadas com multiplos da penalidade do *gap* (-2 para o nosso caso). Isto justifica o fato de que só poderá existir um único alinhamento possível, caso alguma das sequências esteja vazia.

Ao começarmos a preencher a matriz, observamos que podemos computar o valor para uma entrada (i, j) ao reparar três entradas prévias: aquelas por $(i - 1, j)$, $(i - 1, j - 1)$ e $(i, j - 1)$, devido ao fato de que só exitem três formas de se obter um alinhamento entre $[s1..i]$ e $t[1..j]$, e cada um deles utiliza um desses três resultados prévios. Portanto, para preencher a matriz e conseguir um alinhamento entre $s[1..i]$ e $t[1..j]$, teremos uma das seguintes opções para escolher:

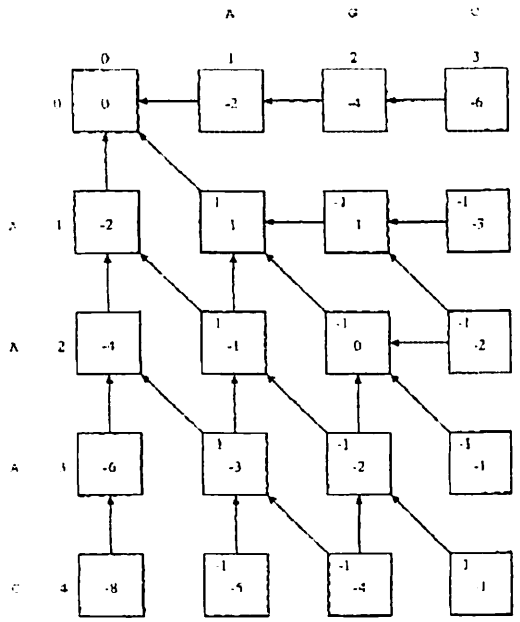


Figura 22: Matriz gerada para computar alinhamentos ótimos.

- Alinhar $s[1..i]$ com $t[1..j-1]$ e casar $t[j]$ com um *gap*.
- Alinhar $s[1..i-1]$ com $t[1..j-1]$ e casar $s[i]$ com $t[j]$.
- Alinhar $s[1..i-1]$com $t[1..j]$ e casar $s[i]$ com um *gap*.

Consequentemente, a busca pela similaridade é realizada através da seguinte fórmula:

$$sim(s[1..i], t[1..j]) = max \begin{cases} sim(s[1..i], t[1..j-1]) - 2 \\ sim(s[1..i-1], t[1..j-1]) + p(i,j) \\ sim(s[1..i-1], t[1..j]) - 2 \end{cases} \quad (3.1)$$

A função $p(i,j)$ retorna +1 caso $s[i] = t[j]$, e -1 caso $s[i] \neq t[j]$. Os valores de $p(i,j)$ estão escritos no canto esquerdo superior das caixas na figura 22.

O preenchimento da matriz não precisa seguir uma ordem específica. Ela poderá ser preenchida linha por linha, coluna por coluna, da esquerda para direita, de cima para baixa e vice-versa, contanto que a restrição da função de similaridade, descrita acima, seja garantida.

Finalmente, as setas ilustradas na figura 22 indicam o caminho de origem do *valor máximo* gerado de acordo com a função de similaridade. Denominando a matriz da função de similaridade por a, temos que o valor de $a[1,2]$ foi obtido a partir do maior valor gerado dentre os seguintes valores da função de similaridade:

$$a[1,2] = max \begin{cases} a[1,2-1] - 2 \\ a[1-1,2-1] + p(1,2) \\ a[1-1,2] - 2 \end{cases}$$

O maior valor obtido pela função de similaridade vem da posição $a[1,1]$, com valor de -1, que será o dado de entrada para a posição $a[1,2]$, como ilustra a seta na figura 22.

Ao preenchermos toda a matriz com a utilização da função de similaridade, estaremos finalizando o processo de calcular todas as similaridades entre as duas seqüências. A seguir, realizaremos uma busca para encontrar o alinhamento ótimo entre elas.

As setas da figura 22 ajuda a visualizar a busca pela construção do processo do alinhamento ótimo. O primeiro passo é começar lendo pelo fim da matriz, ou seja, a partir da última entrada (m,n) e seguir as setas até que elas cheguem a entrada $(0,0)$. Cada seta corresponde a comparação de uma coluna do alinhamento a ser gerado. Consideremos o seguinte fato:

- Temos uma seta saindo da entrada (i,j).
- A seta aponta para a horizontal: ela corresponde a uma coluna do alinhamento com um *gap* em s, casado com uma base em $t[j]$.
- A seta aponta para a vertical: ela corresponde a uma coluna do alinhamento com um *gap* em t, casado com uma base em $s[i]$.
- A seta aponta na direção diagonal: ela corresponde a uma coluna do alinhamento onde as bases de $s[i]$ e $t[j]$ se casam.

Como já explicamos no início do capítulo, poderão existir vários alinhamentos ótimos entre duas seqüências. Ao percorremos cada entrada da matriz em busca deste alinhamento, poderemos nos deparar com três ramificações diferentes de acordo com a direção das setas. A cada escolha de uma das três ramificações, estaremos gerando um novo alinhamento ótimo.

Por exemplo, na figura 22, iniciando a busca pela entrada $a[4,3]$, só temos um caminho a percorrer, a seta na diagonal que nos leva à $a[3,2]$. Nela, temos outras duas ramificações: uma com a seta para a vertical, em direção à $a[2,2]$, e uma outra seta para a diagonal, em direção à $a[2,1]$. Continuando o trajeto por $a[2,2]$ até o nosso destino $a[0,0]$, geramos o seguinte alinhamento ótimo ilustrado na figura 23.

$$\begin{array}{cccc} A & G & - & C \\ | & | & & \\ A & A & A & C \end{array}$$

Figura 23: Possível alinhamento ótimo obtido através da matriz ilustrada na figura 22, passando por $a[4,3] \rightarrow a[3,2] \rightarrow a[2,2] \rightarrow a[0,0]$.

3.2 Inconsistência no alinhamento de sequências de DNA Mitocondrial

A ideia de caracterizar perfis de sequências de DNAmt humano através da comparação de uma sequência de amostra com uma sequência padrão, a rCRS, tem sido de grande valia pois, desta maneira, provemos de uma linguagem comum para descrever as variações observadas nos perfis de DNAmt humano das populações.

As regras de nomenclatura, para anotar as diferenças entre uma sequência de amostra e a sequência rCRS, já foram descritas nas seções 2.7 e 2.12. Mas, enquanto estas recomendações de nomenclatura atendem a maioria das situações encontradas em estudos populacionais da região controle do DNAmt, ainda assim existem algumas outras situações mais complexas onde será preciso um tratamento mais explícito no alinhamento de suas sequências (BUDOWLE; DIZINNO; WILSON, 1999; TULLY et al., 2001).

O alinhamento entre sequências de DNAmt, relativas a rCRS, deverá ser padronizado para que amostras idênticas sejam analisadas da mesma forma senão diferentes laboratórios poderão caracterizar diferentemente uma mesma sequência.

Mesmo com a metodologia utilizada no alinhamento de sequências, já consolidada para a geração de variações entre elas, ainda existem situações onde poderão ocorrer alternativas de inserções e deleções, provocando a inconsistência na anotação dos polimorfismos entre as sequências (SALAS; LAREU; CARRACEDO, 2001).

Em Wilson et al. (2002) são propostas recomendações adicionais para garantir a consistência no alinhamento de sequências de DNAmt para sua utilização em aplicações de bancos de dados forenses e na comparação de perfis. É abordado um alinhamento baseado no contexto filogenético, utilizando diferentes prioridades para transições, transversões, inserções e deleções, mesmo que estes não reflitam o seu mecânismo biológico natural. As complicações começam a surgir quando encontramos dois ou mais alinhamentos ótimos.

3.2.1 Problemáticas do alinhamento

A seguir apresentaremos um número de exemplos que passam por uma leve revisão na estratégia do alinhamento de sequências do DNAmt com a sequência padrão, a rCRS. São observadas diversas instâncias do alinhamento onde cada uma delas apresenta uma diferente caracterização para o perfil de DNAmt da sequência em análise (WILSON et al., 2002).

Exemplo 1

Uma pequena repetição de binucleotídeos é encontrada na região controle do DNAmt humano próximo ao gene tRNA-*phenylalanine* (BODENTEICH et al., 1992). A rCRS contêm cinco repetições de AC, e indivíduos na população apresentaram ter de três a sete repetições. Neste exemplo 1, verificamos uma variação de transição de G-A na posição 513.

Amostra:	A C C C A A C A C A C A C A C C G C T G		
rCRS:	A C C C A G C A C A C A C A C A C C G C T G		
Posição:		510	520

Exemplo 1: sequências a serem alinhadas.

O alinhamento 1a ilustra um possível alinhamento entre as duas sequências indicando a deleção do binucleotídeo AC.

Amostra:	A C C C A A C A C A C A C A C C - - G C T G		
rCRS:	A C C C A G C A C A C A C A C A C C G C T G		
Posição:		510	520

Alinhamento 1a, com tres diferenças: 513A, 521D, 522D.

Pode-se eliminar a transição inserindo dois *gaps* na sequência de amostra de acordo com as recomendações 1 e 2, gerando assim o alinhamento 1b.

Amostra:	A C C C A - - A C A C A C A C A C C G C T G		
rCRS:	A C C C A G C A C A C A C A C A C C G C T G		
Posição:		510	520

Alinhamento 1b, com duas diferenças: 513D, 514D.

O alinhamento 1a perde logo na 1° recomendação com uma diferença a mais ao alinhamento 1b, sendo então, este último, o alinhamento mais adequado, com as suas diferenças anotadas: 513D, 514D.

Exemplo 2

A variação de bases dentre uma região de repetição afeta consideravelmente a consistência do alinhamento. Por exemplo, o alinhamento 2 ilustra uma transição C-T na posição 514 que precede as repetições AC.

Amostra:	A C C C A G T A C A C A C A C C G C T G		
rCRS:	A C C C A G C A C A C A C A C A C G C T G		
Posição:		510	520

Exemplo 2: sequencias a serem alinhadas.

O alinhamento 2a resulta em duas deleções: 514D, 515D, seguido de uma transição 516T.

Amostra:	A C C C A G – – T A C A C A C A C C G C T G		
rCRS:	A C C C A G C A C A C A C A C A C C G C T G		
Posição:		510	520

Alinhamento 2a, com tres diferenças: 514D, 515D, 516T.

Neste caso, utilizando-se as recomendações 1 e 2, o T poderá ser posicionado em ambos os lados da repetição. Posicionando-o antes das deleções, estas passam a fazer parte da repetição AC e assim, a recomendação 3 é aplicada no alinhamento 2b. Caso a transição seja posicionada depois das deleções, estas então passam a não ser encontradas dentro da repetição e a recomendação 3 não poderá ser aplicada (alinhamento 2a).

Amostra:	A C C C A G T – – A C A C A C A C C G C T G		
rCRS:	A C C C A G C A C A C A C A C A C C G C T G		
Posição:		510	520

Alinhamento 2b, com três diferenças: 514T, 515D, 516D.

De acordo com a recomendação 3, as inserções e deleções deverão ser posicionadas o quanto possível ao fim da 3´ mantendo o mesmo número de diferenças em relação a

rCRS. Portanto, o alinhamento 2c é o preferido, com suas diferenças anotadas: 514T, 523D, 524D.

Amostra:	A C C C A G T A C A C A C A C – – C G C T G		
rCRS:	A C C C A G C A C A C A C A C A C C G C T G		
Posição:		510	520

Alinhamento 2c, com três diferenças: 514T, 523D, 524D.

Exemplo 3

Neste exemplo, observa-se uma transição T-C na posição 16189, dentre o comprimento de citosinas (poly-C) e a substituição de uma transversão A-C na posição 16183.

Amostra:	A A A C C C C C C C T C C C A T G C T		
rCRS:	A A A A C C C C C T C C C C A T G C T		
Posição:		16180	16190

Exemplo 3: seqüências a serem alinhadas.

Um possível alinhamento deste perfil, em relação a rCRS, mantêm os dois T alinhados no centro em meio ao poly-C, como ilustra o alinhamento 3a.

Amostra:	A A A C C C C C C C C T C C C – A T G C T		
rCRS:	A A A A C C C C C – – T C C C C A T G C T		
Posição:		16180	16190

Alinhamento 3a, com quatro diferenças: 16183C, 16189D, 16190D, 16195T.

O alinhamento 3a resulta em quatro diferenças, uma transversão e três *indels*. Já o alinhamento 3b resulta em apenas três diferenças, sendo este o mais adequado.

Amostra:	A A A C C C C C C C T C C C A T G C T		
rCRS:	A A A A C C C C C C T C – C C C A T G C T		
Posição:		16180	16190

Alinhamento 3b, com tres diferenças: 16183C, 16189C, 16190.1T.

Portanto, as diferenças geradas a partir do alinhamento 3b são uma transversão, uma transição e uma inserção: 16183C, 16189C, 16190.1T .

Exemplo 4

Neste exemplo, temos a ocorrência de uma transição C-T na posição 16179 que precede a primeira base A na posição 16180.

Amostra:	T T A A C C C C C T C C C C C A T G C T		
rCRS:	T C A A A A C C C C C T C C C C A T G C T		
Posição:		16180	16190

Exemplo 4: sequencias a serem alinhadas.

Veremos que existem duas maneiras de alinhar essa sequência, em relação a rCRS, com o mesmo número de diferenças.

Amostra:	T T A A – – C C C C C T C C C C C A T G C T		
rCRS:	T C A A A A C C C C C T C C C C – A T G C T		
Posição:		16180	16190

Alinhamento 4a, com quatro diferenças: 16179T, 16182D, 16183D, 16193.1C.

As quatro diferenças no alinhamento 4a são uma transição 16179T, duas deleções 16182D, 16183D e uma inserção 16193.1C .

Amostra:	T – – T A A C C C C C T C C C C C A T G C T		
rCRS:	T C A A A A C C C C C T C C C C – A T G C T		
Posição:		16180	16190

Alinhamento 4b, com quatro diferenças: 16181T, 16179D, 16180D, 16193.1C.

O alinhamento 4b resulta em uma transversão A-T (16181T), duas deleções (16179D e 16180D) e uma inserção (16193.1C). Já o alinhamento 4a possui uma transição (16179T) ao invés de uma transversão (16181T), como é visto 4b. Portanto, de acordo com a recomendação 2, o alinhamento 4a é o mais adequado.

Exemplo 5

Considere as duas seqüências abaixo:

Amostra:	A A A C C C C C C C C C G C	
rCRS:	A A A C C C C C C C T C C C C C G C	
Posição:	300	310

Exemplo 5: sequencias a serem alinhadas.

Temos as seguintes estratégias de alinhamento:

Amostra:	A A A C C C C C C C C – – – – G C	
rCRS:	A A A C C C C C C C C T C C C C C G C	
Posição:	300	310

Alinhamento 5a, com cinco diferenças: 309C, 311D, 312D, 313D, 314D.

Amostra:	A A A C C C C C C C C – – – – C C G C	
rCRS:	A A A C C C C C C C C T C C C C C G C	
Posição:	300	310

Alinhamento 5b, com quatro diferenças: 309D, 310D, 311D, 312D.

Amostra:	A A A C C C C – – – – C C C C C G C	
rCRS:	A A A C C C C C C C C T C C C C C G C	
Posição:	300	310

Alinhamento 5c, com quatro diferenças: 306D, 307D, 308D, 309D.

A princípio, o alinhamento 5a aparenta ser o mais adequado com cinco diferenças em relação a rCRS, com uma transição T-C (310) e mais quatro deleções da posição 312 até a 314.

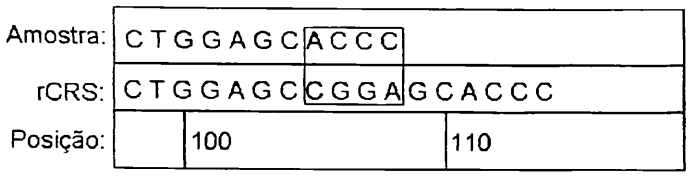
Porém, existe a possibilidade de outros alinhamentos que resultam em apenas quatro diferenças, 5b e 5c.

O alinhamento 5b organiza suas deleções de forma contígua no fim da 3′ após o T deletado na posição 310. Outras estratégias de alinhamento arrumariam as deleções de uma forma não contígua e portanto, são descartadas.

Já o alinhamento 5c não organiza suas deleções contíguas ao fim da 3´ como é exigido pela recomendação 3, sendo então o alinhamento 5b o mais adequado.

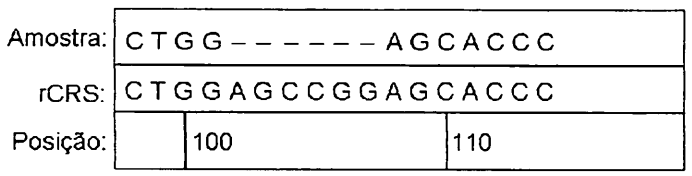
Exemplo 6

Algumas seqüências de DNAmt exibem até seis deleções de pares de base na região controle em HVII. Esta região inicia na posição 98 e vai até a posição 114, como é ilustrado abaixo.

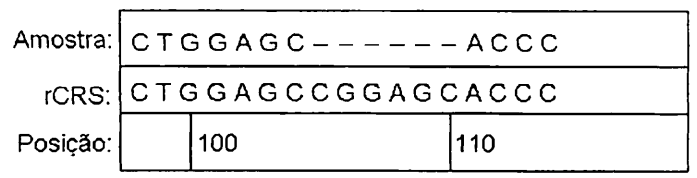


Exemplo 6: sequencias a serem alinhadas.

Existem apenas duas formas potenciais de alinhamento como ilustram os alinhamentos 6a e 6b.



Alinhamento 6a, com seis diferenças: 102D, 103D, 104D, 105D, 106D, 107D.



Alinhamento 6b, com seis diferenças: 105D, 106D, 107D, 108D, 109D, 110D.

O alinhamento 6a resulta em um total de seis diferenças. No entanto, o G pode ser organizado em ambas as extremidades das deleções contíguas ou até mesmo no meio. Porém, as deleções podem ser movidas para o fim da 3´ gerando o alinhamento 6b, sendo este último o mais adequado.

Exemplo 7

Em alguns casos as deleções demonstradas no exemplo 6 poderão ser diferentes. Duas substituições de pares de base TC nas posições 96-97 é acompanhada de seis deleções, levemente modificadas do exemplo 6. As sequências abaixo iniciam na posição 96 e terminam na posição 114 .

Amostra:	A G A T C C T G G A G C C C C C		
rCRS:	A G A C G C T G G A G C C G G A G C A C C C		
Posição:		100	110

Exemplo 7: sequências a serem alinhadas.

Na primeira inspeção, o alinhamento 7a poderá aparentar ser o melhor alinhamento de acordo com as recomendações. Porém, as *indels* são preferidas em relação a transições e transversões. Assim, o alinhamento 7b resulta em dois *indels* ao invés de duas substituições, como é ilustrado abaixo.

Amostra:	A G A T C C T G G A G C C – – – – – – C C C		
rCRS:	A G A C G C T G G A G C C C G G A G C A C C C		
Posição:		100	110

Alinhamento 7a, com oito diferenças: 96T, 97C, 106D, 107D, 108D, 109D, 110D, 111D.

Amostra:	A G A T C – C T G G A G C C – – – – – – C C C		
rCRS:	A G A – C G C T G G A G C C G G A G C A C C C		
Posição:		100	110

Alinhamento 7b, com oito diferenças: 95.1T, 97D, 105D, 106D, 107D, 108D, 109D, 110D.

Ao contrário de uma combinação de transição/transversão, o alinhamento 7b resulta na inserção de um T (95.1T), seguido de uma citosina e depois, seguido de uma delação na posição 97. Portanto, o alinhamento 7b é o mais adequado.

As recomendações e os exemplos demonstrados, nesta seção, tem como propósito refinar e padronizar o tratamento de variantes do DNAmt humano para ser utilizado pela a comunidade forense. Estas práticas tem como o objetivo ajudar, a comunidade forense, a manter e gerar banco de dados de alta qualidade e, com isto, possibilitar a comparação de perfis a serem interpretados em estudos de caso forense.

3.2.2 Ferramentas para o alinhamento de sequências

O objetivo desta seção é descrever algumas ferramentas encontradas na literatura que alinham sequências de DNA. Demonstraremos a inviabilidade do uso destas ferramentas em estudos de caso forenses, especialmente em se tratando da análise de sequências do DNAmt em comparação com a sequência de referência, a rCRS.

Por último, ainda demonstraremos que o uso da ferramenta comercial de alinhamento o SeqScape (Applied Biosystems), cujo propósito é a sua utilização específica na análise de sequências forenses, também apresentou deficiência quanto ao alinhamento de sequências de DNAmt, na geração de seus resultados, para serem armazenados em bancos de dados forenses.

A maioria das ferramentas aqui citadas poderão ser encontradas para uso *online* ou para *download* através do portal de bioinformática da *European Bioinformatics Institute* (EMBL-EBI, 2005) (figura 21).

Ao acessar a seção *Toolbox*, no portal do EBI, o geneticista se depara com uma grande variedade de ferramentas. Elas estão organizadas e agrupadas, em diferentes categorias, de acordo com o seu tipo de funcionalidade.

É importante salientar que a categoria que nos interessa, e que está mais relacionada com o foco deste trabalho, é a da "*Sequence Analysis*".

Apesar das demais categorias estarem relacionadas com a predição de funções ou estruturas de proteínas, cuja utilidade não abrange o objetivo deste trabalho; a maioria, senão todas, possui, como base, a função do alinhamento de sequências, o que poderá confundir o usuário final.

As categorias são:

- **Similarity & Homology:** Este grupo apresenta mais de 15 ferramentas cujo propósito é o de comparar sequências de nucleotídeos ou de proteínas contra os diversos bancos de dados, na Internet, em busca de similaridades entre as sequências. Neste grupo, podemos encontrar algumas das ferramentas mais conhecidas como o Blast e o Fasta. Ambas possuem diversas ferramentas para diferentes funcionalidades, por exemplo: o *WU-Blast2* é uma ferramenta de alinhamento local e busca, para encontrar regiões de similaridades de sequências em bancos de dados nucléicos ou de proteínas (figura 25). Já o Fasta apresenta duas ferramentas independentes onde cada uma executa uma funcionalidade diferente como o *Fasta Nucleotide* e o *Fasta*

Protein, dentre muitas outras. Observe que na figura 25 até mesmo um geneticista experiente poderá se passar por um usuário leigo ao se deparar com os diversos tipos de configurações que podem ser estabelecidos para uma determinada tarefa em que ele esteja em busca. Nota-se facilmente que a maioria dessas opções de configuração não estão diretamente relacionadas com os conceitos de biologia, o que poderá dificultar ainda mais o seu manuseio.

- **Protein Functional Analysis:** Este grupo apresenta 8 ferramentas cujos propósitos são o de determinar a funcionalidade de proteínas através de diversos métodos e bancos de dados, bastante conhecidos, como o UniProtKB Swiss-Prot e o UniProtKB TrEMBL. Uma das ferramentas de destaque é a *InterProScan* que reune vários métodos através de uma interface "amigável" (figura 26).

- **Proteomic Services:** Apresenta duas ferramentas para visualizar informações de seqüências de proteína e outra para os pesquisadores poderem demonstrar os resultados de suas pesquisas no contexto de anotação do UniProt.

- **Sequence Analysis:** É nesta categoria que se encontram as ferramentas, propriamente ditas, para alinhar seqüências de DNA Proteína, com o objetivo de elucidar o grau de similaridade entre elas e a sua origem evolucionária, tal como faz o *ClustalW* (figura 27), cujo enfoque é na utilização de alinhamentos múltiplos, de 2 a 3 seqüências ou mais. Encontramos, também, ferramentas para determinar estrutura de genes e os tipos de proteínas que eles codificam, tal como faz o *Transeq*. Neste grupo, dentre as 15 ferramentas de análises disponíveis apenas uma delas, a Align-EMBOSS, está voltada para o alinhamento local e global entre duas seqüências. Sendo portanto, esta última ferramenta, a única apropriada para alinhar seqüências de amostras de DNAmt, com a rCRS, em estudos de casos forenses.

- **Structural Analysis:** Este grupo apresenta as ferramentas para a determinação de estruturas 2D 3D das proteínas, o que facilita, desta forma, o estudo da sua funcionalidade.

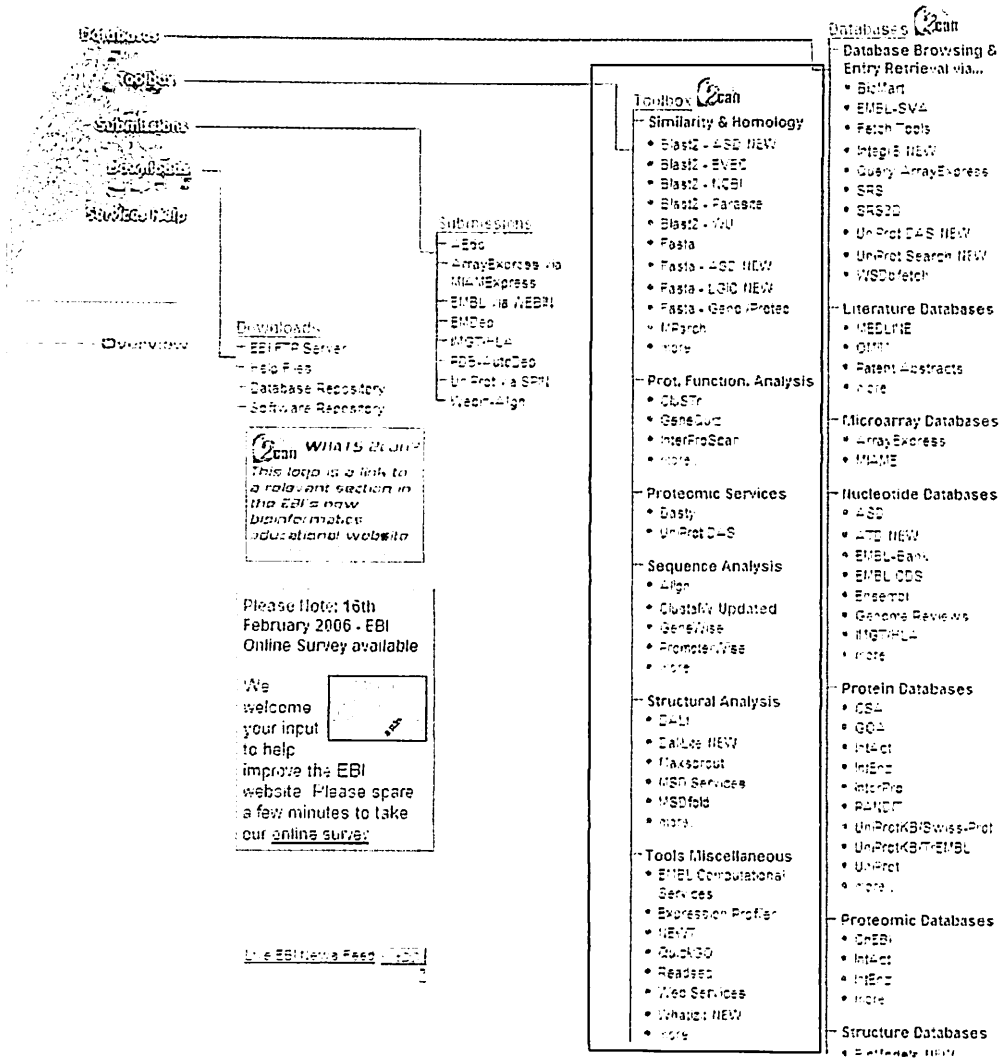


Figura 24: Serviços de Bioinformática oferecidos pelo EBI.

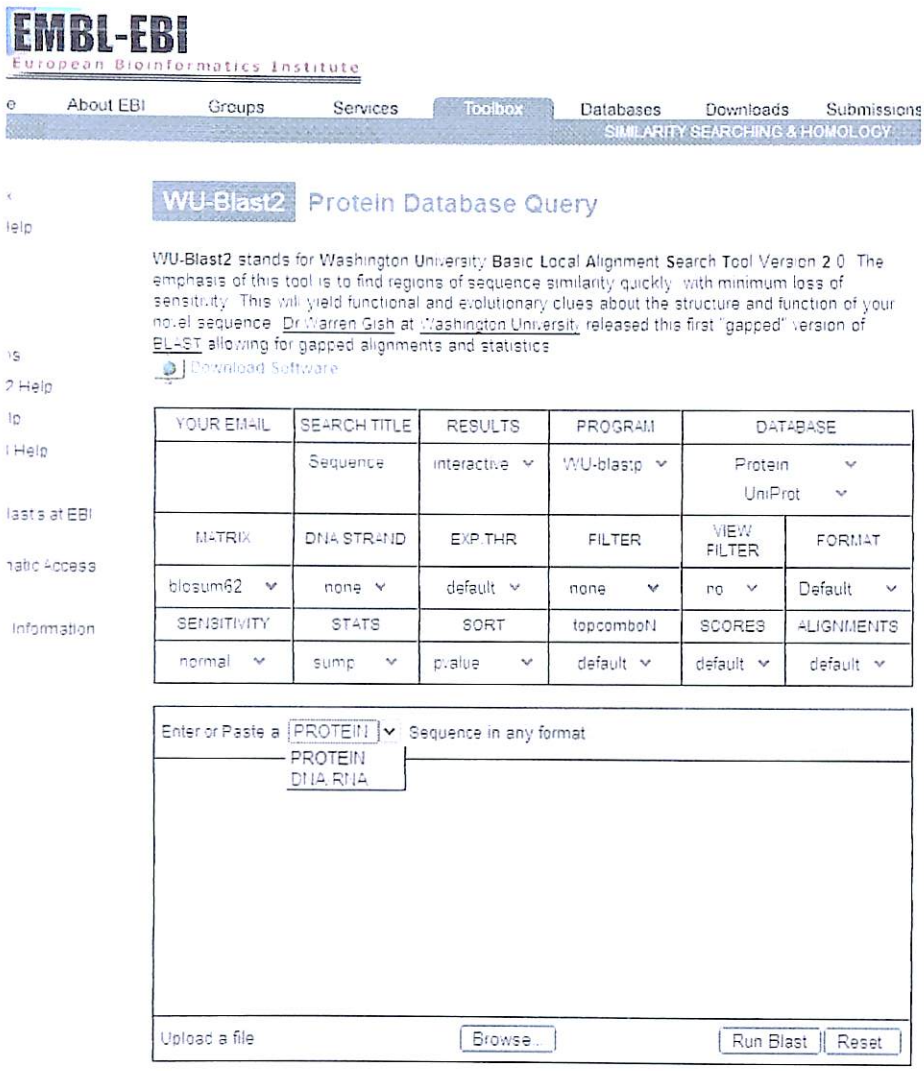


Figura 25: Interface da ferramenta WU-Blast2 com suas diversas opções de configuração do alinhamento, entre DNA ou proteínas, para a busca de similaridades nos bancos de dados.

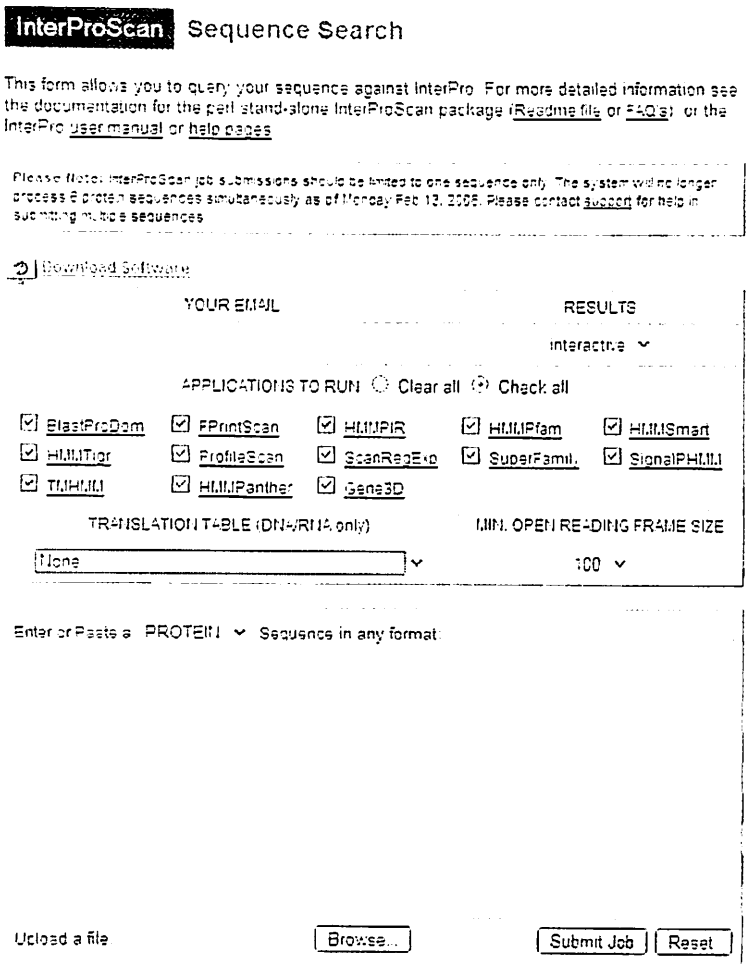


Figura 26: Interface "amigável" do InterProScan.

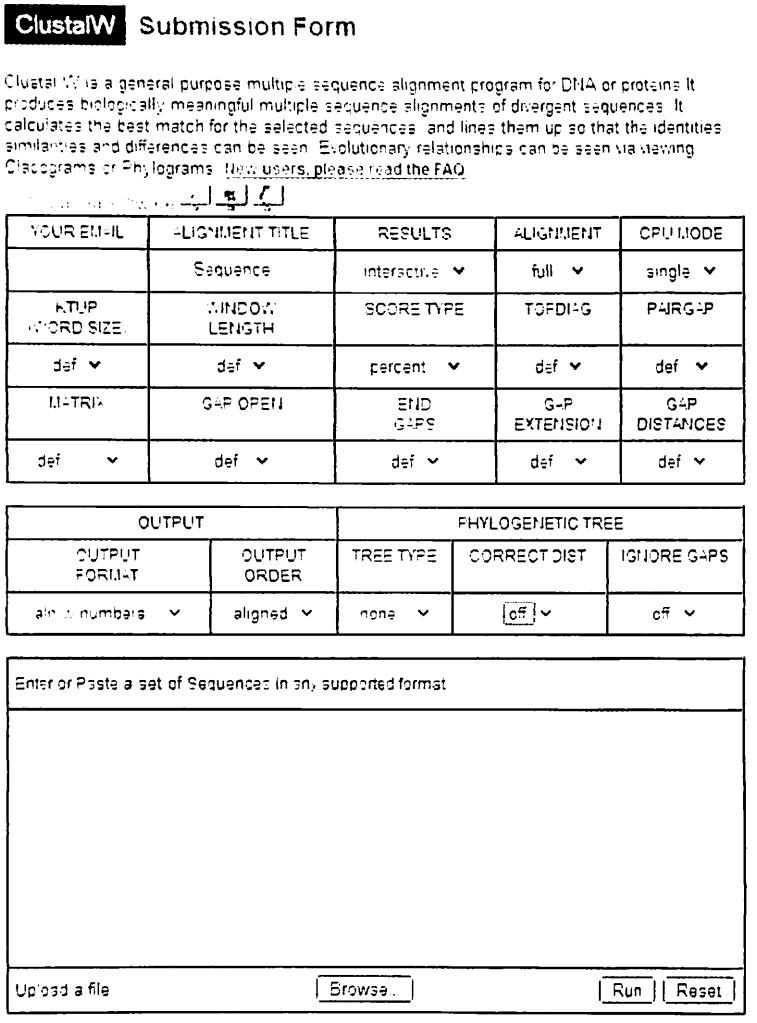


Figura 27: Interface grática do ClustalW para o alinhamento de múltiplas sequências, com suas diversas opções de configuração do alinhamento.

3.2.3 Alinhando sequências nucleotídeas do DNA Mitocondrial

A seguir, utilizaremos como exemplo duas ferramentas de análises para alinhar sequências de DNAmt humano como é realizado em estudos de casos forenses. O objetivo é provar que ainda não existem ferramentas de análises para alinhar sequências de DNAmt humano, em estudos de casos forenses, de acordo com as três recomendações (seção 2.12), como foi apresentado nos exemplos da seção anteiror.

O simples fato de existirem diversas ferramentas deste gênero, disponíveis gratuitamente *(online* ou *standalone)* e até mesmo comerciais, já significa uma enorme preocupação para a comunidade forense, pois, o estudo do DNAmt necessita da padronização e da comunicação comum entre os laboratórios para que se possa obter êxito em suas análises.

O risco de se produzir diferentes perfis de DNAmt, sendo eles originados de uma mesma amostra, é muito grande quando um ou mais laboratórios utilizam diferentes ferramentas de análises para alinhar suas sequências. Esta preocupação se dá pelo fato de que as diversas ferramentas de análises são genéricas e não específicas para a utilização em casos forenses. Por serem genéricas, elas poderão apresentar inúmeras opções de configuração e diferentes formas de utilização, provocando a ocorrência de uma ou várias das 5 classes de erros, como é descrito na seção 2.13.

Antes de iniciarmos os testes com as duas ferramentas, vamos determinar a amostra e suas sequências que serão utilizadas. O penúltimo item aponta para o resultado final do perfil, desta amostra, o qual objetivamos atingir:

- A identidade da amostra é a LUC-01.
- Suas sequências são:

```
TTCTTTCATGGGGAAGCAGATTTGGGTAC
CACCCAAGTATTGACTCACCCATCAACAA
CCGCTATGTATTTCGTACATTACTGCCAG
TCACCATGAATATTGTACGGTACCATAAA
TACTTGACCACCTGTAGTACATAAAAACC
CAATCCACATCAAAACCCCCTCCCCATGC
TTACAAGCAAGCACAGCAATCAACCTTCA
ACTATCACACATCAACTGCAACTCCAAAG
CCACCCCCTCACCCACTAGGATACCAACAA
ACCTATCCACCCTTAACAGTACATAGTAC
ATAAAACCATTTACCGTACATAGCACATT
ACAGTCAAATCCCTTCTCGCCCC
```

HVI

TGTGCACGCGATAGCATTGCGAGACGCTGG
AGCCGGAGCACCCTATGTCGCAGTATCTGT
CTTTGATTCCTGCCCCATCCTGTTATTTAT
CGCACCTACGTTCAATATTACAGGCGAACA
TACTTACTAAAGTGTGTTGATTAATTAATG
CTTGTAGGACATAGTAATAACAATTGAATG
TCTGCACAGCCGCTTTCCACACAGACATCA
TAACAAAAAATTTCCACCAAACCCCCCCT
CCCCCCGCTTCTGGCCACAGCACTTAAACAC

HVII

- O resultado final do perfil deverá possuir o seguinte haplótipo:

HVI - 16111T, 16209C, 16223T, 16290T, 16319A, 16362C
HVII - 73G, 146C, 153G, 210G, 235G, 263G, 309.1C, 315.1C

- A sequência rCRS poderá ser obtida através da referência: Brandon et al. (2001).

3.2.3.1 A ferramenta de análise Align-EMBOSS

Ao acessar a interface gráfica desta ferramenta (figura 28), o usuário geneticista forense irá se deparar com 5 opções de configuração: *Method, Gap Open, Gap Extend, Molecule* e *Martrix*). Existem duas caixas de texto para inserir cada uma da sequência de amostra e da rCRS.

A configuração de *Method* disponibiliza duas opções para a escolha do tipo de método do alinhamento desejado: Local ou Global. Como foi apresentado na seção 3.1.3, o alinhamento local tem o propósito de caracterizar o alinhamento de sequências de uma maneira oposta ao alinhamento global. A partir daqui, o usuário já poderá cometer erros, dependendo do tipo de dado que ele utilizará para representar a sequência *rCRS*. Exitem duas práticas:

1. Representar a sequência rCRS com todas as bases do genoma.
2. Representar a sequência rCRS como sendo duas sequências independentes, uma para HVI (com apenas 342) e outra para HVII (com apenas 268pb).

A primeira prática significa inserir as 16569 pares de bases nucléicas, acarretando em uma maior complexidade do alinhamento e, portanto, acarretando em uma maior utilização do tempo de processamento para sua conclusão.

Comumente, os geneticistas forenses analisam uma sequência de cada vez, primeiro HVI (comprime as posições 73-340) e depois HVII (comprime as posições 16024-16365). Como cada sequência da amostra (HVI e HVII) tem, em média, 300pb (muito menor em relação ao tamanho do genoma mitocondrial), a escolha do método de alinhamento apropriado, para este caso, será o local (figura 29 e 30).

A escolha do método de alinhamento global causará problemas, visto que o mesmo passará a considerar todo o genoma mitocondrial como sendo a região controle de HVI ou HVII, ocasionando um resultado, em média, com 95% de gaps (figura 31).

A segunda prática significa inserir em torno de 16270pb a menos do que a primeira, resultando em um ganho considerável do custo e do tempo final de processamento do alinhamento.

O método de alinhamento adequado a escolher nesta situação é o global, pois agora trabalhamos especificamente com as duas regiões, HVI e HVII, da sequência de amostra, relativas a da sequência de referência (rCRS), objetivando comparar suas similaridades por inteiro. Porém, nem o método global e nem o local resultará em um alinhamento satisfatório, visto que a ferramenta falha ao determinar corretamente a numeração das bases das devidas regiões de HVI e HVII (figuras 32 e 33 para o alinhamento local) (figuras 34 e 35 para o alinhamento global).

Foram realizados outros testes de alinhamento com diferentes valores de configuração para *Gap Open*, *Gap Extend* e *Matrix*. No entanto, estas alterações não apresentaram nenhuma diferença no resultado em relação aos seus valores *default*.

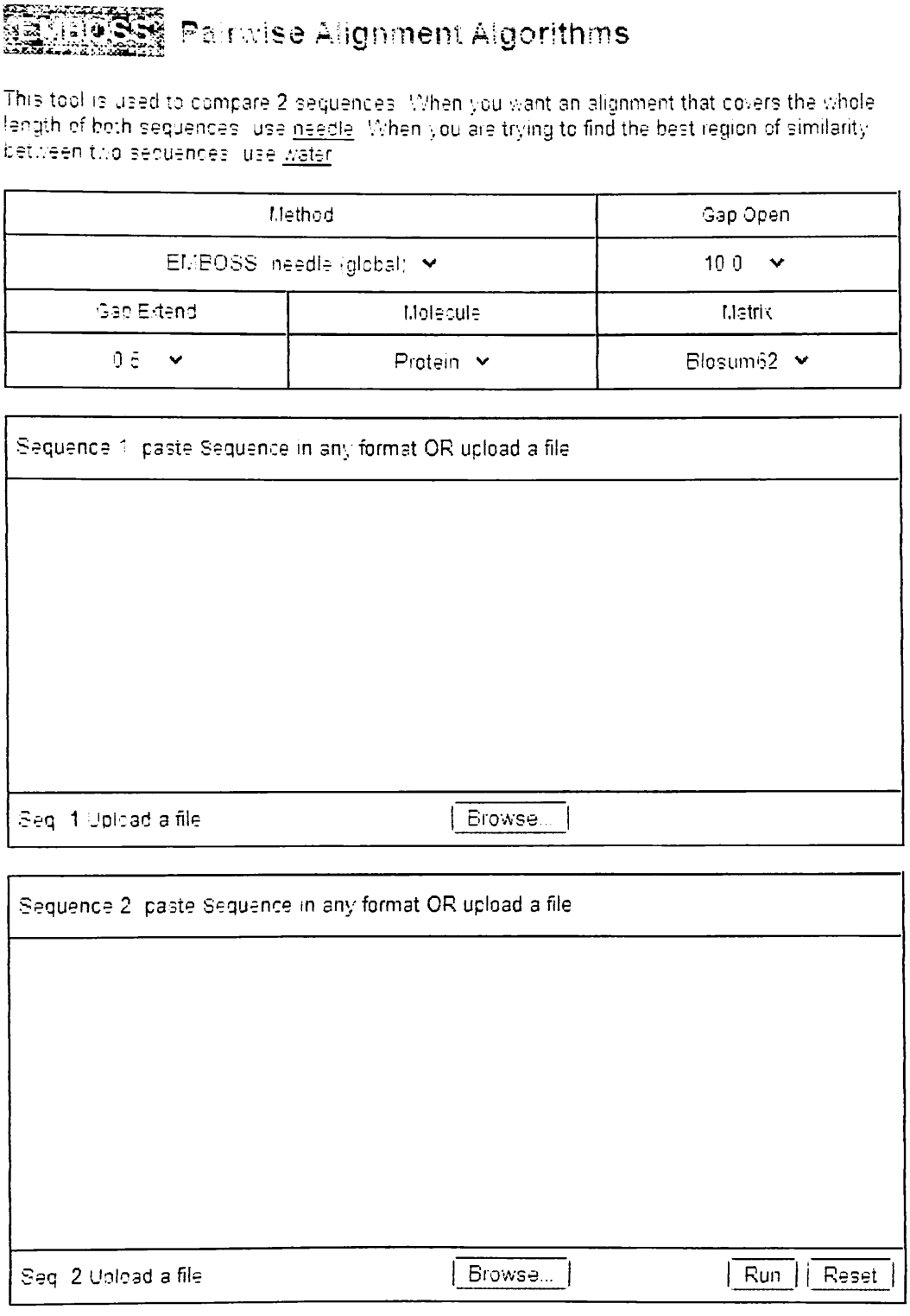


Figura 28: Interface do Align-EMBOSS com suas configurações de alinhamento pré-determinadas para a análise de moléculas de proteína, um grande risco para geração de erros em estudos de casos forenses.

```
########################################
# Program:  water
# Rundate:  Thu Mar 23 17:34:14 2006
# Align_format: srspair
# Report_file: /ebi/extserv/old-work/water-20060323-17341358680340.output
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 342
# Identity:      336/342 (98.2%)
# Similarity:    336/342 (98.2%)
# Gaps:            0/342 ( 0.0%)
# Score: 1656.0
#
#
#=======================================

EMBOSS_001          1 TTCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCACCC     50
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      16024 ttctttcatggggaagcagatttgggtaccacccaagtattgactcaccc 16073

EMBOSS_001         51 ATCAACAACCGCTATGTATTTCGTACATTACTGCCAGTCACCATGAATAT    100
                      ||||||||||||||||||||||||||||||||||||.||||||||||||
EMBOSS_001      16074 atcaacaaccgctatgtatttcgtacattactgccagccaccatgaatat 16123

EMBOSS_001        101 TGTACGGTACCATAAATACTTGACCACCTGTAGTACATAAAAACCCAATC    150
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      16124 tgtacggtaccataaatacttgaccacctgtagtacataaaaacccaatc 16173

EMBOSS_001        151 CACATCAAAACCCCCTCCCCATGCTTACAAGCAAGCACAGCAATCAACCT    200
                      ||||||||||||||||||||||||||||||||||.||||||||||||.|
EMBOSS_001      16174 cacatcaaaaccccctccccatgcttacaagcaagtacagcaatcaaccct 16223

EMBOSS_001        201 TCAACTATCACACATCAACTGCAACTCCAAAGCCACCCCTCACCCACTAG    250
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      16224 tcaactatcacacatcaactgcaactccaaagccacccctcacccactag 16273

EMBOSS_001        251 GATACCAACAAACCTATCCACCCTTAACAGTACATAGTACATAAAACCAT    300
                      ||||||||||||||||.|||||||||||||||||||||||||.||||||
EMBOSS_001      16274 gataccaacaaacctacccaccccttaacagtacatagtacataaagccat 16323

EMBOSS_001        301 TTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGCCCC    342
                      ||||||||||||||||||||||||||||||||||||.|||
EMBOSS_001      16324 ttaccgtacatagcacattacagtcaaatcccttctcgtccc 16365
```

Figura 29: Alinhamento local para HV1 da amostra LUC-01, contra os 16569pb da rCRS, gerado pelo Align-EMBOSS. Foram mantidos os valores *default* para *Gap Open*, *Gap Extend* e *Matrix*, com *Molecule* para DNA. As entre linhas amarelas destacam a ambiguidade de numeração das bases geradas pela ferramenta, possibilitando ao usuário de cometer os erros das classes 1, 2 e 4, por exemplo: 87T ou 87C (seção 2.13). As entre linhas verdes destacam o alinhamento correto, de acordo com as tres recomendações, produzindo os seguintes polimorfismos para HV1: 16111T, 16209C, 16223T, 16290T, 16319A, 16362C.

```
########################################
# Program:  water
# Rundate:  Thu Mar 23 18:12:39 2006
# Align_format:  srspair
# Report_file: /ebi/extserv/old-work/water-20060323-18123808503683.output
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 271
# Identity:      262/271 (96.7%)
# Similarity:    262/271 (96.7%)
# Gaps:            2/271 ( 0.7%)
# Score: 1271.5
#
#
#=======================================

EMBOSS_001         1 TGTGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCG     50
                     | ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001        72 tatgcacgcgatagcattgcgagacgctggagccggagcaccctatgtcg    121

EMBOSS_001        51 CAGTATCTGTCTTTGATTCCTGCCCCATCCTGTTATTTATCGCACCTACG    100
                     |||||||||||||||||||||||||. ||||| ||||||||||||||||||
EMBOSS_001       122 cagtatctgtctttgattcctgcctcatcctattatttatcgcacctacg    171

EMBOSS_001       101 TTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTGATTAATTAATG    150
                     |||||||||||||||||||||||||||||||||||||. ||||||||||
EMBOSS_001       172 ttcaatattacaggcgaacatacttactaaagtgtgttaattaattaatg    221

EMBOSS_001       151 CTTGTAGGACATAGTAATAACAATTGAATGTCTGCACAGCCGCTTTCCAC    200
                     |||||||||||||. |||||||||||||||||||||||.||||||||
EMBOSS_001       222 cttgtaggacataataataacaattgaatgtctgcacagccactttccac    271

EMBOSS_001       201 ACAGACATCATAACAAAAAATTTCCACCAAACCCCCCCCTCCCCCCGCTT    250
                     |||||||||||||||||||||||||||||||||||||||| . |||||||||
EMBOSS_001       272 acagacatcataacaaaaaatttccaccaaaccccccc--tCCCCCgctt    319

EMBOSS_001       251 CTGGCCACAGCACTTAAACAC    271
                     |||||||||||||||||||
EMBOSS_001       320 ctggccacagcacttaaacac    340
```

Figura 30: Alinhamento local para HVII da amostra LUC-01, contra os 16569pb da rCRS, gerado pelo Align-EMBOSS. Foram mantidos os valores *default* para *Gap Open*, *Gap Extend* e *Matrix*, com *Molecule* para DNA. As entre linhas amarelas destacam a ambiguidade de numeração das bases geradas pela ferramenta, possibilitando ao usuário de cometer os erros das classes 1, 2 e 4, por exemplo: 202G ou 202A (seção 2.13). As entre linhas verdes destacam o alinhamento correto e as entre linhas vermelhas os incorretos, de acordo com as três recomendações, produzindo os seguintes polimorfismos para HVII: 73G, 146C, 153G, 210G, 235G, 263G, 309.1C, 309.2T, 310C.

```
########################################
# Program:  needle
# Rundate:  Thu Mar 23 18:45:13 2006
# Align_format: srspair
# Report_file: /ebi/extserv/old-work/needle-20060323-18451162865055.output
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 16569
# Identity:      336/16569 ( 2.0%)
# Similarity:    336/16569 ( 2.0%)
# Gaps:        16227/16569 (97.9%)
# Score: 1656.0
#
#
#=======================================

EMBOSS_001         1 gatcacaggtctatcaccctattaaccactcacgggagctctccatgcat     50

EMBOSS_001         1                                                      0

EMBOSS_001        51 ttggtattttcgtctgggggggtatgcacgcgatagcattgcgagacgctg    100

EMBOSS_001         1                                                      0

EMBOSS_001       101 gagccggagcaccctatgtcgcagtatctgtctttgattcctgcctcatc    150

EMBOSS_001         1                                                      0

EMBOSS_001       151 ctattatttatcgcacctacgttcaatattacaggcgaacatacttacta    200

EMBOSS_001         1                                                      0

EMBOSS_001       201 aagtgtgttaattaattaatgcttgtaggacataataataacaattgaat    250

EMBOSS_001         1                                                      0
```

Figura 31: Alinhamento global para HVI da amostra LUC-01, contra os 16569pb da rCRS, gerado pelo Align-EMBOSS. Foram mantidos os valores *default* para *Gap Open*, *Gap Extend* e *Matrix*, com *Molecule* para DNA. A entre linha vermelha destaca a enorme quantidade de *gaps* inserido no alinhamento, o que não vai de acordo com as três recomendações. O alinhamento não pôde ser exibido por inteiro pois ultrapassou as dimensões do monitor, devido aos 16569pb alinhados. O mesmo experimento foi realizado para HVII da amostra LUC-01 (com as mesmas configurações), apresentando uma taxa de 98% de inserções de *gaps* no alinhamento.

```
########################################
# Program:  water
# Rundate:  Thu Mar 23 19:29:23 2006
# Align_format: srspair
# Report_file: /ebi/extserv/old-work/water-20060323-19292205141118.output
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 342
# Identity       336/342 (98.2%)
# Similarity:    336/342 (98.2%)
# Gaps:            0/342 ( 0.0%)
# Score: 1656.0
#
#
#=======================================

EMBOSS_001         1 TTCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCACCC     50
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001         1 ttctttcatggggaagcagatttgggtaccacccaagtattgactcaccc     50

EMBOSS_001        51 ATCAACAACCGCTATGTATTTCGTACATTACTGCCAGTCACCATGAATAT     100
                     |||||||||||||||||||||||||||||||||||.|||||||||||||
EMBOSS_001        51 atcaacaaccgctatgtatttcgtacattactgccagccaccatgaatat     100

EMBOSS_001       101 TGTACGGTACCATAAATACTTGACCACCTGTAGTACATAAAAACCCAATC     150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       101 tgtacggtaccataaatacttgaccacctgtagtacataaaaacccaatc     150

EMBOSS_001       151 CACATCAAAACCCCCTCCCCATGCTTACAAGCAAGCACAGCAATCAACCT     200
                     ||||||||||||||||||||||||||||||||||.||||||||||||.
EMBOSS_001       151 cacatcaaaaccccctccccatgcttacaagcaagtacagcaatcaaccc     200

EMBOSS_001       201 TCAACTATCACACATCAACTGCAACTCCAAAGCCACCCCTCACCCACTAG     250
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       201 tcaactatcacacatcaactgcaactccaaagccaccctcacccactag     250

EMBOSS_001       251 GATACCAACAAACCTATCCACCCTTAACAGTACATAGTACATAAAGCCAT     300
                     ||||||||||||||||.||||||||||||||||||||||.|||
EMBOSS_001       251 gataccaacaaacctacccacccttaacagtacatagtacataaagccat     300

EMBOSS_001       301 TTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGCCCC     342
                     |||||||||||||||||||||||||||||||||||.||
EMBOSS_001       301 ttaccgtacatagcacattacagtcaaatcccttctcgtccc     342
```

Figura 32: Alinhamento local para HVI da amostra LUC-01 contra HVI da rCRS, gerado pelo Align-EMBOSS. Foram mantidos os valores *default* para *Gap open*, *Gap Extend* e *Matrix*, com *Molecule* para DNA. As entre linhas amarelas destacam a numeração incorreta das bases geradas pela ferramenta, possibilitando ao usuário de cometer os erros das classes 1, 2 e 4 (seção 2.13). As entre linhas verdes destacam o alinhamento correto, de acordo com as três recomendações. Porém, a geração incorreta da numeração das bases resulta em uma falha total do processo de alinhamento, produzindo os seguintes polimorfismos para HVI: 88T, 186C, 200T, 267T, 296A, 339C.

```
########################################
# Program:  water
# Rundate:  Thu Mar 23 19:33:49 2006
# Align_format: srspair
# Report_file: /ebi/extserv/old-work/water-20060323-19334842741681.output
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 269
# Identity:      261/269 (97.0%)
# Similarity:    261/269 (97.0%)
# Gaps:            2/269 ( 0.7%)
# Score: 1270.5
#
#
#=======================================

EMBOSS_001          3 TGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCA     52
                      |||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001          2 tgcacgcgatagcattgcgagacgctggagccggagcaccctatgtcgca     51

EMBOSS_001         53 GTATCTGTCTTTGATTCCTGCCTCATCCTGTTATTTATCGCACCTACGTT    102
                      |||||||||||||||||||||.|||||.||||||||||||||||||||
EMBOSS_001         52 gtatctgtctttgattcctgcctcatcctattatttatcgcacctacgtt    101

EMBOSS_001        103 CAATATTACAGGCGAACATACTTACTAAAGTGTGTTGATTAATTAATGCT    152
                      ||||||||||||||||||||||||||||||||||.||||||||||||||
EMBOSS_001        102 caatattacaggcgaacatacttactaaagtgtgttaattaattaatgct    151

EMBOSS_001        153 TGTAGGACATAGTAATAACAATTGAATGTCTGCACAGCCGCTTTCCACAC    202
                      ||||||||||.||||||||||||||||||||||||||.||||||||||
EMBOSS_001        152 tgtaggacataacaataacaattgaatgtctgcacagccactttccacac    201

EMBOSS_001        203 AGACATCATAACAAAAAATTTCCACCAAACCCCCCCCTCCCCCCGCTTCT    252
                      ||||||||||||||||||||||||||||||||||||  .||||||||||
EMBOSS_001        202 agacatcataacaaaaaatttccaccaaacccccccc--tccccgcttct    249

EMBOSS_001        253 GGCCACAGCACTTAAACAC    271
                      ||||||||||||||||||
EMBOSS_001        250 ggccacagcacttaaacac    268
```

Figura 33: Alinhamento local para HVII da amostra LUC-01 contra HVII da rCRS, gerado pelo Align-EMBOSS. Foram mantidos os valores *default* para *Gap Open*, *Gap Extend* e *Matrix*, com *Molecule* para DNA. As entre linhas amarelas destacam a numeração incorreta das bases geradas pela ferramenta, possibilitando ao usuário de cometer os erros das classes 1, 2 e 4 (seção 2.13). As entre linhas verdes destacam o alinhamento correto e a vermelha o incorreto, de acordo com as três recomendações. Porém, a geração incorreta da numeração das bases resulta em uma falha total do processo de alinhamento.

```
########################################
# Program:  needle
# Rundate:  Thu Mar 23 19:16:27 2006
# Align_format: srspair
# Report_file: /ebi/extserv/old-work/needle-20060323-19162635630612.output
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 342
# Identity:      336/342 (98.2%)
# Similarity:    336/342 (98.2%)
# Gaps:            0/342 ( 0.0%)
# Score: 1656.0
#
#
#=======================================

EMBOSS_001         1 TTCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCACCC     50
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001         1 ttctttcatggggaagcagatttgggtaccacccaagtattgactcaccc     50

EMBOSS_001        51 ATCAACAACCGCTATGTATTTCGTACATTACTGCCAGTCACCATGAATAT    100
                     ||||||||||||||||||||||||||||||||||||.||||||||||||
EMBOSS_001        51 atcaacaaccgctatgtatttcgtacattactgccagccaccatgaatat    100

EMBOSS_001       101 TGTACGGTACCATAAATACTTGACCACCTGTAGTACATAAAAACCCAATC    150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       101 tgtacggtaccataaatacttgaccacctgtagtacataaaaacccaatc    150

EMBOSS_001       151 CACATCAAAACCCCCTCCCCATGCTTACAAGCAAGCACAGCAATCAACCT    200
                     ||||||||||||||||||||||||||||||||||.||||||||||||.
EMBOSS_001       151 cacatcaaaaccccctccccatgcttacaagcaagtacagcaatcaaccc    200

EMBOSS_001       201 TCAACTATCACACATCAACTGCAACTCCAAAGCCACCCCTCACCCACTAG    250
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       201 tcaactatcacacatcaactgcaactccaaagccacccctcacccactag    250

EMBOSS_001       251 GATACCAACAAACCTATCCACCCTTAACAGTACATAGTACATAAAAACCAT    300
                     ||||||||||||||||.||||||||||||||||||||||||.|||
EMBOSS_001       251 gataccaacaaacctaccccaccttaacagtacatagtacataaagccat    300

EMBOSS_001       301 TTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGCCCC    342
                     ||||||||||||||||||||||||||||||||||||.|||
EMBOSS_001       301 ttaccgtacatagcacattacagtcaaatcccttctcgtccc    342
```

Figura 34: Alinhamento global para HVI da amostra LUC-01 contra HVI da rCRS, gerado pelo Align-EMBOSS. Foram mantidos os valores *default* para *Gap Open*, *Gap Extend* e *Matrix*, com *Molecule* para DNA. As entre linhas amarelas destacam a numeração incorreta das bases geradas pela ferramenta, possibilitando ao usuário de cometer os erros das classes 1, 2 e 4 (seção 2.13). As entre linhas verdes destacam o alinhamento correto, de acordo com as três recomendações. Porém, a geração incorreta da numeração das bases resulta em uma falha total do processo de alinhamento, produzindo os seguintes polimorfismos para HVI: 88T, 186C, 200T, 267T, 296A, 339C.

```
########################################
# Program:  needle
# Rundate:  Thu Mar 23 19:24:06 2006
# Align_format: srspair
# Report_file: /ebi/extserv/old-work/needle-20060323-19240687457724.output
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 271
# Identity:      261/271 (96.3%)
# Similarity:    261/271 (96.3%)
# Gaps:            3/271 ( 1.1%)
# Score: 1266.5
#
#
#=======================================

EMBOSS_001         1 TGTGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCG     50
                     .||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001         1  atgcacgcgatagcattgcgagacgctggagccggagcaccctatgtcg     49

EMBOSS_001        51 CAGTATCTGTCTTTGATTCCTGCCTCATCCTGTTATTTATCGCACCTACG    100
                     |||||||||||||||||||||||||.|||||.||||||||||||||||||
EMBOSS_001        50 cagtatctgtctttgattcctgcctcatcctattatttatcgcacctacg     99

EMBOSS_001       101 TTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTGATTAATTAATG    150
                     ||||||||||||||||||||||||||||||||||||||.||||||||||||
EMBOSS_001       100 ttcaatattacaggcgaacatacttactaaagtgtgttaattaattaatg    149

EMBOSS_001       151 CTTGTAGGACATAGTAATAACAATTGAATGTCTGCACAGCCGCTTTCCAC    200
                     |||||||||||||.|||||||||||||||||||||||||.||||||||
EMBOSS_001       150 cttgtaggacataataataacaattgaatgtctgcacagccgctttccac    199

EMBOSS_001       201 ACAGACATCATAACAAAAAATTTCCACCAAACCCCCCCTCCCCCCGCTT    250
                     |||||||||||||||||||||||||||||||||||||   .|||||||||
EMBOSS_001       200 acagacatcataacaaaaaatttccaccaaaccccccc--tcccccgctt    247

EMBOSS_001       251 CTGGCCACAGCACTTAAACAC    271
                     |||||||||||||||||||||
EMBOSS_001       248 ctggccacagcacttaaacac    268
```

Figura 35: Alinhamento global para HVII da amostra LUC-01 contra HVII da rCRS, gerado pelo Align-EMBOSS. Foram mantidos os valores *default* para *Gap Open*, *Gap Extend* e *Matrix*, com *Molecule* para DNA. As entre linhas amarelas destacam a numeração incorreta das bases geradas pela ferramenta, possibilitando ao usuário de cometer os erros das classes 1, 2 e 4 (seção 2.13). As entre linhas verdes destacam o alinhamento correto e a vermelha os incorretos, de acordo com as três recomendações. Porém, a geração incorreta da numeração das bases resulta em uma falha total do processo de alinhamento.

3.2.3.2 A ferramenta de análise SeqScape

Esta ferramenta de análise vem acompanhada do seqüenciador da Applied Biosystems. A grande maioria dos laboratórios de DNA forense utilizam equipamentos da Applied Biosystems, especialmente pelo fato desses equipamentos serem acompanhados de *softwares* voltados para a análise do DNA forense, como o SeqScape.

O SeqScape, como toda ferramenta para fins comerciais, foi modelado com o intuito de facilitar a utilização do processo de análise de seqüências para o usuário final, o geneticista forense. O seu processo de análise de seqüências é realizado em cinco passos. Eles são:

1° passo: Importar os quatro arquivos '.AB1' do seqüenciador, com as seqüências F e R das duas regiões de HVI e HVII. Cada arquivo contêm informações relacionadas ao seqüenciamento de cada região da amostra em estudo, como também as informações visuais do seu eletroferograma (figura 37).

2° passo: Adicionar os quatro arquivos '.AB1' no SeqScape para visualização dos dados.

3° passo: Este envolve o processo de geração da seqüência consenso (figura 10) a partir das duas seqüências F e R, de cada região do D-loop (HVI e HVII). A seguir, cada seqüência consenso de HVI e HVII é alinhada com a rCRS e, em seguida, é gerado o resultado do processo da análise (figura 38).

4° passo: Consiste em visualizar o alinhamento gerado da seqüência de amostra com a rCRS e o seu eletroferograma. Neste passo o usuário poderá fazer uma revisão da qualidade dos dados, observando os picos de cada base do eletroferograma. Ele também poderá inspecionar e editar os polymorfimos gerados pelo alinhamento (figura 39).

5° passo: Nesta última etapa o usuário poderá gerar relatórios com as informações e os resultados da análise, destacando os polimorfismos que foram obtidos (figura 40).

Como exemplo de teste para a ferramenta do SeqScape, utilizaremos os seguintes dados:

- A identidade da amostra é a AL09.
- Suas seqüências são:

```
TTCTTTCATGGGGAAGCAGATTT
GGGTACCACCCAAGTATTGACTC
ACCCATCAACAACCGCTATGTAT
TTCGTACATTACTGCCAGTCACC
ATGAATATTGTACGGTACCATAA
ATACTTGACCACCTGTAGTACAT
AAAAACCCAATCCACATCAAAAC
CCCCTCCCCATGCTTACAAGCAA
GTACAGCAATCAACCTTCAACTA
TCACACATCAACTGCAACTCCAA
AGCCACCCCTCACCCACTAGGAT
ACCAACAAACCTATTCACCCTTA
ACAGTACATAGTACATAAAACCA
TTTACCGT-CATAGCACATTACA
GTCAAATCCCTTCTCGCCCC
```

HVI

```
GTGCACGCGATAGCATTGCGAG
ACGCTGGAGCCGGAGCACCCTA
TGTCGCAGTATCTGTCTTTGAT
TCCTGCCCCATCCTGTTATTTA
TCGCACCTACGTTCAATATTAC
AGGCGAACATACTTACTAAAGT
GTGTTAATTAATTAATGCTTGT
AGGACATAGCAATAACAATTGA
ATGTCTGCACAGCCGCTTTCCA
CACAGACATCATAACAAAAAAT
TTCCACCAAACCCCCCCCCTCC
CCCCGCTTCTGGCCACAGCACT
TAAACAC
```

HVII

- O resultado final do perfil deverá possuir o seguinte haplótipo:

HVI - 16111T, 16223T, 16290T, 16291T, 16319A, 16362C

HVII - 73G, 146C, 153G, 235G, 236C, 263G, 309.1C, 309.2C, 315.1C

- A seqüência rCRS poderá ser obtida através da referência: (BRANDON et al., 2004).

O SeqScape já vem com o genoma completo do DNAmt humano embutido como padrão, possibilitando que o usuário apenas se preocupe em inserir as seqüências de HVI e HVII da amostra. Porém ele utiliza outro genoma mitocondrial como referência, o Yoruba (Africano), indo totalmente de encontro com a prática exercida pela comunidade forense, onde a seqüência de referência utilizada é a rCRS. Assim, o geneticista forense, ao utilizar o SeqScape, precisará entrar com a seqüência rCRS. Caso contrário, suas seqüências de

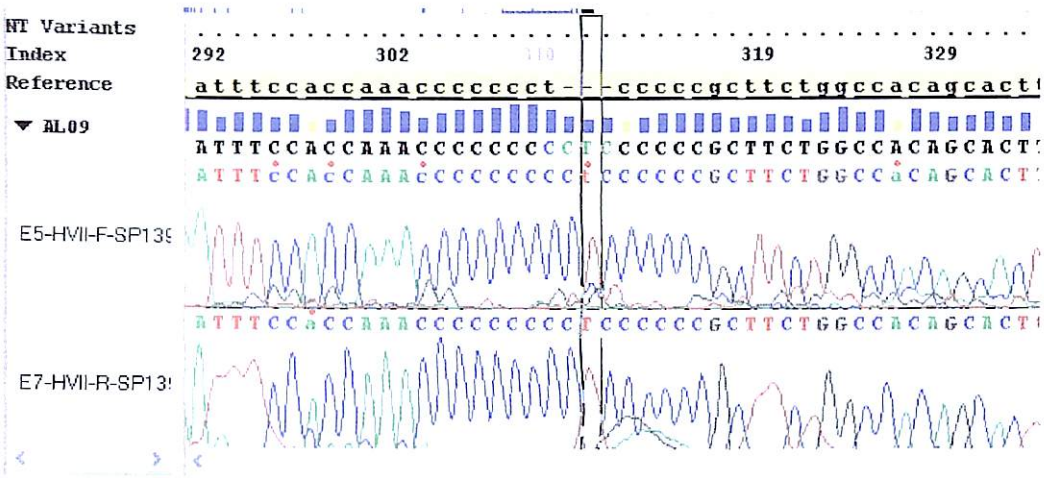


Figura 36: Alinhamento confuso gerado pelo SeqScape da sequência HVII da amostra AL09. O resultado gerado foi o: HVII - 73G, 146C, 153G, 235G, 236C, 263G, 310.1C, 310.2T, 310.3C.

amostra serão analisadas com a Yoruba, gerando resultados inesperados. Da mesma forma que a ferramenta Align-EMBOSS, o SeqScape também obteve sucesso em alinhar a sequência de HVI. No entando, ela também foi incapaz de alinhar a sequência de HVII (figura 36). O alinhamento obteve os polimorfismos '... 310.1C, 310.2T, 310.3C' ao invés dos '... 309.1C, 309.2C, 315.1C'.

Com o fim dos testes, concluímos que ambas as ferramentas (Align-EMBOSS e o SeqScape) são incapazes de alinhar corretamente a sequência de HVII quando há inserções na região do poly-C, não atendendo as três recomendações propostas para o tratamento de variantes do DNAmt humano em estudos de casos forenses (seção 2.12). Mesmo que a ferramenta SeqScape possibilite a edição manual desta falha no alinhamento, esta mesma possibilidade poderá acarretar em novos erros, visto que a falha humana é inerente.

Diferente do Align-EMBOSS, o SeqScape destaca em relatórios os polimorfismos obtidos do resultado do alinhamento. Este mesmo relatório fornece informaçãoes adicionas que não serão utilizados no estudo de casos forenses e a sua anotação dos polimorfismos não está de acordo com as padronizações propostas por (WILSON et al., 2002), como pode ser observado na figura 40.

No final das análises, o usuário terá que anotar à mão ou copiar e colar, em algum editor de texto, cada polimorfismo da amostra para poder compor o seu perfil de DNAmt. Esta prática leva ao grande risco do usuário cometer alguns dos erros das classes 1, 2 e 4.

No capítulo 4 é apresentado uma solução para os problemas aqui tratados, referentes ao alinhamento de sequências de DNAmt humano, para o estudo de casos forenses.

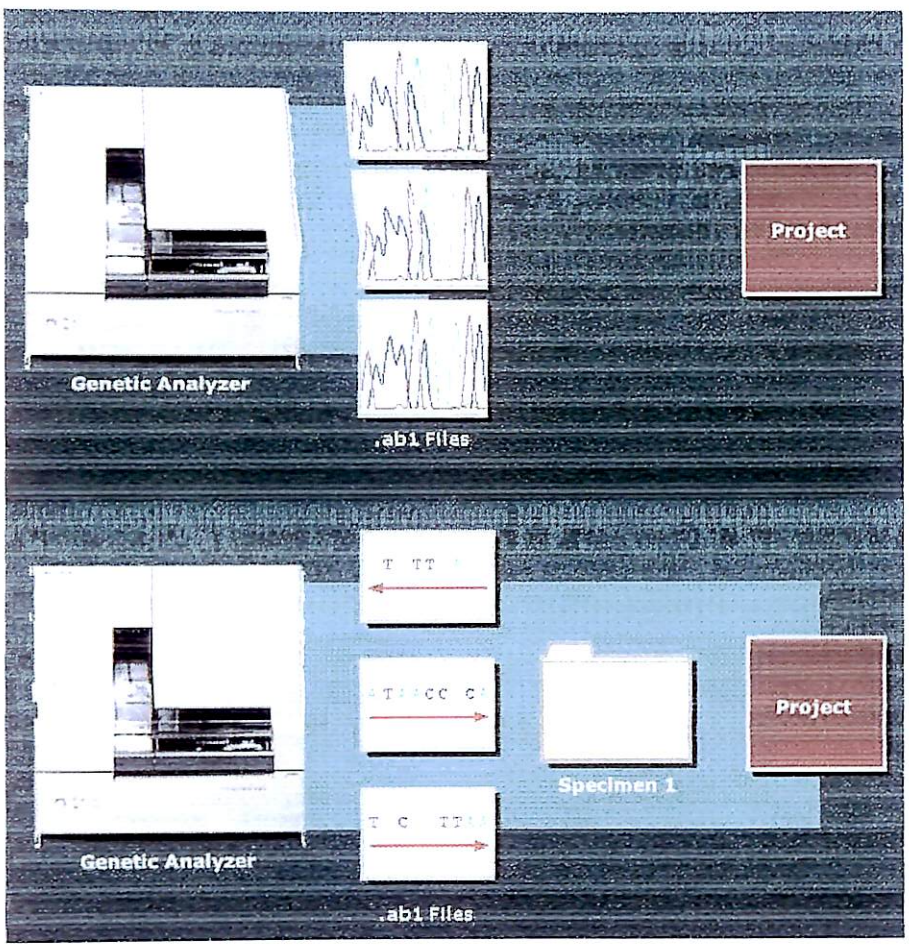


Figura 37: Exemplo ilustrativo de como são gerados e importados os arquivos AB1 do sequênciador, como projeto, para o SeqScape.

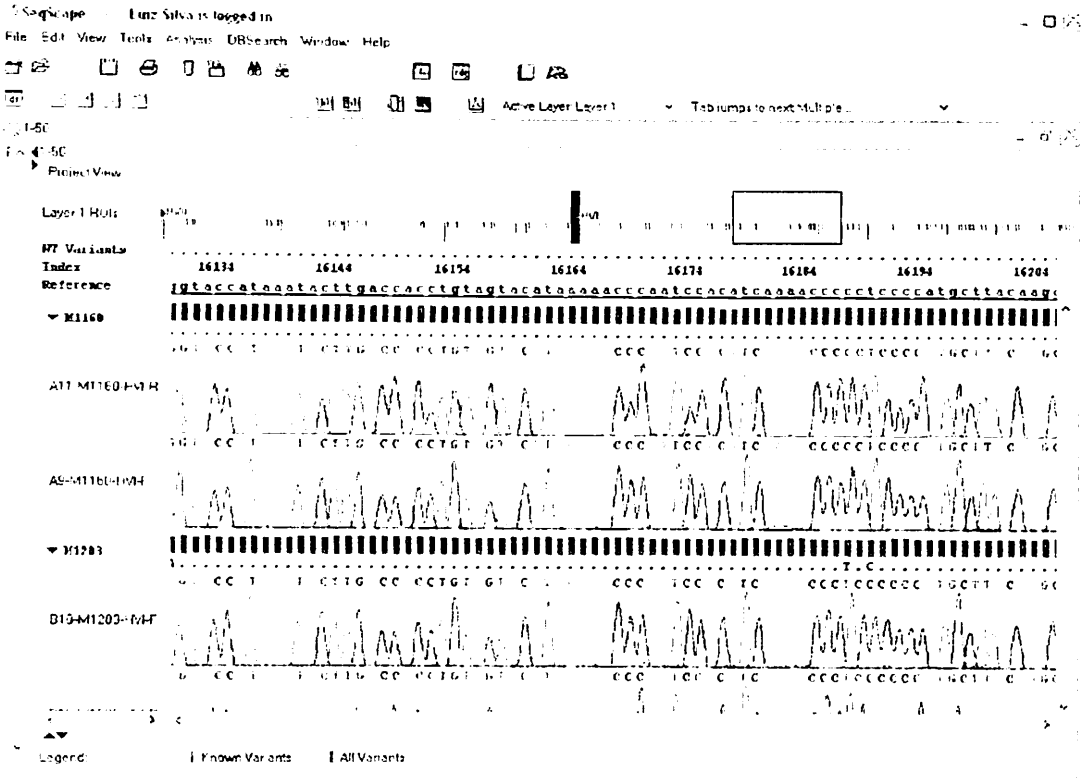


Figura 38: Interface gráfica do SeqScape, demonstrando o resultado da análise de sequências. Observe que podemos visualizar o alinhamento relativo a rCRS e o seu eletroferograma, para que se possa reavaliar a qualidade dos picos das bases das sequências.

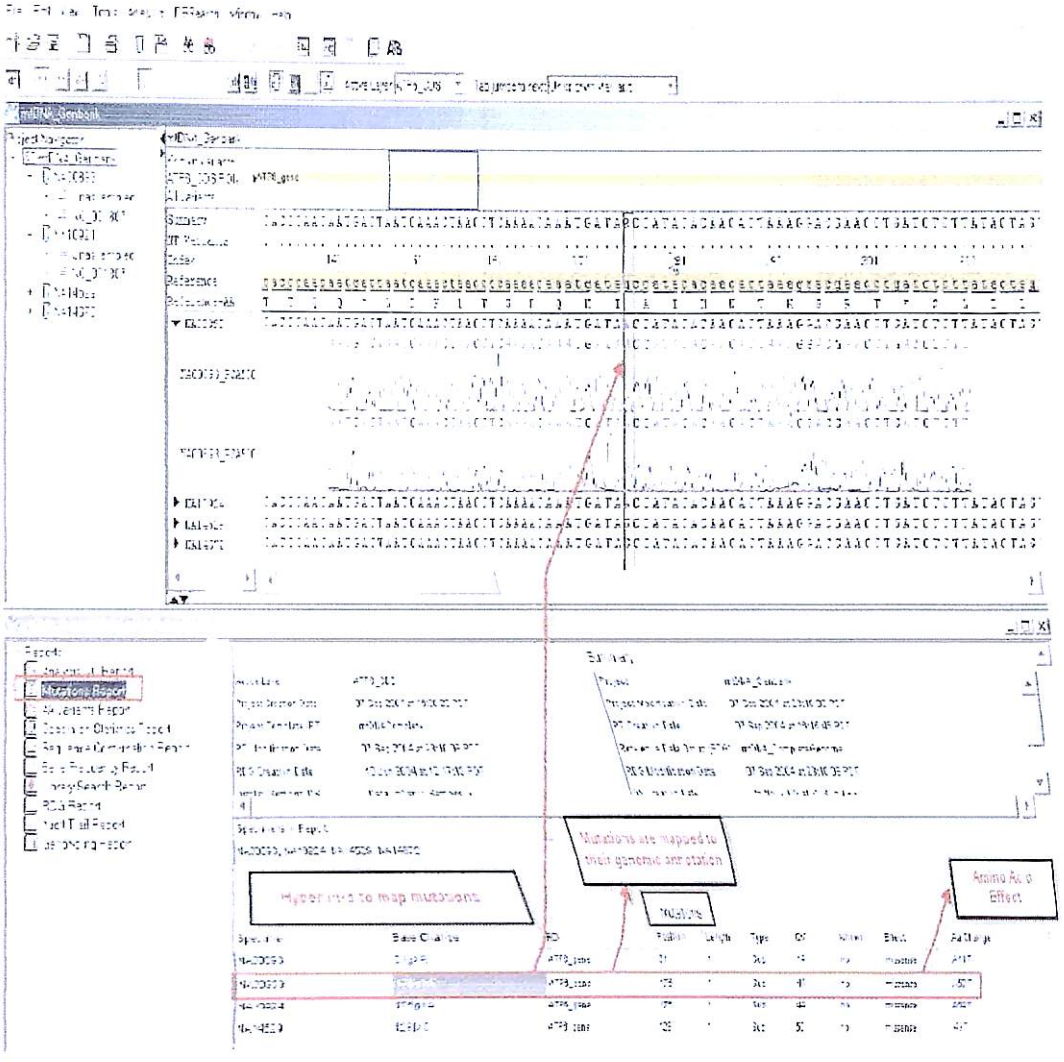


Figura 39: O SeqScape possibilida a edição dos resultados gerado pelo alinhamento.

Applied Biosystems

Generated at 28 Fev 2006 at 13:55:24 GMT-03:00

Summary

Active Layer	Layer 1	Project	KB48-100
Project Creation Date	07 Ago 2005 at 14:03:20 GMT-03:00	Project Modification Date	18 Out 2005 at 15:52:51 GMT-03:00
Project Template (PT)	KB	PT Creation Date	03 Ago 2005 at 15:09:03 GMT-03:00
PT Modification Date	15 Out 2005 at 15:52:51 GMT-03:00	Reference Data Group (RDG)	Mitocondrial_CRSCopy2
RDG Creation Date	07 Ago 2005 at 14:00:02 GMT-03:00	RDG Modification Date	18 Out 2005 at 15:52:51 GMT-03:00
Display Settings (DS)	Mitocondria-HV1	DS Creation Date	02 Mar 2006 at 16:04:50 GMT-03:00
DS Modification Date	15 Out 2005 at 15:52:51 GMT-03:00	Analysis Defaults (AD)	KB
AD Creation Date	03 Ago 2005 at 15:08:12 GMT-03:00	AD Modification Date	18 Out 2005 at 15:52:51 GMT-03:00

Specimens in Report
SP982

Mutations

Specimen	Base Change	ROI	Position	Length	Type	QV	Known	Effect	Aa Change	Description
SP982	73a>G	HV1	73	1	Sub	46	no	missense	T>C	
SP982	146c>T	HV1	146	1	Sub	42	no	silent	-	
SP982	153a>G	HV1	153	1	Sub	41	no	missense	I29V	
SP982	263a>G	HV1	263	1	Sub	36	no	silent	-	
SP982	310c>T	HV1	310	1	Sub	36	no	missense	-	
SP982	311c>T	HV1	311	1	Sub	33	no	missense	-	

For Research Use Only. Not for use in diagnostic procedures

Project Creation: mtDNA Laboratório de DNA Forense
Printed by: seqscape

Page 1

Figura 40: Relatório gerado pelo SeqScape com os dados do resultado da análise.

4 O Sistema Eva

"Computers are to biology what mathematics is to physics."
Harold Morowitz

Este capítulo apresenta uma solução de software, específico, com suporte a gerência, a automatização das análises e ao controle de erros de dados biológicos do DNA Mitochondrial humano, possibilitando a geração de seus perfis genéticos para serem comparados e avaliados em estudos de casos forenses no Brasil ...

4.1 Concepção

A utilização do DNA Mitocondrial como instrumento para a identificação de seres humanos, através de restos mortais encontrados em estado de decomposição, já é uma realidade em países como os Estados Unidos, o Canadá e em diversos outros da Europa, como a Inglaterra, Espanha, Alemanha, dentre outros.

No Brasil esta prática ainda não é muito conhecida e são poucos os laboratórios de DNA forense em atuação, onde apenas um ou outro se destaca pela sua capacitação e perícia no uso de equipamentos e softwares apropriados para diagnosticar e identificar seres humanos, através do estudo do seu DNAn /DNAmt, nas diversas condições e situações em que uma amostra biológica poderá se apresentar, conforme o seu estado físico e integral. Dentre os laboratórios de renome no país podemos citar o laboratório de DNA forense da Universidade Federal de Alagoas[1], onde o desenvolvimento deste trabalho foi realizado em parceria.

Como foi explanado nos capítulos anteriores, a comunidade forense possui uma série de dificuldades para utilizar e analisar o DNAmt. O própio FBI não obteve êxito ao produzir um sistema com um banco de dados populacional para poder estimar a frequência de um dado perfil de DNAmt em questão.

O fator decisivo que contribuiu para a má qualidade do banco populacional do CODIS foi justamente aceitar dados de sequências de vários laboratórios ao redor do mundo, e inseri-los no banco, sem que ao menos estas informações passassem por um controle de qualidade. Isto comprova que diferentes práticas de análises de sequências do DNAmt são utilizadas entre os laboratórios.

Um outro fator se deve a ausência de ferramentas que extraíssem os polymorfimos gerados pelo alinhamento, em formatos padronizados, das duas regiões de HVI e HVII. Esta prática é realizada manualmente pelo geneticista onde ele busca e anota estes resultados, a olho nú, em uma folha de papel, editores de texto ou em planilhas eletronica.

Um hábito bastante comum realizado para armazenar os perfis genéticos de DNAmt é comportar esses dados em formato de tabelas, utilizando a planilha do *software* Excel da Microsoft (JIN et al., 2006). Consequentemente, o método de comparação destas informações será altamente precário onde, possivelmente, deve se usar a opção de "localizar", "buscar", "ir para" ou, até mesmo, realizar a comparação de perfil por perfil manualmente. A figura 41 ilustra bem as etapas desse processo.

[1]URL - `http://www.mhn.ufal.br`

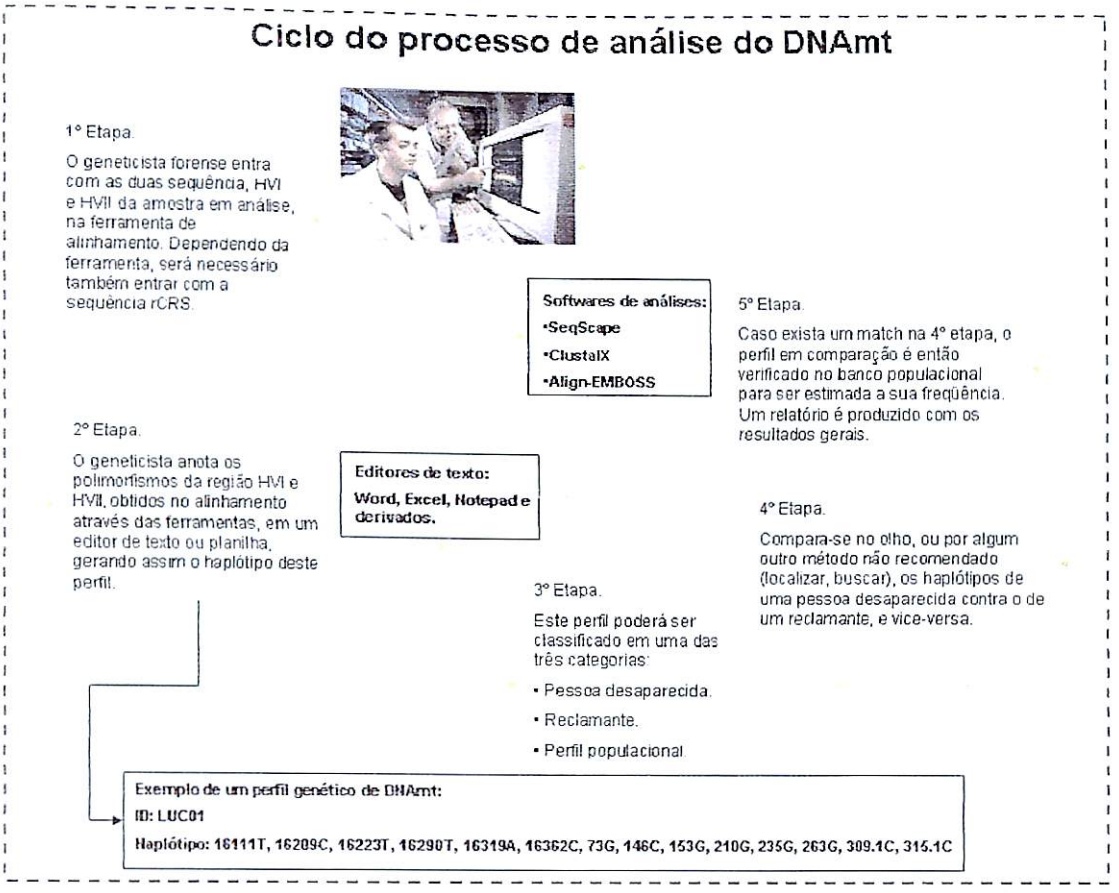


Figura 41: Ilustração das etapas executadas pelos geneticistas para a análise do DNAmt em estudos de casos forenses.

Ausentamos a etapa da verificação de erros dos polimorfismos gerados pelo alinhamento, na ilustração do ciclo do processo de análise do DNAmt, devido as problemáticas e a quantidade de erros que já foram apontados nos dados de DNAmt na literatura (seções 2.10 e 2.13). Pode-se dizer que esta prática não é ainda bem familiarizada, e se for, também não é estritamente adotada pelos laboratórios como uma etapa crítica e obrigatória.

Alguns geneticistas até desconfiam, por prepotência ou por falta de conhecimento, da utilidade da análise filogenética para apontar alguns dos possíveis erros que poderão vir a ocorrer. Talvez essa desconfiança seja apenas uma desculpa para evitar a perda de tempo com o minucioso e demorado trabalho que a análise filogenética exige, no caso, apenas para verificar se um simples polimorfismo suspeito poderá ter sido ocasionado por alguma falha em um dos processos de tipagem, análise e/ou na anotação dos dados.

O *ambiente de apóio e análise à identificação humana através do DNA Mitocondrial* abrange a concepção de um sistema a ser utilizado pelos diversos laboratórios de DNA forense no Brasil, provendo acesso *online* a um banco de dados centralizado. Esta abordagem tem como propósito, integrar os diversos laboratórios nacionais, padronizar os métodos de análises e automatizar o processo de geração dos perfis genéticos de DNAmt para serem armazenados e comparados através de uma base de dados comum.

Para o desenvolvimento do sistema proposto, elaboramos um ambiente que agrupe as diversas ferramentas necessárias que serão utilizadas em estudos de casos forenses para a geração e validação de perfis de DNAmt humano. São elas, ferramentas para alinhar seqüências de DNAmt de acordo com as três recomendações (capítulo 3), identificar e validar os haplótipos livres de erros e também, para funcionalidades que dizem respeito a gestão das informações a serem manipuladas, armazenadas e comparadas em um banco de dados central.

O ambiente constitui no desenvolvimento de um sistema *Web*, em conjunto com um banco de dados relacional, no esforço de integrar os laboratórios nacionais de DNA forense, em escopo nacional, possibilitando a utilização mútua, através de um ambiente homogêneo, acessando a uma base de dados comum, com ferramentas e dados padronizados, onde a troca e o compartilhamento de informações será crucial para o sucesso da identificação humana através dos dados genéticos do DNAmt (figura 42).

A escolha de desenvolvimento de um sistema *Web* se deu pelas vantagens que essa abordagem oferece como, por exemplo:

- Possibilita o acesso ao sistema através de um simples *browser*, de qualquer localidade, onde exista uma conexão de Internet.
- Todo o arcabouço do sistema é centralizado, evitando qualquer necessidade de instalação de componentes ou módulos nos terminais do cliente, referentes ao sistema.
- As atualizações no servidor refletem para todos os clientes, evitando a necessidade de atualizar cada um dos terminais de acesso individualmente.
- A necessidade do poder de processamento do aplicativo não reflete no lado do cliente.
- Possibilita a distribuição e o compartilhamento de dados, ao mesmo tempo que estes se encontram centralizados.

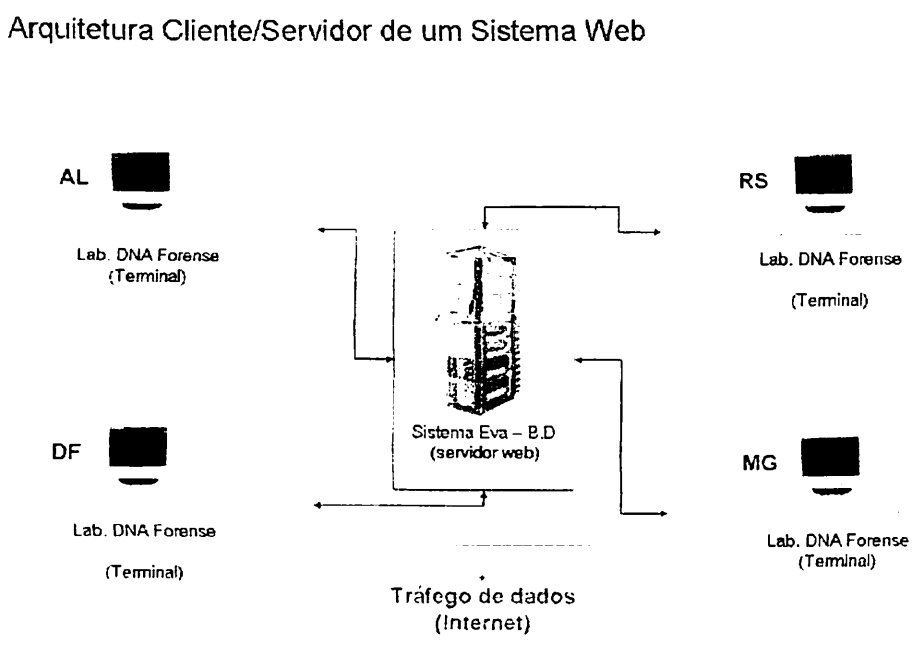


Figura 42: Esquema de funcionamento de um sistema *Web*. A linha verde representa o fluxo da Internet, a laranja a rede local dos terminais e a azul, a rede local do servidor *Web*. As setas pretas representam o canal de comunicação que interliga as redes.

A aplicação foi implementada utilizando a linguagem de programação PHP[2] versão-5.1.2. O PHP é uma linguagem de programação bastante poderosa e simples, com suporte para o desenvolvimento da programação orientada a objetos. Esta linguagem permite criar aplicações *Web* dinâmicas, possibilitando uma interação com o usuário através de formulários, parametros de entrada, entre outras características.

O código PHP é executado como linguagem de script no servidor, sendo enviado para o cliente apenas o código HTML. Desta maneira, é possível interagir com bancos de dados e aplicações existentes nos servidores.

Utilizamos o PHP em conjunto com o servidor *Web* Apache[3] versão-2.0.55 mais o banco de dados Access[4] (somente protótipo). A linguagem também se destaca pela possibilidade de interagir com uma vasta lista de banco de dados. Dentre eles estão um dos dois mais bem conceituados como o Oracle, PostgreSQL e o MySQL. Um recurso avançado do PHP é permitir conexões persistentes de banco de dados, minimizando a necessidade de constantes conexões (operações que aumentam o tempo de resposta das aplicações).

[2]URL - `http://www.php.net`
[3]URL - `http://www.apache.org`
[4]URL - `http://www.microsoft.com/access`

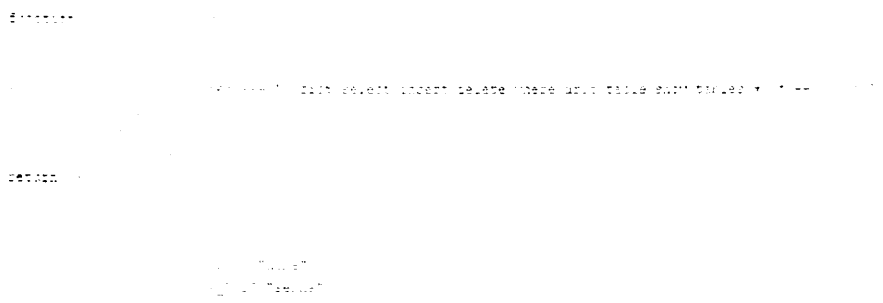


Figura 43: Função elaborada para evitar ataques de SQL *Injection*. Observe que o objetivo é anular qualquer sintaxe de SQL, tags de HTML ou da própria linguagem de programação, e qualquer caractere que venha a possibilitar a ocorrência de um erro da instrução de código utilizado pela aplicação que possa vir a revelar informações cruciais do seu banco de dados.

A quantidade de aplicações já existentes em PHP, a sua robustez, a rapidez na codificação e a facilidade de aprendizagem, fazem dela uma excelente escolha para o desenvolvimento de aplicações para a *Web*.

Em se tratando de um sistema *Web*, algumas medidas de segurança foram tomadas em relação ao controle de acesso do Eva. Por exemplo, um sistema de *Login* foi implementado onde só será permitido o acesso de laboratórios credenciados que solicitarem um pedido de permissão para uso, via email. Cada laboratório recebe um limite de usuários e todas as ações e os dados gerados serão de seu uso exclusivo e de total responsabilidade, onde nenhum outro laboratório poderá alterar ou modificar as informações geradas e armazenadas de segundos ou terceiros, sendo possível apenas a visualização, a análise e a comparação destes dados entre eles.

Medidas de segurança também foram tomadas para evitar SQL *Injections* (figura 43). SQL *Injection* é uma forma básica de ataque utilizada para obter acesso não autorizado a um banco de dados ou até mesmo para obter informações ocultas do banco que possam comprometer a sua integridade e segurança. O atacante tenta inserir uma string na aplicação, com um comando SQL manipulado a seu favor, para tentar obter informações sobre as instruções de SQL que estão sendo utilizadas pela aplicação atacada. Desta forma, o atacante poderá obter informações como, por exemplo, o nome e os atributos de uma tabela do banco de dados que contenha informações de *Login* dos usuários (KOST, 2004).

4.2 Modelagem

O sistema proposto neste trabalho, que a título de referência passa a ser denominado: Eva - *Identificação humana através do DNA Mitocondrial*, leva em conta a centralização das informações genéticas do DNAmt, coletadas em âmbito nacional, para serem analisadas e armazenadas em uma base de dados padronizada e livre de erros. Desta forma o Eva gerencia estas informações, no esforço de agilizar o processo da comparação destes dados, evitando redundâncias, visto que dois ou mais laboratórios poderão estar tipando e comparando as mesmas amostras de um mesmo caso, em diferentes estados. Isto aprimora a abrangência da identificação humana em todo o território nacional.

Por exemplo: um pai de família que mora no nordeste é levado sequestrado, morto e seu corpo incendiado, no sul do país. A identificação desse indivíduo poderá ser rapidamente agilizada pela inserção da análise do seu perfil DNAmt no Eva. Partindo da idéia de que através da doação de amostras do DNAmt de seus familiares (da mesma linhagem materna), estas serão também inseridas, no sistema Eva, diretamente de um dos laboratórios de DNA forense daquela região do país, e eventualmente possibilitando a identificação. A figura 44 ilustra este fluxo.

O sistema Eva visa otimizar o desempenho de todo o processo de identificação humana, desde a coleta da amostra biológica de um parente que mora em outro estado até ao envio da coleta das amostras biológicas de seus parentes que também residem em uma outra região distante, independente do laboratório que irá executar a extração e as análises desses perfis de DNAmt para serem comparados.

A figura 45 ilustra a arquitetura projetada e desenvolvida para o sistema Eva, composta pelos quatro módulos descritos abaixo:

- Alinhamento de sequências.
- Validador de polimorfismos.
- Buscador de similaridades.
- Calculador probabilístico.

Cada módulo é descrito nas seções seguintes.

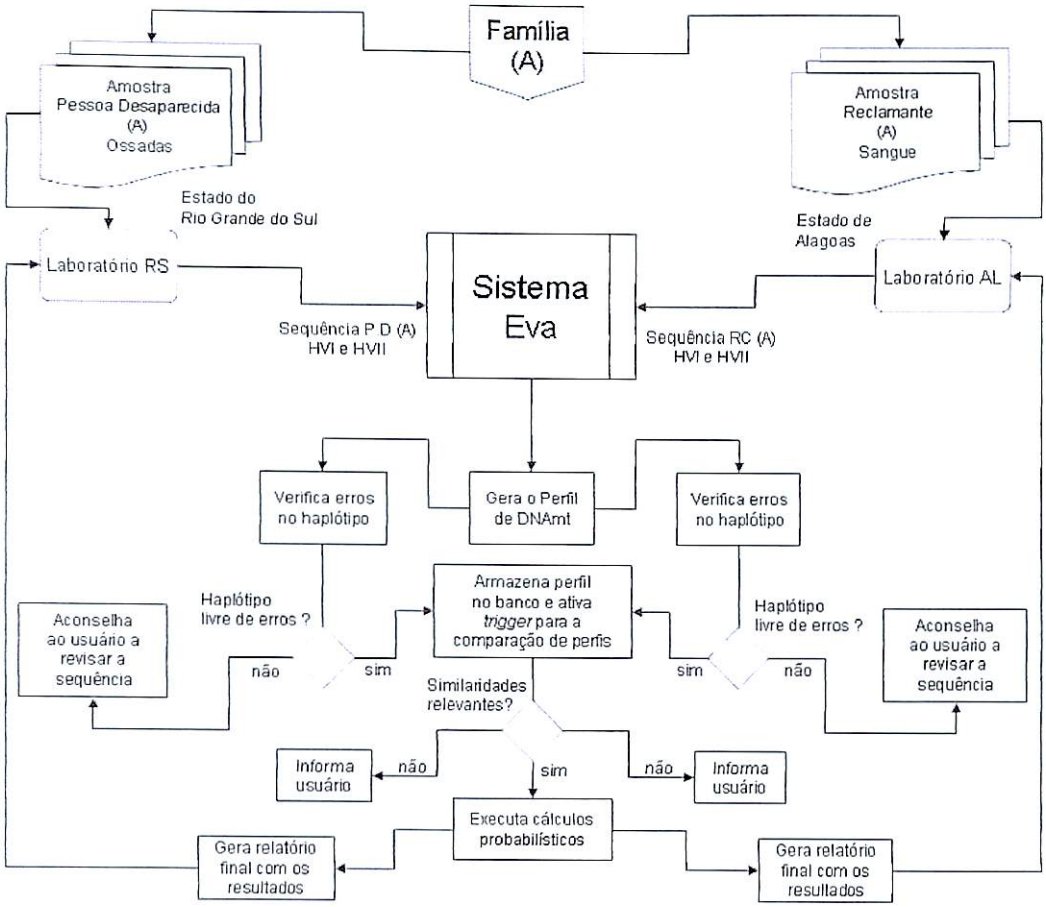


Figura 44: Diagrama do fluxo de dados no sistema Eva.

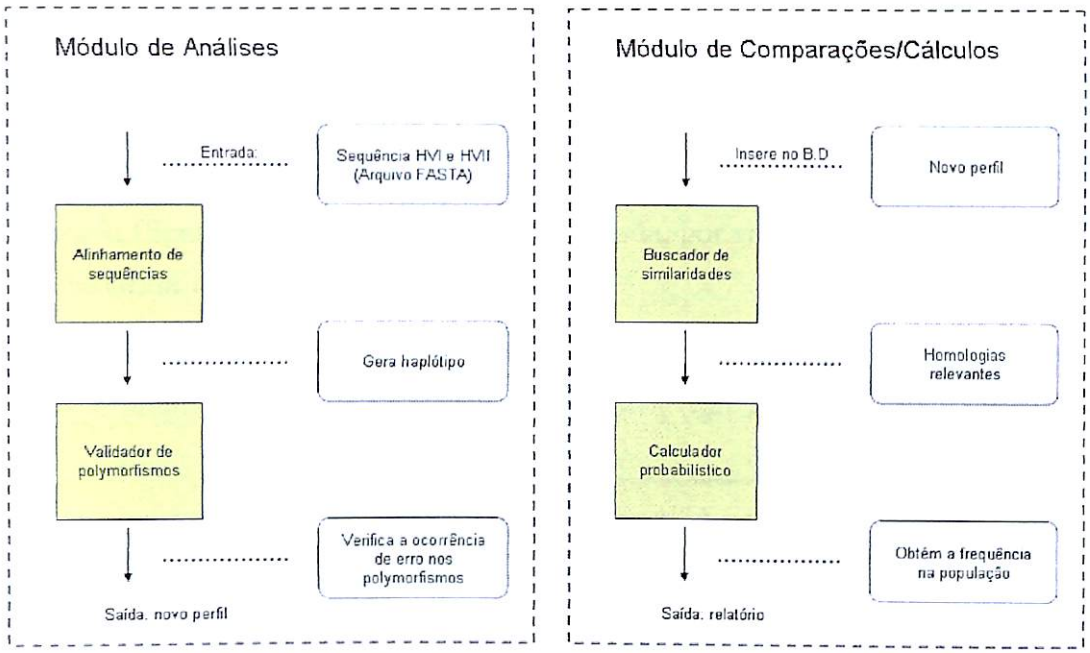


Figura 45: Arquitetura do sistema Eva - Identificação humana através do DNA Mitocondrial, composta pelos quatro módulos.

4.2.1 Módulo - Alinhamento de sequências

Diante dos fatos descritos nas seções 3.2.2, 3.2.3 e em suas subseções, fica evidente que a análise do alinhamento de sequências do DNAmt para fins de estudo de casos forenses ainda é muito precária. Essa precariedade se dá devido a ausência de uma ferramente específica para o alinhamento de suas sequências. Uma outra problemática existente é a carência de padronização para a anotação dos polimorfismos do haplótipo, gerados a partir do alinhamento.

Na seção 2.12 é apresentada uma hierarquia que deverá ser seguida para relatar os polimorfismos, e nas seções 3.2 e 3.2.1 essa problemática é ainda mais explicitada, com exemplos que apontam as recomendações surgeridas para que se mantenha uma padronização.

Portanto, este módulo foi desenvolvido com o objetivo de alinhar, de forma específica, as sequências de DNAmt para estudos de casos forenses. Além disto, após a execução do alinhamento, o módulo gera o haplótipo com os polimorfismos anotados, conforme os padrões, para posteriormente ser armazenado e comparado com os demais haplótipos no banco de dados.

O módulo tem como entrada as duas sequências HVI e HVII no formato Fasta. Este formato é constituído de um arquivo texto onde, por padrão, a primeira linha representa o cabeçalho e a partir da segunda linha coloca-se as bases da sequência. Os arquivos no formato Fasta vem sempre precedido com o símbolo ">" como primeiro caractere na primeira linha do arquivo. Após esse símbolo, ainda na primeira linha, a primeira palavra representa o nome ou o código dado a sequência; a segunda, o tipo de molécula (nucleotídeo ou proteína) utilizado; por fim anota-se a descrição da região e o tamanho da sequência (figura 46). Cada sequência é analisada, por vez, com a sequência rCRS que já vem embutida no Eva (figura 47).

O módulo utiliza o algoritmo de alinhamento global, descrito na seção seção 3.1.4, visto que é alinhado HVI da sequência de amostra com HVI da rCRS (mesma tática

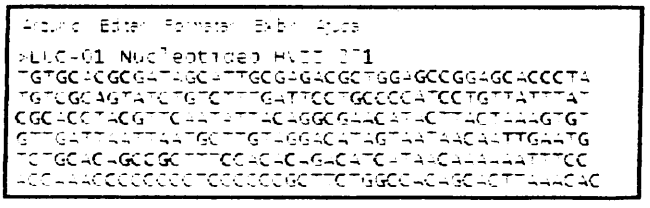


Figura 46: Exemplo de um arquivo Fasta formatado.

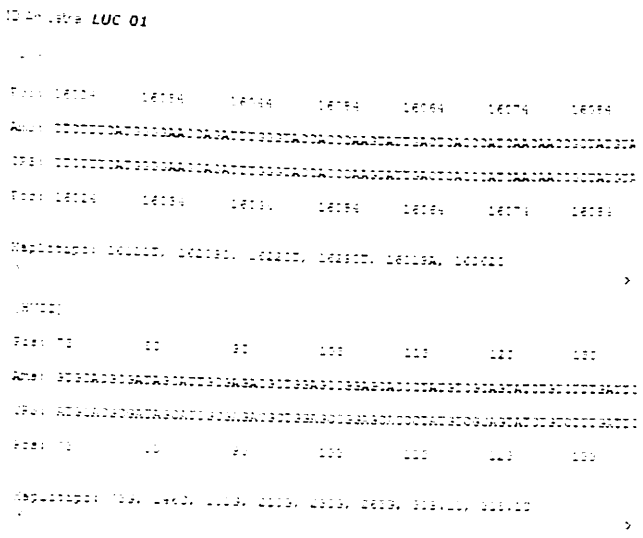


Figura 47: Geração de cada alinhamento das regiões HVI e HVII com os haplótipos anotados de acordo com a padronização, realizado pelo módulo - Alinhamento de sequências.

para HVII). O algoritmo gera a matriz com as similaridades entre as duas sequências, e assim, criando vários alinhamentos ótimos que poderão ser obtidos a partir da sua leitura. A problemática começa a surgir a partir do momento em que precisamos obter um alinhamento específico dentre os vários alinhamentos ótimos.

Tratar esta problemática gerando todos os alinhamento ótimos e a partir daí escolher, através de heurísticas, o alinhamento mais apropriado, poderá ser inviável devido à complexidade deste algoritmo na construção de sua matriz que é $O(mn)$ e é de $O(m+n)$ para obter cada um dos alinhamentos onde $(m+n)$ representa o tamanho do alinhamento a ser gerado. Como se trata de uma análise global do algoritmo, tem-se (na maioria dos casos) duas sequências, digamos n, com quase o mesmo tamanho. Portanto, a sua complexidade é considerada quadrática $O(n^2)$.

Em contrapartida, o módulo retorna apenas um alinhamento, lendo-se as setas da matriz a partir da entrada (m, n) até $(0, 0)$, de acordo com as três recomendações descritas na seção 2.12.

A heurística utilizada na leitura da matriz dá preferências a inserções (*gap* em s), deleções (*gap* em t) e como útlima alternativa, considera-se um alinhamento da base s com a base de t (figura 48). Concluído o processo de alinhamento, o módulo então anota as variações entre as sequências s e t de acordo com a padronização (seção 2.12):

Haplótipo - 16111T, 16209C, 16223T, 16290T, 16319A, 16362C, 73G, 146C, 153G, 235G, 236C, 263G, 309.1C, 309.2C, 315.1C

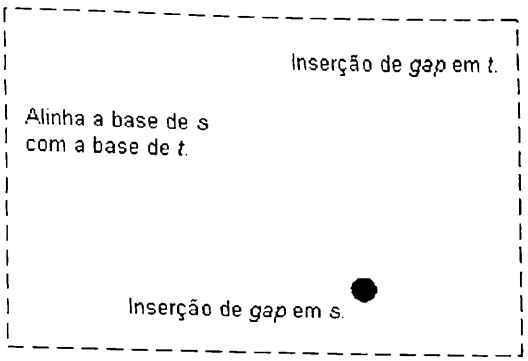


Figura 48: Heurística adotada pelo módulo - Alinhamento de sequencias, na leitura da matriz para obter o alinhamento ótimo. A preferencia das setas se dá de acordo com a sequencia das cores: verde, laranja e azul.

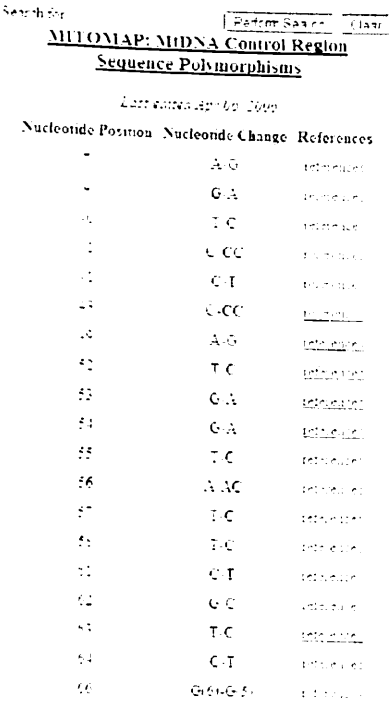
4.2.2 Módulo - Validador de polimorfismos

Todo o processo da análise do DNAmt é bastante minucioso, desde a tipagem até geração do seu perfil genético. Pois, em cada uma destas etapas, o DNAmt estará sujeito a erros que poderão ser provocados por uma eventual contaminação de laboratório devido a prática de manuseio da amostra ou, até mesmo, nas etapas de alinhamento de suas seqüências e na geração do seu haplótipo.

Como é descrito na seção 2.13, existem cinco classes de erros que classificam os tipos mais comuns que poderão ser gerados durante a análise do DNAmt. Nesta mesma seção, também é explanado que a técnica mais conhecida para a verificação destes erros é executar uma análise filogenética do haplótipo do perfil suspeito em questão. No entanto, a sua utilização ainda não é muito bem vinda, devido ao tempo que leva para que ela seja concluída e ao seu minucioso trabalho, visto que é necessária a intervenção e verificação pelo o usuário, dos grafos e árvores que são gerados durante o decorrer do seu processo.

Entretanto, o site do MITOMAP (BRANDON et al., 2004) contêm uma tabela com todos os polimorfismos validados pela comunidade genética, provendo também da sua referência literária que os valida (tabela 5).

Com estas informações, o geneticista forense é capaz de verificar se os polimorfismos, das regiões HVI e HVII, do haplótipo, de um perfil genético de DNAmt, estão coerentes com os polimorfismos já verificados e validados pela comunidade genética. Portanto, esta prática se torna uma alternativa para verificar erros que poderão ser provocados pela classe 3, já que o módulo de alinhamento de seqüências (seção anterior) se encarrega de evitar os possíveis erros que poderão ser gerados pelas classes 1, 2 e 4.



MITOMAP: MtDNA Control Region Sequence Polymorphisms

Nucleotide Position	Nucleotide Change	References
-	A-G	
-	G-A	
-	T-C	
-	C-CC	
-	C-T	
-	C-CC	
-	A-G	
-	T-C	
-	G-A	
-	G-A	
-	T-C	
-	A-AC	
-	T-C	
-	T-C	
-	C-T	
-	G-C	
-	T-C	
-	C-T	
-	G(6)-G-5	

Tabela 5: Parte da tabela que contem todos os polimorfismos validados pela literatura, da posição 7 à 16567, através do MITOMAP.

Porém, o geneticista forense, ao usufluir deste serviço oferecido pelo MITOMAP, terá sempre que procurar por cada um dos polimorfismos do haplótipo do perfil génetico em questão e, em seguida, acessar o site do MITOMAP procurando cada posição referente aos polimorfismos que ele deseja comparar para verificar sua autenticação. Desta forma, o geneticista forense estará correndo o risco de provocar erros das classes 2 e 1.

O presente módulo - *Validador de polimorfismos* tem como objetivo acessar o site do MITOMAP, capturar as informações dos polimorfismos e armazená-las no banco de dados do EVA. Com isto o módulo analisa cada posição e a base dos polimorfismos do haplótipo, que foram gerados pelo módulo de alinhamento de sequências, na sua base de dados de polimorfismos. Por fim o resultado da análise é exibido ao usário indicando se os polimorfismos estão de acordo (figura 19). A figura 50 ilustra a execução deste serviço.

No entanto, poderá acontecer de que algum polimorfismo, do haplótipo em questão, não esteja referenciado no site do MITOMAP. Neste caso, o sistema aconselha ao usuário de abortar a operação e reanalisar o eletroferograma, as sequências, executar uma análise filogenética e, até mesmo, resequênciar a amostra, pois o polimorfismo que não estiver validado poderá ter sido gerado através de algum erro durante a tipagem ou no decorrer do seu sequenciamento (classe de erros 3).

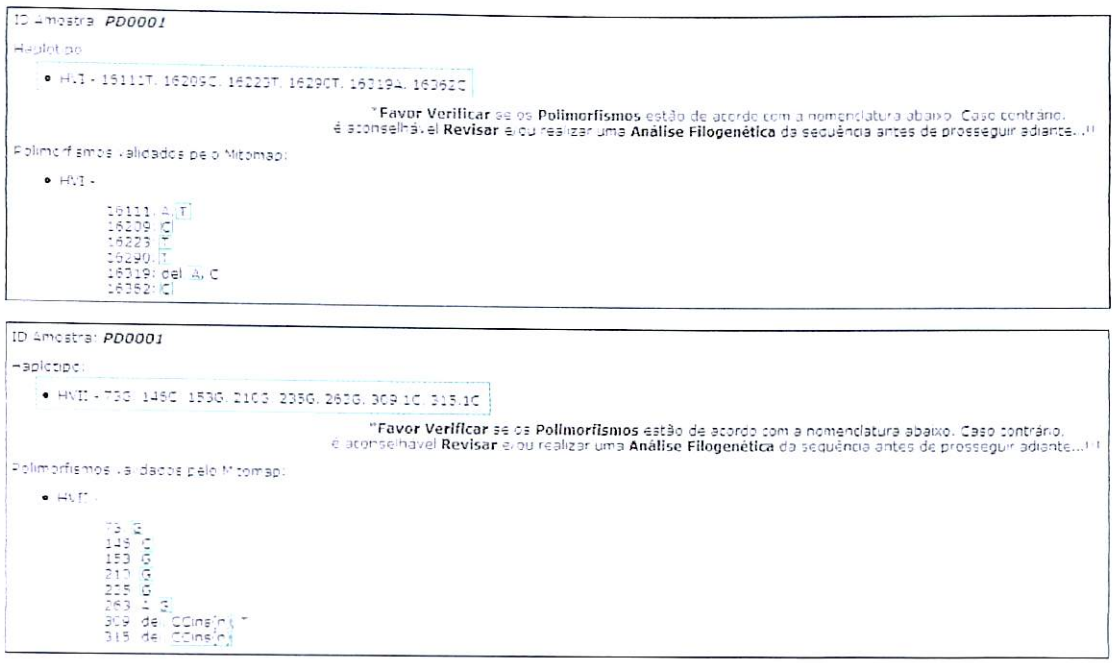


Figura 49: Resultado da análise dos polimorfismos do haplotipo da amostra *PD0001*, gerados pelo módulo - validador de polimorfismos, e exibidos pelo módulo - Alinhamento de sequências.

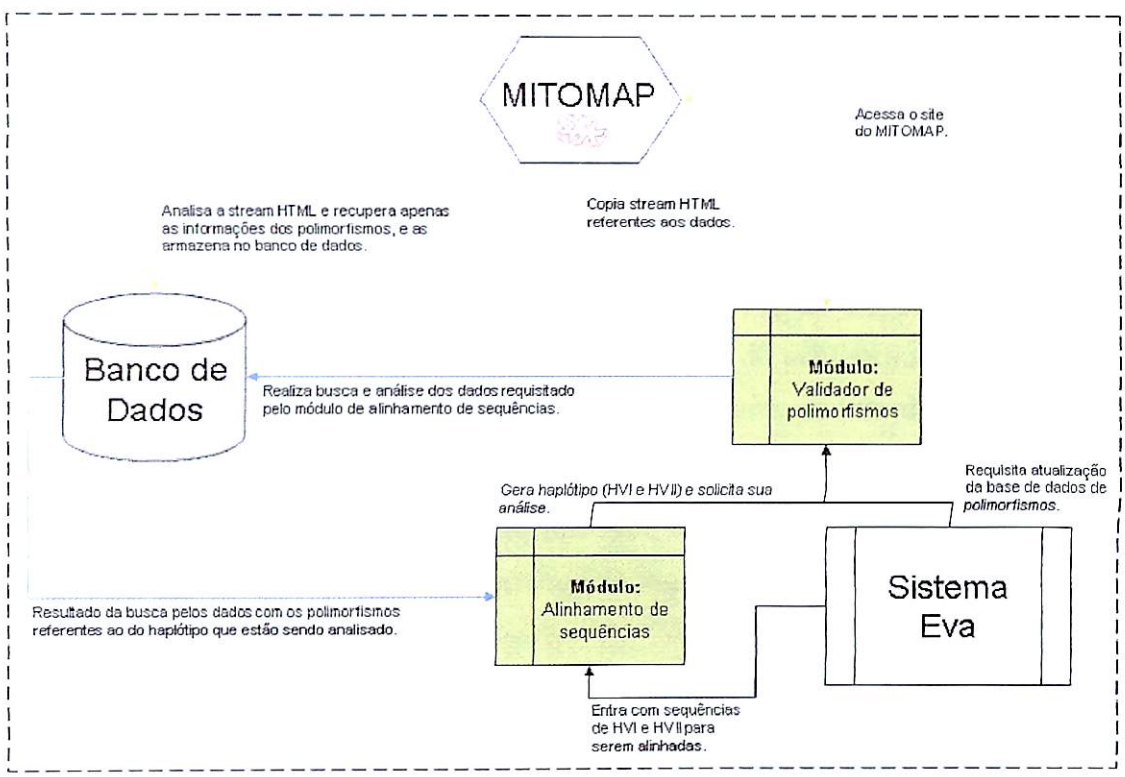


Figura 50: Diagrama do módulo - Validador de polimorfismos.

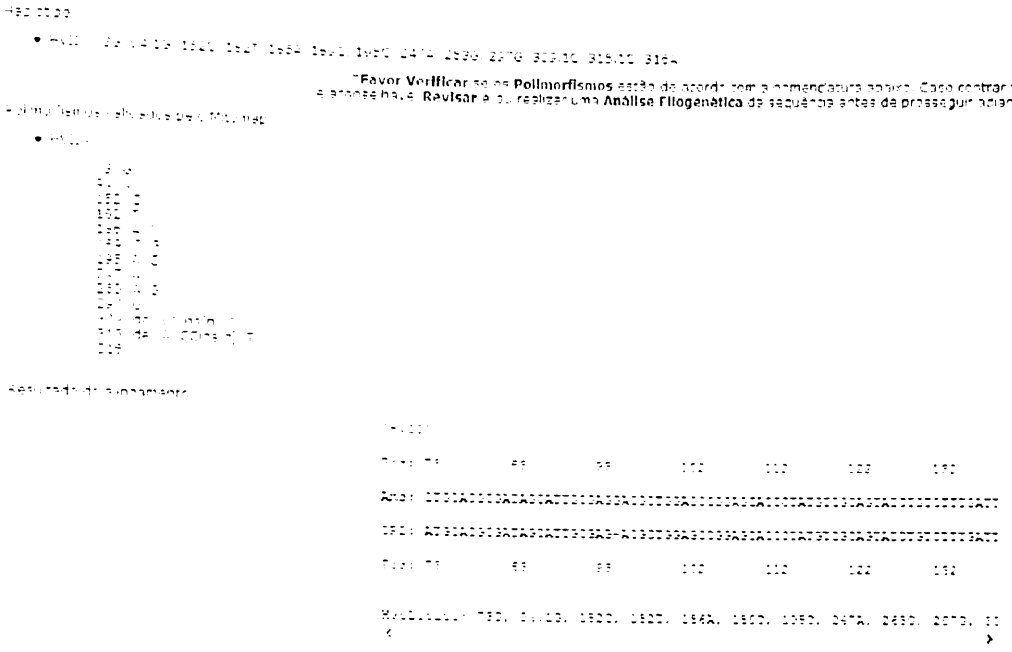


Figura 51: Verificação de um novo polimorfismos detectado pelo módulo - Validador de polimorfismos.

Entretanto, nem sempre, um polimorfismo que não esteja presente no MITOMAP signifique que ele tenha sido gerado por algum erro ou artifício e sim pelo fato de que mutações podem acontecer durante o passar dos anos no genoma mitocondrial, gerando novos polimorfismos ainda desconhecidos, mesmo que seja apenas a mutação de um único sítio.

Eventualmente, neste tipo de caso, pode-se verificar novamente a ocorrência deste novo polimorfismo em uma nova análise da amostra e/ou na análise de outras amostras, até que algum cientista geneticista o valide formalmente.

Por exemplo: ao submeter as sequências HVI e HVII da amostra AL35 para ser analisada no Eva, verificou-se, através da base de dados de polimorfimos (atualizada em 6 de Abril de 2006), a ocorrência de um polimorfismo, no segmento de HVII, ainda não validado pela literatura, o '94.1G'.

A princípio este novo polimorfismo não aparenta ser provocado por algum tipo de erro, visto que a posição '94' está validada como sendo um sítio variante, apresentando apenas a variação de uma única adenina (A) (figura 51). Mesmo assim, um ressequenciamento da amostra foi realizado para certificar a sua ocorrência.

Portanto, em tal situação, o polimorfismo poderá ser considerado válido e o perfil seguro de que está livre de erros da classe 3 para ser armazenado no banco de dados e comparado com os demais perfis, como é descrito na próxima seção.

4.2.3 Módulo - Buscador de similaridades

A criação de um banco de dados se fez necessária diante da concepção do sistema Eva e da importância de se criar dados com qualidade para serem armazenados e gerenciados de maneira apropriada.

A concepção de um sistema para dar suporte a identificação de pessoas desaparecidas, através da análise do DNAmt, necessita da criação de um banco de dados populacional, consistituído de perfis escolhidos ao acaso, utilizados como ferramenta para estimar o peso de uma evidência (*match*), como também a de uma banco de dados forense para dar suporte à gerência dos perfis de pessoas desaparecidas e de seus reclamantes.

Desta forma essa base de dados, em conjunto com o servidor *Web*, permitirá o compartilhamento de perfis de DNAmt, de pessoas desaparecidas (PD) e os de reclamantes (RC), entre os laboratórios, referentes da mesma linhagem materna.

A problemática desta concepção é percebida na medida em que vai aumentando e se acumulando, no banco, o número de perfis de DNAmt de reclamantes e o de pessoas desaparecidas, para serem comparados e analisados. A solução desta problemática se torna inviável com a comparação e a análise destes perfis pelas práticas exercidades atualmente, qual seja utilizando editores de textos ou planilhas e, ou até mesmo a olho nú.

No esforço de resolver esta problemática, o presente módulo, ao armazenar o perfil gerado pelo módulo de alinhamento de sequências, ativa a *trigger* de comparação deste recente perfil com os demais perfis no banco.

Por exemplo: ao se analisar, validar e armazenar um novo perfil de DNAmt de uma pessoa desaparecida, o presente módulo, automaticamente, aciona o seu dispositivo de comparação deste novo perfil com os demais perfis de reclamantes já armazenados na base de dados e vice-versa.

O número de comparações C de um perfil genético em questão se dá pela seguinte equação, onde N é o tamanho da base de dados de PD ou RC vezes x que á a quantidade de perfis genéticos em que se pretende comparar:

$$C = N \cdot x \tag{4.1}$$

O presente módulo compara e analisa os perfis de acordo com as recomendações de interpretação descritas na seção 2.8. São realizadas quatro tipos de análises comparativas para um dado perfil em questão:

- Os perfis em comparação variam em dois ou mais polimorfismos. É notificada a ocorrência de uma exclusão.
- Os perfis em comparação apresentam o mesmo haplótipo. É notificada a ocorrência de um *match* perfeito, não-exclusão.
- Os perfis em comparação diferem apenas pela presença de heteroplasmia. É notificada a ocorrência de um *match*, não-exclusão.
- Os dois perfis em comparação diferem apenas por um único polimorfismo. É notificada a ocorrência de um *match*, inconclusivo.

O presente módulo também realiza análises de busca de similaridades do perfil em questão com os perfis do banco populacional, após a certificação da ocorrência de um *match* na sua comparação com a base de dados de PD ou RC.

Nesta análise, com o banco populacional, o presente módulo interage com o módulo calculador probabilístico para poder estimar o peso da evidência do perfil genético em questão. Esta interação é descrita na seção seguinte.

A figura 52 ilustra o diagrama do serviço executado pelo módulo buscador de similaridades e a figura 53 ilustra o resultado das análises comparativas gerada pelo presente módulo.

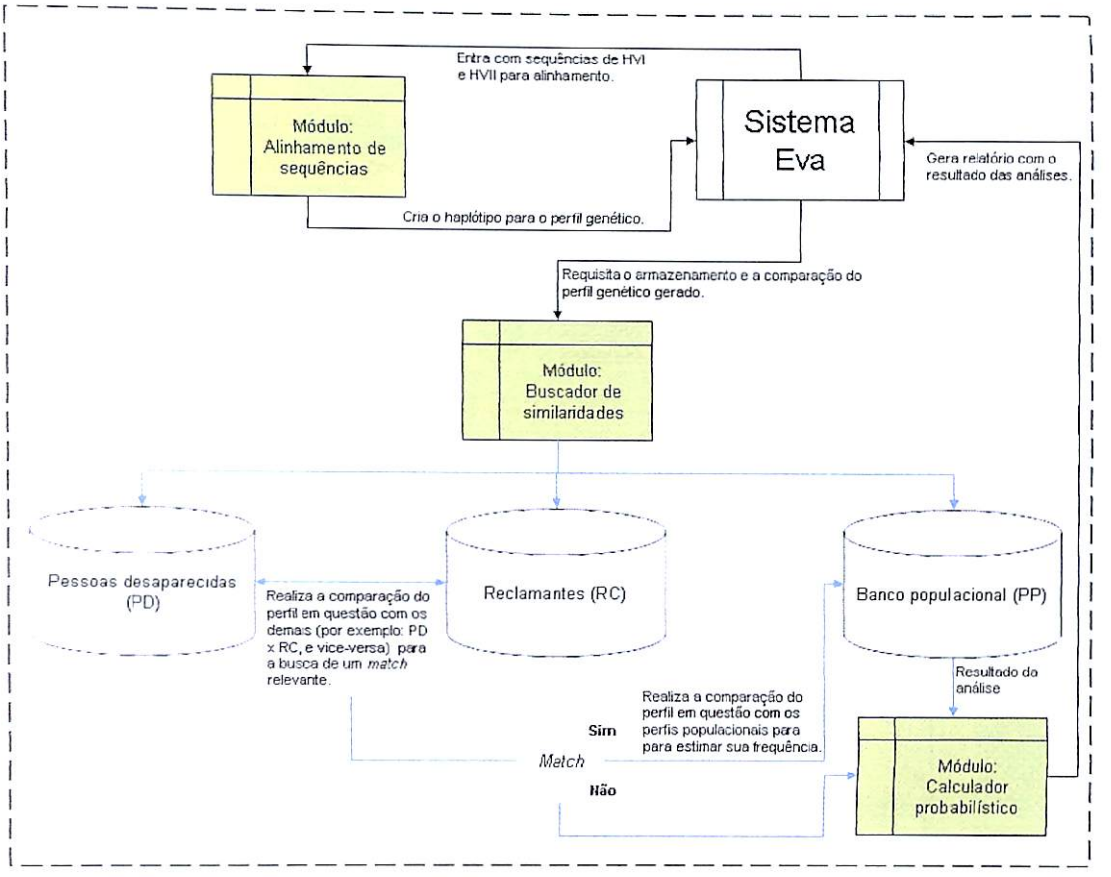


Figura 52: Diagrama do módulo - Buscador de similaridades.

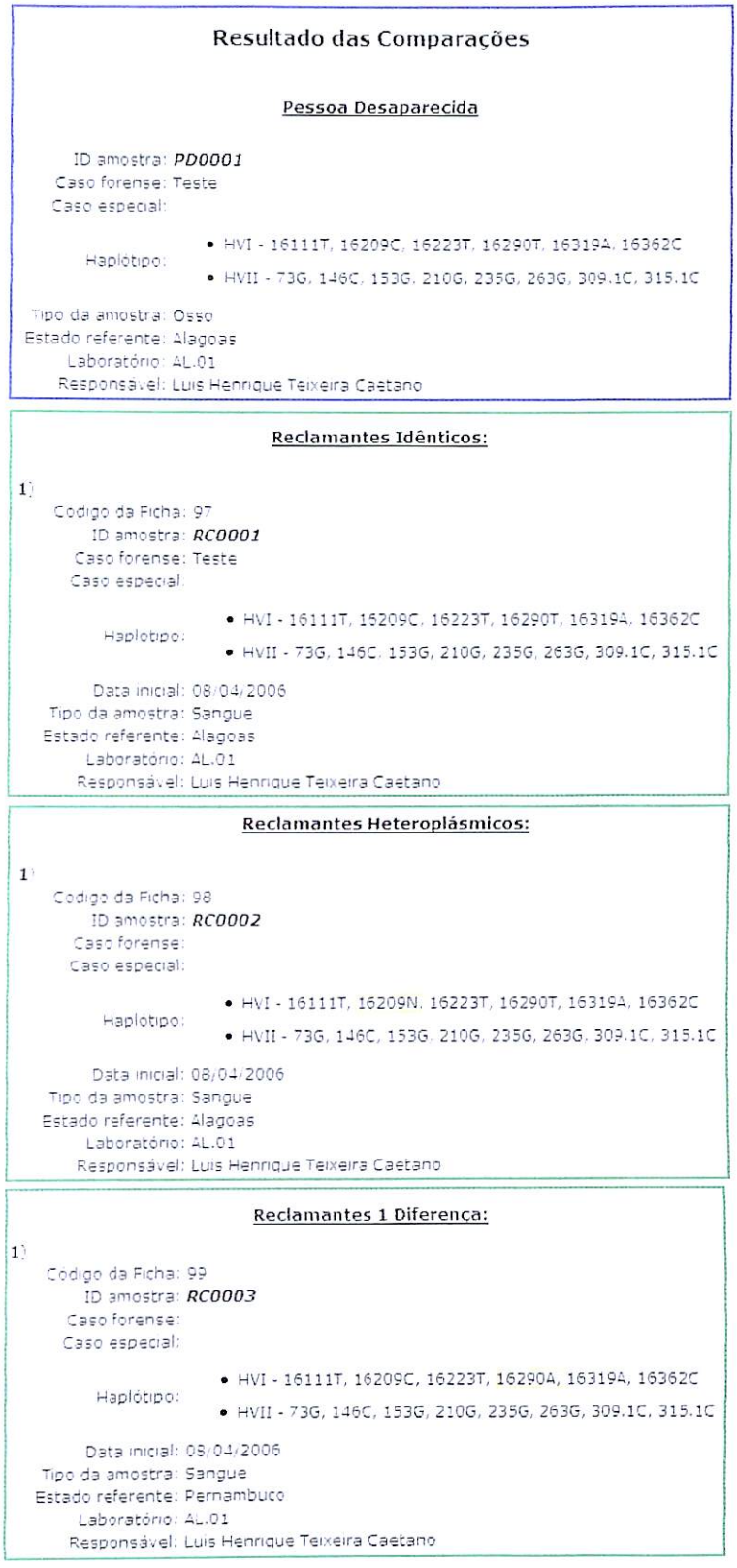


Figura 53: Resultado da análise das comparações entre um perfil de um PD com os perfis de RC, gerado pelo módulo - Buscador de similaridades. As entre linhas verdes representam os perfis RC com os quais se obteve um *match* em relação ao perfil PD. As entre linhas amarelas representam as variantes em relação à PD.

4.2.4 Módulo - Calculador probabilístico

Após ter completado o processo de análise referente à sequência de uma amostra, que engloba a geração do seu perfil genético, a verificação de erros no seu haplótipo e a verificação de um *match* relevante, a etapa final a ser realizada de todo este processo é estimar o peso da evidência desta amostra na população.

Para a execução desta etapa o geneticista forense necessita de uma base de dados composta por perfis da população para poder estimar a frequência da amostra (perfil) em estudo, qual seja ela de um PD ou RC.

Por exemplo, suponha que o Eva possua, na sua base de dados de RC, 50 perfis armazenados a espera de um *match* relevante na sua comparação com um perfil de PD que venha a ser inserido no sistema. Suponhamos ainda que, ao entrar com um novo perfil de um PD, o Eva revele a corrência de um *match* entre dois perfis de RC. Este *match* significa que esta nova amostra de PD poderá ser o parente desaparecido de um destes dois reclamantes, o qual eles estejam à procura. O presente módulo calcula a probabilidade desta amostra PD ser ou não ser este possível parente desaparecido. Para isto, inicialmente, o módulo buscador de similaridades analisa o perfil em questão para verificar a sua presença no banco de dados populacional. Em seguida, o resultado desta análise é então passado para este módulo estimar o peso da evidência deste perfil (PD) gerando, assim, o resultado probabilístico de que esta amostra de uma pessoa desaparecida seja, possivelmente, o parente de um desses dois reclamantes.

O módulo realiza este cálculo de acordo com o número de observações do perfil em questão no banco de dados populacional, de acordo como é descrito no na seção 2.9. A figura 54 ilustra este serviço executado pelo presente módulo em conjunto com o módulo buscador de similaridades.

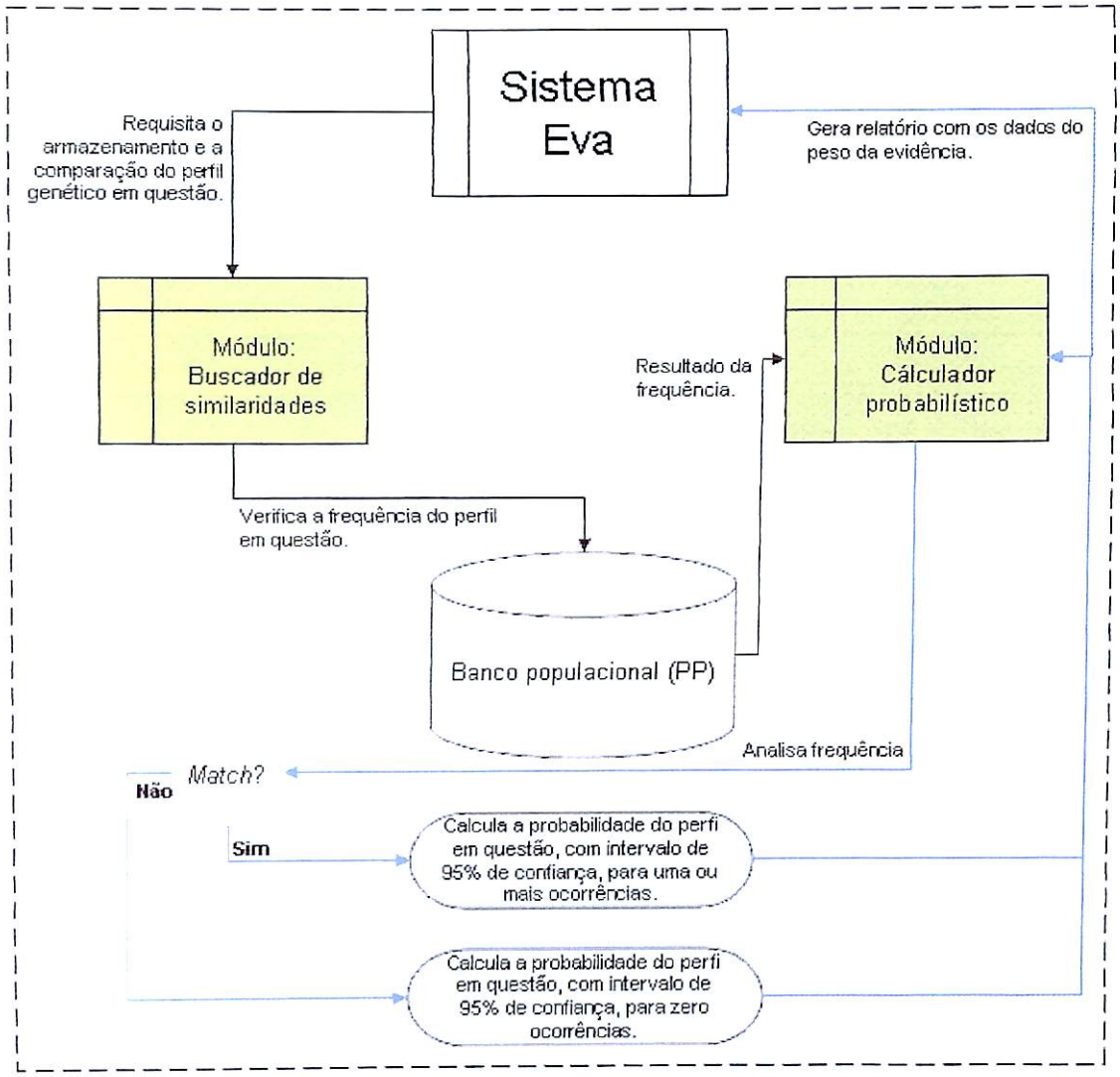


Figura 54: Diagrama do módulo - Calculador probabilístico.

5 *Experimentos e Simulações*

"Só não falha quem não tenta
e quem não tenta jamais terá sucesso."
Reflexão

Neste capítulo são apresentados os dados utilizados nos testes realizados com o sistema Eva para confirmar a acurácia de suas ferramentas ...

5.1 Análise das 123 amostras

No trabalho de Barbosa (2006), também realizado em parceira com laboratório de DNA forense da UFAL, foram escolhidas ao acaso 123 amostras de indivíduos das diversas cidades do estado de Alagoas (figura 55). Maceió e Arapiraca foram as duas cidades com o maior número de amostras coletadas, 33 em Maceió e 11 amostras de Arapiraca. Estas 123 amostras foram sequenciadas e analisadas para compor o banco populacional do Eva. Nestas 123 amostras, encontram-se 110 haplótipos diferentes (perfis) e o haplótipo mais frequente é encontrado em apenas 5 dos 123 indivíduos. Desta forma, a diversidade genética do banco de dados populacional do Eva (composta por estas 123 amostras) foi estimada em 99.7% (SCHNEIDER et al.,).

Estas 123 amostras foram submetidas para serem alinhadas no Eva, como teste, para verificar se o módulo de alinhamento de sequências estaria alinhando perfeitamente de acordo com os padrões propostos pela literatura.

As sequências das duas regiões de HVI e HVII das 123 amostras foram alinhadas, com sucesso, através do Eva. O alinhamento apresentou ser 100% eficaz em todas as 246 sequências, compostas por HVI e HVII, gerando e anotando o perfil de cada uma delas (figura 56) de maneira apropriada. Tendo alcançado este objetivo, as 123 amostras puderam ser alinhadas e os seus perfis genéticos foram gerados, automaticamente, para então serem armazenados com qualidade e segurança, possibilitando assim a criação do banco de dados populacional. A ferramenta de alinhamento, ao gerar e anotar o perfil genético com sucesso, contribui para a eliminação dos erros que podem ser provocados pelas classes 1, 2 e 4 (seção 2.13).

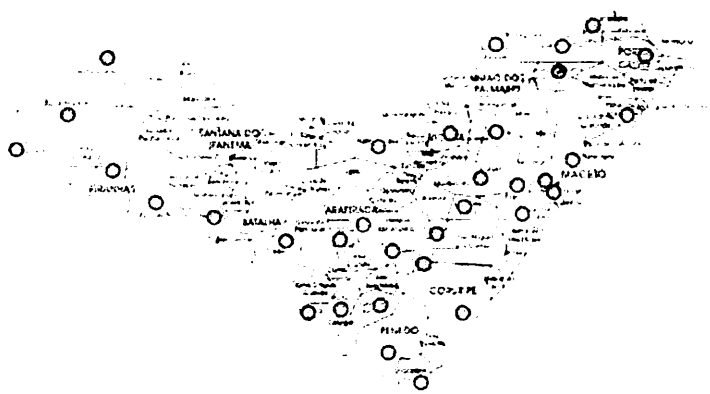


Figura 55: Ilustração do mapa que representa o Estado de Alagoas, indicando os municípios de origem dos 123 indivíduos analisados. Fonte: (BARBOSA, 2006).

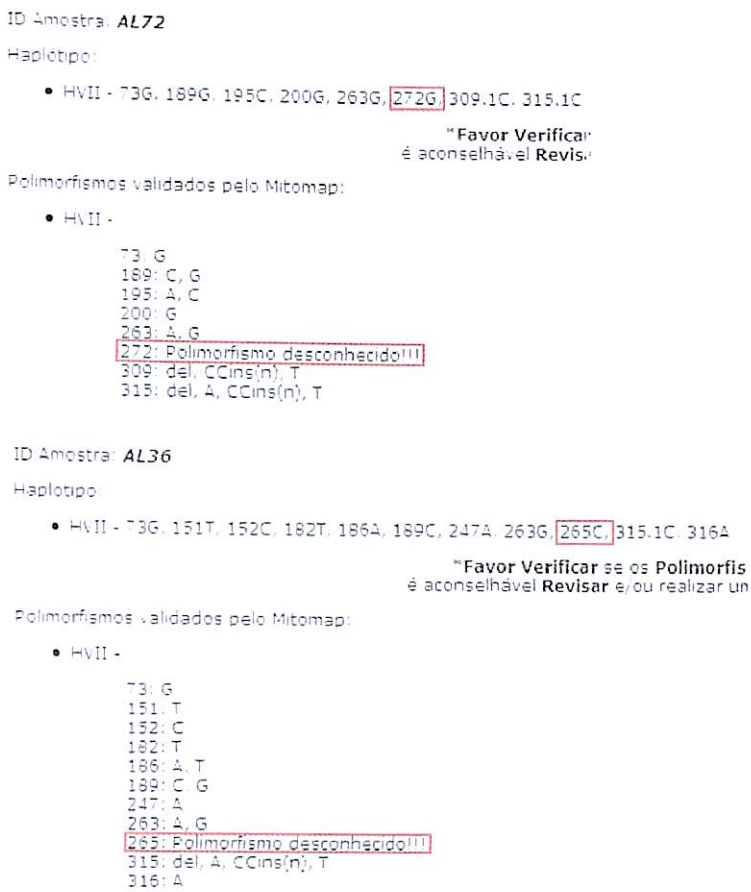


Figura 56: Mutações suspeitas de erro, detectadas pelo Eva.

Neste mesmo estudo (BARBOSA, 2006) verificou-se que dentre as sequências das 123 amostras 96 foram únicas para HVI e 79 para HVII. A região de HVI apresentou 87 substituições onde 77 são transições e as outras 10 são transversões, e mais 2 deleções. Já nas sequências de HVII observou-se 41 substituições com 37 transições e 4 transversões, mais 3 deleções e 4 tipos de inserções. A maioria das mutações (polimorfismos) em ambos os segmentos foram de transições, o que confirma a proposta de Wilson(WILSON et al., 2002) no tratamento de variantes (seção 2.12) que reflete a complexidade da elaboração do alinhamento de sequências de DNA Mitocondrial para o estudo de casos forenses (seção 3.2).

O módulo validador de polimorfismos mostrou ser uma valiosa ferramenta para detectar possíveis erros da classe 3. Os perfis tidos como "suspeitos" foram validados com sucesso pelo módulo. A figura 56 apresenta um resultado gerado pelo Eva onde as substituições na região de HVII, apresentando o polimorfismo 265C no perfil AL36 e a substituição 272G no perfil AL72, foram apontadas como sendo desconhecidas. Porém, o ressequenciamento destas amostras confirmou a aparição das duas mutações inéditas.

5.2 Estudo de caso forense

Serão descritos, nas duas próximas seções, dois casos forenses como exemplo de uso do sistema Eva para a identificação humana, através de restos biológicos degradados, pelo estudo do DNAmt.

O primeiro caso trata de um incidente verídico onde estaremos ligando uma amostra de cabelo degradado (encontrada na cena de um crime) a um suspeito criminoso, foragido, através da sua análise com uma amostra do cabelo da mãe do suspeito. Porém, o nome de todos os envolvidos neste caso serão mantidos em sigilo, com nomes fictícios.

O segundo caso descreve a simulação da queda de um avião onde a grande maioria da identificação dos restos mortais, dos passageiros, só poderão ser identificados através do estudo do DNAmt.

5.2.1 O caso Mercerdes

A secretaria de segurança pública do Estado do Mato Grosso enviou amostras de cabelos, encontradas em cenas de crimes, para o laboratório de DNA forense da UFAL, suspeitas de pertencerem a um único *serial killer*.

A polícia começou a desconfiar destas amostras de cabelos encontradas em algumas cenas de crimes, bastante semelhantes. A polícia verificava que a vítima era morta por estrangulamento e sempre havia, ao lado da sua mão direita, um monte de cabelo cortado.

A partir daí a polícia daquele estado começou a associar que estes crimes estavam sendo cometidos por um mesmo assassino. Após cometer o homicídio de suas vítimas, o *serial killer* cortava seu próprio cabelo e o deixava junto à vítima como sendo uma espécie de marca registrada.

Alguns suspeitos começaram a ser levantados. Um deles estava foragido e já era conhecido pela própria polícia. Providências foram tomadas para que a mãe do suspeito foragido fornecesse amostras do seu cabelo para serem comparadas com as amostras de cabelos encontradas na cena dos crimes.

Estas amostras de cabelos (do principal suspeito e de sua mãe) foram então tipadas pelo laboratório de DNA Forense da UFAL. Foram enviadas duas amostras de cabelo de duas cenas de crime e mais uma amostra de cabelo da mãe do suspeito, denominada Mercedes.

Primeiro foram sequenciadas as regiões hipervariáveis, HVI e HVII do DNAmt, das duas amostras de cabelo, das cenas do crime, denominadas de Cab01 e Cab02. Em seguida o mesmo processo foi realizado para sequênciar as duas sequências de HVI e HVII da amostra de cabelo de sua suposta mãe (Mercedes), denominada de CabM. O objetivo foi gerar o perfil de Cab01 e de Cab02 para serem comparados com o perfil de CabM.

Em seguida as duas sequências de HVI e HVII, de Cab01 e Cab02, foram submetidas ao sistema Eva para serem alinhadas e gerados o seu perfil genético de DNAmt para serem armazenados na base de dados de pessoas desaparecidas, com o rótulo do seu caso forense de: CF M 01.

Suas sequências foram alinhadas e seus perfis gerados com sucesso. O sistema Eva detectou a ocorrência de uma substituição suspeita no sítio 263G que, com uma atualização do banco de polimorfismos, essa suspeita foi eliminada, evitando assim um ressequenciamento das suas amostras (figura 57 e 58).

Por último a amostra CabM foi analisada, gerado o seu perfil e armazenada na base de dados de reclamantes para ser analisada e comparada com as amostras Cab01 e Cab02 (figura 59).

O resultado final das análises comparativas apontaram um *match* de uma não-exclusão (perfis idênticos) entre as três amostras. De acordo com o banco populacional, que na época comportava apenas 10 perfis populacionais, estimou-se que a probabilidade deste perfil ser encontrado na população é de 25,9% e, com 95% de confiança, é possível excluir 74.1% da população como possíveis fontes desta amostra, visto que a sua frequência foi zero no banco populacional. Entretanto, o simples fato do sistema ter apontado:

1. A igualdade entre os dois cabelos, encontrados em cenas de crime semelhantes, dado que a polícia já tinha levantado um suspeito, de seu conhecimento, que estava foragido.

2. Um *match* entre as duas amostras de cabelo, das cenas de crime, com a amostra de cabelo da fonte de referência, ou seja, da mãe deste suspeito.

- Podemos concluir que: O resultado de que as três amostras possuem o mesmo haplótipo já poderá ser utilizado como forte indício de que o *serial killer* suspeito seja, realmente, o responsável por todos os crimes que contenham este mesmo perfil genético nas amostras de cabelo.

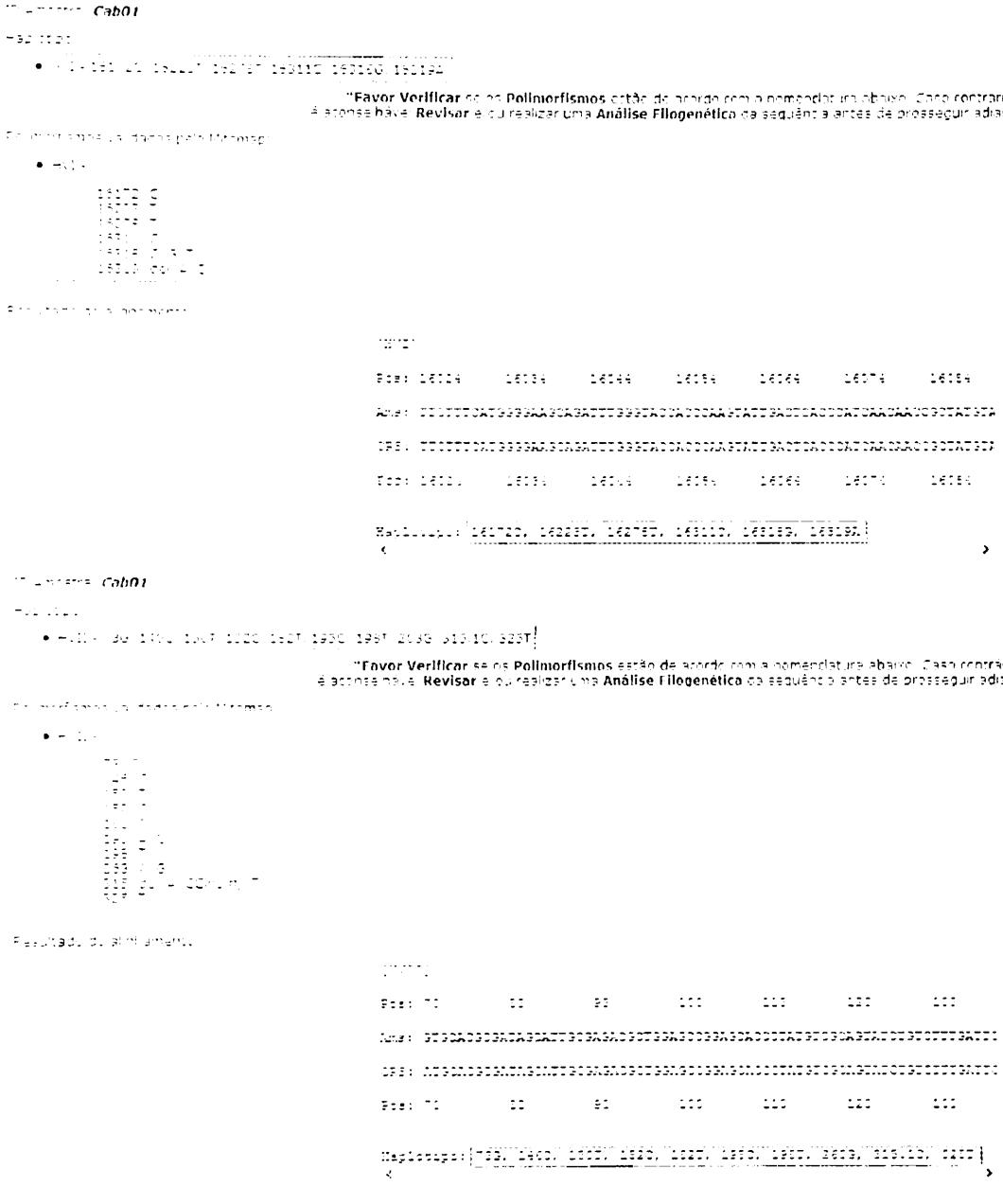


Figura 57: Perfil da amostra Cab01, gerado através da análise das sequencias de HVI e HVII, pelo Eva. As entre linhas laranjas destacam o polimorfismo que, a princípio, era duvidoso, mas acabou sendo confirmado com a atualização da base de dados de polimorfimos.

ID Amostra: **Cab02**

Haplotipo:

- HVI - 16172C, 16223T, 16278T, 16311C, 16318G, 16319A

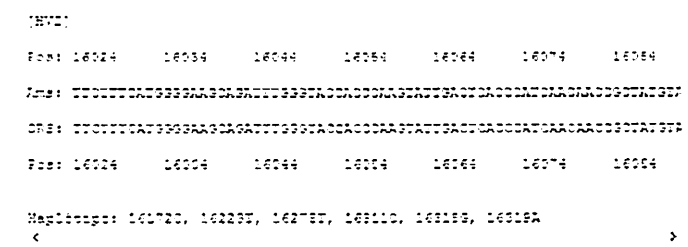
"Favor Verificar se os **Polimorfismos** estão de acordo com a nomenclatura abaixo. Caso contrário, é aconselhável **Revisar** e/ou realizar uma **Análise Filogenética** da sequência antes de prosseguir adiante..."

Polimorfismos validados pelo Mitomap:

- HVI -

16172 C
16223 T
16278 T
16311 C
16318 C, G, T
16319 del A, G

Resultado do alinhamento:



ID Amostra: **Cab02**

Haplotipo:

- HVII - 73G, 146C, 150T, 152C, 182T, 195C, 198T, 263G, 315.1C, 325T

"Favor Verificar se os **Polimorfismos** estão de acordo com a nomenclatura abaixo. Caso contrário, é aconselhável **Revisar** e/ou realizar uma **Análise Filogenética** da sequência antes de prosseguir adiante..."

Polimorfismos validados pelo Mitomap:

- HVII -

73 G
146 C
150 T
152 C
182 T
195 C
198 T
263 A, G
315 del A, C(n=[n]), T
325 T

Resultado do alinhamento:

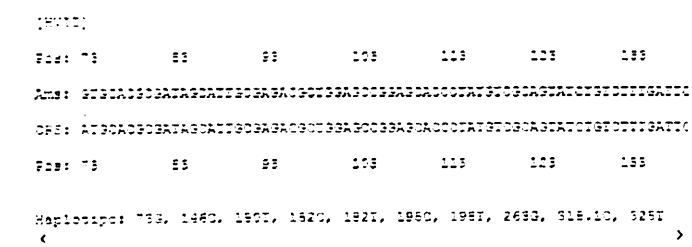


Figura 58: Perfil da amostra Cab02 gerado através da análise das sequências de HVI e HVII, pelo Eva.

Reclamante

Código pericia: 120
ID amostra: ***CabN***
Caso forense: CF/M-01
Caso especial: CF/M-01

Haplótipo:
- HVI - 16172C, 16223T, 16278T, 16311C, 16318G, 16319A
- HVII - 73G, 146C, 150T, 152C, 182T, 195C, 198T, 263G, 315.1C, 325T

Tipo da amostra: Cabelo
Estado referente: Mato Grosso
Laboratório: AL-01
Responsável: Luis Henrique Teixeira Caetano
Resultado probabilístico: 0.259 . Com 95% de confiança, podemos excluir 74.1% da população como possíveis fontes do perfil genético.
Frequência no banco populacional: 0
Tamanho do banco populacional: 10 amostras.

Pessoas Desaparecidas Idênticas:

1:

ID amostra: ***Cab01***
Caso forense: CF/M-01
Caso especial: CF/M-01

Haplótipo:
- HVI - 16172C, 16223T, 16278T, 16311C, 16318G, 16319A
- HVII - 73G, 146C, 150T, 152C, 182T, 195C, 198T, 263G, 315.1C, 325T

Data inicial: 13/04/2006
Tipo da amostra: Cabelo
Estado referente: Mato Grosso
Laboratório: AL-01
Responsável: Luis Henrique Teixeira Caetano

2:

ID amostra: ***Cab02***
Caso forense: CF/M-01
Caso especial: CF/M-01

Haplótipo:
- HVI - 16172C, 16223T, 16278T, 16311C, 16318G, 16319A
- HVII - 73G, 146C, 150T, 152C, 182T, 195C, 198T, 263G, 315.1C, 325T

Data inicial: 13/04/2006
Tipo da amostra: Cabelo
Estado referente: Mato Grosso
Laboratório: AL-01
Responsável: Luis Henrique Teixeira Caetano

Figura 59: Resultado final do caso CF/M/01 realizado através do Eva.

5.2.2 Acidente aéreo

Os laboratórios de DNA forense têm grande dificuldade em analisar casos onde o número de amostras seja muito grande. Essa problemática se dá especialmente pela falta de softwares que dêem suporte para a automatização das análises, do armazenamento e da comparação dos perfis destas amostras (BUDOWLE; BIEBER; EISENBERG, 2005).

Podemos citar vários casos famosos que já ocorreram pelo mundo, como, por exemplo, o desastre do Tsunami na Ásia e o ataque de 11 de Setembro das duas torres gêmeas em Nova York, EUA. Podemos também citar um outro caso envolvendo um grande número de mortos, com corpos parcialmente carbonizados, apresentando difícil identificação, o qual foi o incêndio do supermercado no Paraguai, onde as autoriadades daquele país pediram ajuda aos países vizinhos, especialmente ao Brasil, para ajudar na identificação de suas vítimas através da análise do DNA.

Grandes desastres envolvem um esforço muito grande de vários laboratórios para tipar e gerenciar as amostras de milhares de vítimas para serem identificadas. O sistema Eva pode vir a contribuir para dar este tipo de suporte, tanto na automatização das análises dessas seqüências como na gerência dos seus dados.

Por exemplo, no caso de um acidente aéreo onde mais da metade dos restos mortais dos passageiros não podem ser identificados por outro meio a não ser pelo estudo do seu DNAmt, devido à degradação das amostras e considerando que:

- O avião transportava 200 passageiros a bordo.
- 50 dos 200 passageiros puderam ser identificados através de vestígios, sinais no corpo, dentre outros aspectos físicos, como também através da tipagem do seu DNAn.
- Restando, portanto, 150 passageiros para serem identificados através do estudo do DNAmt.

Considerando ainda que nenhum destes 150 passageiros possuam algum laço de parentesco de 1º grau entre si, teremos, então, mais 150 famílias à procura da identificação do resto mortal do seu familiar que estava abordo.

Portanto, o problema consiste em ter 150 reclamantes r_i que, teoricamente, devem ser comparados com cada uma das N 150 amostras não identificadas (dos passageiros), totalizando em 150 iterações i.

Porém, a cada comparação, teoricamente, ocorrerá uma identificação e, portanto, o valor de N será subtraído pelo somatório de i a cada nova iteração. A equação abaixo expressa a quantidade de análises e comparações C que terão de ser realizadas, neste cenário, para que todas as amostras dos passageiros sejam identificadas e ligadas a cada um dos seus 150 reclamantes:

$$C = \sum_{i=0}^{N-1} \sum_{r=1}^{1} r_i \cdot (N - i) \tag{5.1}$$

Mesmo que seja relacionado a amostra de um reclamante com a de um passageiro, antes que se termine a comparação com o total das $(N - i)$ amostras, a iteração não poderá ser interrompida, pois estamos considerando que poderá existir a possibilidade de termos passageiros com laços de parentesco de 2º grau em diante e da mesma linhagem materna (primos e ou primas por parte da mãe).

O número de análises de sequências e a quantidade de comparações, para associar as amostras dos passageiros aos seus reclamantes, é grandiosa demais ($C = 11325$) para ser realizada em tempo hábil e com segurança com as práticas e ferramentas a que os laboratórios disponham e utilizam até os dias de hoje (capítulos 2 e 3).

O sistema Eva busca contribuir, em relação a estas problemáticas, padronizando e automatizando todo o processo de análise das sequências de DNAmt, o armazenamento e a comparação de seus dados para o estudo de casos forenses (capítulo 1), que envolvam um grande número de corpos e ou amostras biológicas a serem identificadas.

6 *Conclusões e Trabalhos Futuros*

"*Nada se cria,*
nada se perde,
tudo se transforma."
Lavousier

Neste capítulo são apresentados as conclusões e alguns trabalhos futuros ...

6.1 Conclusões

O objetivo primordial deste trabalho foi o de desenvolver uma ferramenta para possibilitar a geração e o armazenamento de perfis de DNA Mitocondrial, com qualidade e segurança, para compor o banco de dados forense e populacional.

Até então, não era possível realizar tal feito devido a ausência de uma ferramenta específica para tratar das sequências e dos dados do DNA Mitocondrial. A falta de integração entre as ferramentas utilizadas pelos laboratórios também contribuia para a deficiência de todo o processo. A maneira como os perfis de DNA Mitocondrial eram gerados e armazenados, em arquivos de texto ou em planilhas, impossibilitava uma gerência adequada para que estes dados pudessem ser analisados e compartilhados entre os laboratórios.

A concepção de se desenvolver um ambiente computacional para dar apóio à análise das sequências do DNA Mitocondrial forense necessitava da integração e automatização das diversas etapas do processo de análise através de um sistema único que pudesse ser comum e acessado pelos diversos laboratórios.

Dentre as diversas problemáticas encontradas durante este estudo, a maior atenção se deu para o desenvolvimento da ferramenta de alinhamento (específica) para as sequências de DNA Mitocondrial. A sua necessidade era extrema, pois todas as outras etapas posteriores de análise dependem do seu serviço. Para isto, pesquisou-se diversos algoritmos e técnicas de alinhamento de sequências (capítulo 3). Esta etapa do estudo foi bastante desgastante e demorada pois diversas sequências, com diferentes características, foram submetidas durante o desenvolvimento do algoritmo de alinhamento para testes.

A definição da estratégia de como realizar o processo de alinhamento também foi um quanto minuciosa devido às diversas abordagens existentes para a comparação de sequências e a elaboração de heurísticas para harmonizar o alinhamento, considerando que existiam dois métodos para atacar o problema:

- O primeiro, através da analise do genoma inteiro do DNA Mitocondrial, alinhando localmente a sua extensão em busca das duas regiões de HVI ou HVII.
- O segundo, ter apenas como referência o segmento destas duas regiões e, consequentemente, adotar a abordagem de um alinhamento global.

A elaboração e o desenvolvimento deste alinhamento, pelo segundo método, tornou possível a geração de perfis de DNA Mitocondrial em formato padronizado, de acordo com as exigências propostas na literatura (seção 2.12).

O DNA Mitocondrial está sujeito a erros, diante das várias etapas e do seu minucioso processo de análise, desde a sua tipagem à geração e o armazenamento de perfis para comparações no bancos de dados. Um ambiente como o proposto neste trabalho não estaria completo sem uma ferramenta de verificação destes possíveis erros. Deixar que apenas a análise filogenética se encarregasse deste processo de verificação ocasionaria na perda de tempo e na dependência de uma outra ferramenta a parte do Eva. Analisar, verificar e validar os polimorfismos de um haplótipo em questão, com os polimorfismos já validados pela literatura, no MITOMAP, através do Eva, foi uma idéia inovadora para a maioria da comunidade forense, como uma ferramenta para livrar o haplótipo da suspeita de erros.

Felizmente, com a ferramenta de alinhamento padronizada e automatizada para gerar o haplótipo de uma amostra em questão, em conjunto com a técnica de validar os polimorfismos, esta abordagem resolveu as quatro primeiras das cinco classes de erros existentes no processo do DNA Mitocondrial. Infelizmente, a quinta classe de erro está relacionada com a recombinação artificial, só podendo então ser detectada através de uma análise filogenética. No entanto esta classe de erro poderá ser evitada, ao se utilizar boas práticas de manuseio das amostras em laboratório.

Para fins de identificação humana através do DNA Mitocondrial, com o intuito de integrar os diversos laboratórios nacionais, fez-se necessário o armazenamento de seus dados em um banco de dados relacional, centralizado, acessado através de um servidor *Web*. O seu armazenamento em banco de dados relacionais possibilita a segurança e a consistência na formatação dos dados, como também permite o relacionamento entre a criação das três bases de dados de pessoas desaparecidas, reclamantes e de perfis populacionais. Esta última base de dados é indispensável em estudos de casos forenses para estimar o peso de uma evidência (seção 2.9).

As 123 amostras tipadas do Estado de Alagoas, utilizadas para compor o banco de dados populacional do Eva, mostraram boa diversidade genética e estão, portanto, aprovadas para serem utilizadas na prática forense para a identificação humana. Cada Estado Brasileiro poderá passar a utilizar o sistema Eva para compor a sua própria base de dados populacional e assim estará contribuindo, desta forma, para aprimorar a estimativa do peso de uma evidência no território nacional.

A realização deste estudo e o desenvolvimento do ambiente computacional demonstraram resultados que avançam muito em relação às práticas exercidas no estudo de casos forenses pela análise do DNA Mitocondrial humano.

A valorização deste projeto se dá ainda mais pela convivência e o trabalho realizado, diretamente, no laboratório de DNA forense da UFAL, em parceria com o Instituto de Computação, destacando a multidisciplinaridade e a aquisição da excelência exercida neste estudo.

6.2 Trabalhos futuros

Apesar do sistema proposto estar modelado para fins de estudos de casos forenses na identificação humana, o estudo do haplótipo do DNA Mitocondrial possui outras finalidades de pesquisa que poderão ser acoplados ao sistema Eva.

As variações do haplótipo, do DNA Mitocondrial, poderão ser tema de estudo em diversas outras ramificações disciplinares, além da ciência forense. Médicos pesquisadores têm ligado a ocorrência de um número de doenças às mutações que poderão ocorrer no genoma mitocondrial.

Estudos evolucionários poderão ser realizados na variação das sequências do DNA Mitocondrial humano com o de outras espécies no esforço de tentar elucidar algum relacionamento entre eles. Um exemplo de um estudo aplicado nesta linha foi a descoberta de que o homem de Neanderthal não é o ancestral do homem moderno, baseado na análise das sequências obtidas da região controle extraídas de suas ossadas (KRINGS et al., 1997).

Antropólogos moleculares estudam as diferenças nas sequências de DNA Mitocondrial de vários grupos populacionais para examinar a ancestralidade e a migração de povos pela história da humanidade. A maioria destes estudos poderão ser realizados utilizando perfis de DNA Mitocondrial. Por exemplo, como foi constatado por Barbosa (2006), das 123 amostras utilizadas para o banco de dados populacional do Eva, 45% são de origem Africana, 27% Ameríndias e 25% Européias, onde apenas 3% delas não puderam ser classificadas. Com isto, podemos concluir que a população de alagoas, em sua maioria (45%), é descendente da linhagem materna do continente Africano (seção 2.11).

Há ainda algumas possíveis melhoras que poderão ser acrescentadas ao Eva para complementar ainda mais o seu ambiente de funcionalidades. Dentre elas, podemos citar:

- A análise das sequências de F e R, que são criadas a partir do sequenciador, para gerar a sequência consenso de HVI e HVII.
- O módulo calculador probabilístico poderá ser incrementado com novos modelos para diferentes tipos de análises estatísticas que poderão ser realizadas nos perfis encontrados no banco, como, por exemplo:
  - Calcular a diversidade genética, a probabilidade de dois indivíduos escolhidos ao acaso apresentarem haplótipos diferentes.

  - Calcular a diversidade de nucleotídeos, a probabilidade de dois nucleotídeos escolhidos ao acaso serem diferentes.
  - Calcular a probabilidade de coincidência, dado que dois indivíduos, tomados ao acaso, exibam o mesmo haplótipo.

- Adicionar novos marcadores do genoma mitocondrial para a identificação humana, como a região HVIII e também o estudo de SNP´s (minúsculas variações que poderão ocorrer pelo genoma).

- Melhorar a interface gráfica do Eva tornando-a casa vez mais intuitiva para o usuário geneticista.

- Realizar um estudo de complexidade dos algoritmos utilizados pelo Eva, visando diminuir o tempo de resposta, quando o banco populacional estiver com uma grande massa de dados.

- Aprimorar e fortalecer, cada vez mais, a segurança do Eva, visto que o sistema se trata de uma aplicação *Web* e será utilizado por diversos laboratórios do país (Brasil) ou até do mundo. Restringir o seu uso, futuramente, através de uma *virtual private network* (VPN) poderá contribuir, ainda mais, para uma maior segurança no seu controle de acesso, visto que o único controle implementado até agora, desta natureza, se resume a uma conta de *Login* com nome de usuário e senha.

Estas e dentre outras modificações futuras irão fortalecer e ajudar a comunidade forense a realizar estudos de casos para identificar amostras e ou restos mortais humanos através da análise do seu DNA Mitocondrial.

Referências

ACHILLI, A. et al. The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. *American Journal of Human Genetics*, v. 75, p. 910–918, 2004.

ALLARD, M. W. et al. Characterization of the caucasian haplogroups present in the SWGDAM forensic mtDNA dataset for 1771 human control region sequences. *Journal of Forensic Science*, v. 47, 2002.

ALLARD, M. W. et al. Characterization of human control region sequences of the African American SWGDAM forensic mtDNA data set. *Forensic Science Internacional*, v. 148, p. 169–179, 2005.

ALLARD, M. W. et al. Control region sequences for East Asian individuals in the Scientific Working Group on DNA Analysis Methods forensic mtDNA data set. *Legal Medicine*, v. 6, p. 11–24, 2004.

ALVES-SILVA, J. et al. The Ancestry of Brazilian mtDNA Lineages. *American Journal of Human Genetics*, v. 67, p. 444–461, 2000.

ANDERSON, S. et al. Sequence and organization of the human mitochondrial genome. *Nature*, v. 290, p. 457–465, 1981.

ANDREWS, R. M. et al. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics*, v. 23, p. 147, 1999.

ATTIMONELL, M. et al. MitBASE: a comprehensive and integrated mitochondrial DNA database. *Nucleic Acids Reserach*, v. 28, p. 148–142, 2000.

BANDELT, H.-J. et al. Detecting errors in mtDNA data by phylogenetic analysis. *International Journal of Legal Medicine*, v. 115, p. 64–69, 2001.

BANDELT, H.-J. et al. The Fingerprint of Phantom Mutations in Mitochondrial DNA Data. *American Journal of Human Genetics*, v. 71, p. 1150–1160, 2002.

BANDELT, H.-J.; SALAS, A.; BRAVI, C. Problems in the mtDNA database. *Science*, v. 305, p. 1402–1404, 2004.

BANDELT, H.-J.; SALAS, A.; LUTZ-BONENGEL, S. Artificial Recombination in Forensic mtDNA population databases. *International Journal of Legal Medicine*, v. 118, p. 267–273, 2004.

BARBOSA, A. B. de G. *Determinação do polimosrfismo das seqüências de DNA mitocondrial humano na população de Alagoas, Brasil*. Dissertação (Mestrado) — UFPE, 2006.

BARNES, M. R.; GRAY, I. C. *Bioinformatics for Geneticists*. [S.l.]: British Library, 2003.

BENTON, D. Bioinformatics: principles and potential of a new multidisciplinary tool. *Trends in Biotechnology*, v. 14, p. 261–272, 1996.

BODENTEICH, A. et al. Dinucleotide repeat in the human mitochondrial D-loop. *Human Molecular Genetics*, v. 1, p. 140, 1992.

BRANDON, M. C. et al. *MITOMAP: a human mitochondrial genome database*. 2004. Último acesso: 2006. Disponível em: <http://www.mitomap.org>.

BRANDSTäTTER, A.; PARSONS, T. Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups. *International Journal of Legal Medicine*, v. 117, p. 291–298, 2003.

BUDOWLE, B. et al. Forensic mitochondrial DNA: applications, debates and foundations. *Annual review of Genomics and Human Genetics*, v. 4, p. 119–143, 2003.

BUDOWLE, B.; BIEBER, F. R.; EISENBERG, A. J. Forensic aspects of mass disasters: Strategic considerations for DNA-based human identification. *Legal Medicine*, v. 7, p. 230–243, 2005.

BUDOWLE, B.; DIZINNO, J.; WILSON, M. Interpretation guidelines for mitochondrial DNA sequencing. In: *Proceedings of the Tenth International Symposium on Human Identification*. Madison - WI: Promega, 1999. Disponível em: <http: www.promega.com/ussymp10proc/default.html>.

BUDOWLE, B.; POLANSKEY, D.; ALLARD, M. W. Addressing the Use of Phylogenetics for Identification of Sequences in Error in the SWGDAM Mitochondrial DNA Database. *Journal of Forensic Science*, v. 49, 2004.

BUDOWLE, B. et al. Mitochondrial DNA regions HVI and HVII Population Data. *Forensic Science Internacional*, v. 103, p. 23–35, 1999.

BUTLER, J. M. *Forensic DNA Typing*. 2ª. ed. [S.l.: s.n.], 2005.

CAETANO, L. H. T. et al. A Bioinformatic tool to assist in human mtDNA profiles for forensic purposes in Brazil. In: *X-meeting 1º Internacional Conference on the AB3C*. [S.l.: s.n.], 2005.

CAETANO, L. H. T. et al. Development of a Bioinformatic tool for analysis, data storage and comparisons of human mtDNA profiles for forensic purposes in Brazil. In: *BIOMAT International Symposium on Mathematical and Computational Biology*. [S.l.: s.n.], 2005.

CARRACEDO, A. et al. Results of a collaborative study of the EDNAP group regarding the reproducibility and robustness of the Y-chromosome STRs DYS19, DYS389 I and II, DYS390 and DYS393 in a PCR pentaplex format. *Forensic Science International*, v. 119, p. 24–41, 2001.

CARRACEDO, A. et al. DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA. *Forensic Science Internacional*, v. 110, p. 79–85, 2000.

CARRACEDO, A. et al. Reproducibility of mtDNA analysis between laboratories: a report of the European DNA profiling group (EDNAP). *Forensic Science International*, v. 97, p. 165–170, 1998.

CHEN, X. et al. Rearranged mitochondrial genomes are present in the human oocytes. *American Journal of Human Genetics*, v. 57, p. 239, 1995.

COBLE, M. D. et al. Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. *Internacional Journal of Legal Medicine*, v. 118, p. 137–146, 2004.

DEBENHAM, P. Heteroplasmy and the Tsar. *Nature*, v. 380, p. 484–485, 1996.

DENNIS, C. Error reports threaten to unravel databases of mitochondrial DNA. *Nature*, v. 121, p. 773–774, 2003.

EMBL-EBI. *European Bioinformatic Institute*. 2005. Disponível em: <http://www.ebi.ac.uk/Tools/>.

FORSTER, P. *To Err is Human*. [S.l.]: Annals of Human Genetics, 2003.

GIBAS, C.; JAMBECK, P. *Desenvolvendo Bioinformática*. [S.l.]: O'Reilly, 2001.

GILL, P. et al. Identification of the remains of the Romanov family by DNA analysis. *Nature Genetics*, v. 6, p. 130–135, 1994.

HOLLAND, M.; PARSONS, T. Mitochondrial DNA Sequence Analysis Validation and use for Forensic Casework. *Forensic Science Rev.*, v. 11, p. 21–50, 1999.

INGMAN, M. et al. Mitochondrial genome variation and the origin of modern humans. *Nature*, v. 408, p. 708–713, 2000.

ISENBERG, A. *Forensic Mitochondrial DNA Analysis*. [S.l.]: Forensic Science Handbook, 2004.

JIN, H. J. et al. Forensic genetic analysis of mitochondrial DNA hypervariable region I/II sequences: An expanded Korean population database. *Forensic Science International*, v. 158, p. 125–130, 2006.

KOST, S. *An Introduction to SQL Injection Attacks for Oracle Developers*. January 2004. Disponível em: <http://www.integrigy.com>.

KRINGS, M. et al. Neandertal DNA Sequences and the Origin of Modern Humans. *Cell*, v. 90, p. 19–30, 1997.

LECOMPTE, O. et al. Multiple alignment of complete sequences (MACS) in the post-genomic era. *Gene*, v. 270, p. 17–30, 2001.

LESK, A. M. *Introduction to Bioinformatics*. [S.l.]: British Library, 2002.

MELTON, T. Mitochondrial DNA heteroplasmy. *Forensic Science Review*, v. 16, p. 1–20, 2004.

MILLER, K. W.; BROWN, B. L.; BUDOWLE, B. The Combined DNA Index System. In: *International Congress Series*. [S.l.: s.n.], 2003. p. 617–620.

MONSON, K. L. et al. *The mtDNA population database: an integrated software and database resource*. 2002. Forensic Science Communications [online]. Disponível em: <http: www.fbi.gov hq lab fsc backissu april2002 miller1.htm>.

MONTERA, L. *Regiões ortólogas em múltiplos genomas*. Dissertação (Mestrado) — UFMS, 2004.

NEEDLEMAN, S.; WUNSCH, C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, v. 48, p. 443–453, 1970.

PARSON, W. et al. The EDNAP mitochondrial DNA population database (EMPOP) collaborative exercises: organisation, results and perspectives. *Forensic Science International*, v. 139, p. 215–226, 2004.

PARSON, W. et al. EMPOP - the EDNAP mtDNA population database concept for a new generation, high quality mtDNA database. *Internacional Congress Series*, v. 1261, p. 106–108, 2004.

PARSONS, T. J. et al. A high observed substitution rate in the human mitochondrial DNA control region. *Nature Genetics*, v. 15, p. 363–368, 1997.

PEVZNER, P. A. *Computational Molecular Biology*. [S.l.]: MIT press, 2001.

RAND, S.; SCHüRENKAMP, M.; BRINKMANN, B. The GEDNAP (German DNA profiling group) blind trial concept. *International Journal of Legal Medicine*, v. 116, p. 199–206, 2002.

RöHL, A. et al. An annotated mtDNA Database. *International Journal of Legal Medicine*, v. 115, p. 29–39, 2001.

RUIZ-PESINI, E. et al. Effects of Purifying and Adaptive Selection on Regional Variation in Human mtDNA. *Science*, v. 303, p. 223–226, 2004.

SALAS, A. et al. A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. *Biochemical and Biophysical Research Communications*, v. 335, p. 891–899, 2005.

SALAS, A.; LAREU, M.; CARRACEDO, A. Heteroplasmy in mtDNA and the weight of evidence in forensic mtDNA analysis: a case report. *International Journal of Legal Medicine*, v. 114, p. 186–190, 2001.

SANKOFF, D. Minimal mutation trees of sequences. *Journal on Applied Mathematics*, v. 28, p. 35–42, 1975.

SATOH, M.; KUROIWA, T. Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. *Experimental Cell Research*, v. 196, p. 137–140, 1991.

SCHNEIDER, S. et al. *Arlequin ver. 2.000: A software for population genetics data analysis*. University of Geneva, Switzerland.

SETUBAL, J.; MEIDANIS, J. *Introduction to Computacional Molecular Biology*. [S.l.]: Brooks, United States., 1997.

SILVA, L. A. F. da; PASSOS, N. S. *DNA Forense: Coleta de Amostras Biológicas em locais de Crime para Estudo do DNA*. 1°. ed. [S.l.]: edUFAL, 2002.

SMITH, T. F.; WATERMAN, M. S. Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, v. 147, p. 195–197, 1981.

STEWART, J. et al. Length variation in HV2 of the human mitochondrial DNA control region. *Journal of Forensic Science*, v. 46, p. 862–870, 2001.

STONEKING, M. et al. Population variation of human mitochondrial DNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *American Journal of Human Genetics*, v. 48, p. 370–382, 1991.

STONEKING, M. et al. Establishing the identity of Anna Anderson Manahan. *Nature Genetics*, v. 9, p. 9–10, 1995.

SWGDAM. Guidelines for the mitochondrial DNA (mtDNA) nucleotideo sequence interpretation. *Forensic Science Communication*, v. 5, 2003. Disponível em: <http: www.fbi.gov hq lab fsc backissu april2003 swgdammitodna.htm>.

TICONA, W. G. C. *Aplicação de algoritmos genéticos multi-objetivo para alinhamento de sequências biológicas*. Dissertação (Mestrado) — USP, 2003.

TULLY, G. et al. Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretations of mitochondrial DNA profiles. *Forensic Science Internacional*, v. 124, p. 83–91, 2001.

TULLY, L. A. et al. A sensitive denaturing gradient-gel electrophoresis assay reveals a high frequency of heteroplasmy in hypervariable region 1 of the human mtDNA control region. *The American Journal of Human Genetics*, v. 67, p. 1029–1032, 2000.

WALLACE, D.; BROWN, M.; LOTT, M. Mitochondrial DNA variation in human evolution and disease. *Gene*, v. 238, p. 211–230, 1999.

WILSON, M. R. et al. Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region. *Forensic Science Internacional*, v. 129, p. 35–42, 2002.

WILSON, M. R. et al. Guidelines for the use of mitochondrial DNA sequencing in forensic science. *Crime Laboratory Digest*, v. 20, p. 68–77, 1993.

WITTIG, H. et al. Mitochondrial DNA in the central european population human identification with the help of the forensic mtDNA DLoop-Base Database. *Forensic Science International*, v. 113, p. 113–118, 2000.

YAO, Y.-G.; BRAVI, C. M.; BANDELT, H.-J. A call for mtDNA data quality control in forensic science. *Forensic Science International*, v. 141, p. 1–6, 2004.

Este documento foi preparado utilizando LaTeX em conjunto com o editor LyX.
As referências bibliográficas foram administradas com o editor JabRef.
As referências estão no estilo ABNT da norma 6023.
Contato: zootalk@gmail.com