RITA DE CÁSSIA DE LIMA IDALINO

HOMOLOGIAS EM GENES RELACIONADOS À RESISTÊNCIA À MASTITE EM VACAS, OVELHAS E CABRAS

RECIFE-PE

DEZEMBRO/2010

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

HOMOLOGIAS EM GENES RELACIONADOS À RESISTÊNCIA À MASTITE EM VACAS, OVELHAS E CABRAS

Dissertação apresentada ao Programa de Pós-Graduação em Biometria e Estatística Aplicada como exigência parcial à obtenção do título de Mestre.

Área de Concentração: Modelagem Estatística e Computacional

Orientador: Prof. Dr Kleber Régis Santoro

RECIFE-PE

DEZEMBRO/2010

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

HOMOLOGIAS EM GENES RELACIONADOS À RESISTÊNCIA À MASTITE EM VACAS, OVELHAS E CABRAS

RITA DE CÁSSIA DE LIMA IDALINO

Dissertação julgada adequada para obtenção do título de mestre em Biometria e Estatística Aplicada, defendida e aprovada por unanimidade em 20/12/2010 pela Comissão Examinadora.

Orientador:

_____

Prof. Dr. Kleber Régis Santoro
Universidade Federal Rural de Pernambuco

Banca Examinadora:

_____

Prof. Dr. Manoel Adrião Gomes Filho
Universidade Federal Rural de Pernambuco

_____

Profª. Dra. Tatijana Stosic
Universidade Federal Rural de Pernambuco

_____

Prof. Dr. Tiago Alessandro Espínola Ferreira
Universidade Federal Rural de Pernambuco

Ficha catalográfica

I18h Idalino, Rita de Cássia de Lima
Homologias em genes relacionados à resistência à mastite em vacas, ovelhas e cabras / Rita de Cássia de Lima Idalino. – 2010.
79 f.: il.

Orientador: Kleber Régis Santoro.
Dissertação (Mestrado em Biometria e Estatística Aplicada) - Universidade Federal Rural de Pernambuco, Departamento de Estatística e Informática, Recife, 2010.
Inclui referências, anexo e apêndice.

1. Homogeneidade 2. HMM 3. Seqüências genéticas 4. Biometria I. Santoro, Kleber Régis, orientador II. Título

CDD 574.018

Dedico

Ao meu esposo Jeremias Leão,
pela motivação e incansável dedicação
...disciplina, disciplina e disciplina.

Agradecimentos

Agradeço a Deus, por ter chegado até aqui.

A minha família que sempre esteve ao meu lado e que soube conviver com minha ausência.

Ao professor Kleber pela orientação e amizade e aos demais professores do programa de pós-Graduação em Biometria e Estatística Aplicada pelos ensinamentos e o convívio.

À Zuleide e a Marco, sempre dispostos a ajudar.

À Dona Marlene, Andréia, Camilo, Melzinho e Leocadia que conviveram comigo nos primeiros meses em Recife. Obrigada!

Aos amigos Dâmocles, Rogério (Coleguinha), Gabriel, Samuel, Cícero, Rodrigo, Alvino, Amanda, Jéssika e Anderson pelos bons momentos, vocês são amizades pra uma vida inteira.

À Soraya Fárias, por toda paciência em me explicar o mundo da genética. Obrigada Amada.

Agradeço imensamente a Dani, minha Paixão! Obrigada por todos os momentos, por gastar horas estudando comigo. Deixo aqui todo meu carinho e respeito por toda família Roges (Lú, Andréia, Ricardo, Socorro, Cassinha, Aninha, Mel, Negão e Costela). Vocês são uma verdadeira família pra mim.

À Luciano (Luc) por todo carinho e disposição para ajudar. Uma pessoa com o coração do tamanho do mundo.

Aos meus ex-alunos da UFRPE. Obrigada a cada um que tive a oportunidade de ensinar um pouquinho de Estatística.

À Josimar, por ter me incentivado a ter vindo até aqui. Você tem grande responsabilidade nisso tudo.

À Julianne, alguém que cuidou muito de mim nos momentos mais difíceis que passei estando longe de casa.

À Raphael, amigo de todas as horas que preciso e até quando não preciso. Obrigada Rapha!

Aos amigos que ficaram em Natal, mas que nunca me deixaram de lado, Eliza, Paulinha, Maricélia, Leandro, Mirlândia, Mikarla, Jusci, Dany e Michell.

Por fim, agradeço ao meu amado esposo, Jeremias, que trabalhou muito comigo. Deixo aqui toda minha admiração, respeito e Amor que tenho por você. Te Amo!

Ao apoio financeiro do CNPq (Processo 505912/2008-2).

Eu não sei se você se recorda do seu primeiro caderno
Eu me recordo do meu
Com ele eu aprendi muita coisa
Foi nele que descobri que a experiência dos erros,
É tão importante quanto à experiência dos acertos

Por que vistos de um jeito certo, os erros, nos preparam para nossas vitórias e conquistas futuras. Por que não à aprendizado na vida que não passe pela experiência dos erros

Caderno é uma metáfora da vida, quando os erros cometidos eram demais eu me recordo que nossa professora nos sugeria que a gente virasse a pagina
Era um jeito interessante de descobrir a graça que há nos recomeços
Ao virar a pagina os erros cometidos deixavam de nos incomodar e a partir deles a gente seguia um pouco mais crescido

O caderno nos ensina que erros não precisam ser fontes de castigos
Erros podem ser fontes de virtudes
Na vida é a mesma coisa
O erro tem que esta a serviço do aprendizado
Nenhum tem que ser fonte de culpas ou de vergonhas.
Nenhum ser humano pode ser verdadeiramente grande sem que seja capaz de reconhecer os erros que cometeu na vida.

Uma coisa é a gente se arrepender do que fez
Outra coisa é a gente se sentir culpado
Culpas nos paralisam, arrependimentos não.
Eles nos lançam pra frente, nos ajuda a corrigir os erros cometidos.

Deus é semelhante a um caderno
Eles nos permite os erros pra que a gente aprenda pra fazer do jeito certo
Você tem errado muito? Não importa. Aceite de Deus
esta nova pagina de vida que tem nome de hoje
Recorde-se das lições do seu primeiro caderno
Quando os erros são demais vire a pagina

O Caderno, por Pe. Fabio de Melo

Resumo

Diante da grande massa de dados que é gerada na área da genética molecular, é de suma importância que técnicas que possibilitem a organização e interpretação desses dados sejam desenvolvidas e amplamente divulgadas. Inicialmente, neste trabalho, foi realizada uma análise da composição de três sequências genéticas, das espécies Bovina (*Bos taurus*), Caprina (*Capra hircus*) e Ovina (*Ovis aries*), em seguida aplicamos técnicas de alinhamentos para identificação de similaridades entre estas. Posteriormente, utilizamos a teoria das cadeias de Markov com estados ocultos, HMM's (Hidden Markov Models), na aplicação do problema de discriminação de regiões homogêneas em sequências de DNA. Utilizamos o algoritmo de Viterbi como uma ferramenta auxiliar para obtenção de regiões homogêneas e em seguida o algoritmo *Baum-Welch* para maximizar as probabilidades de uma sequência de observações. Foram analisados trechos dos genes HSP70.1 e NRAMP-1 para três espécies diferentes.

Palavras-chave:.homogeneidade, HMM, sequências genéticas

Abstract

Given the large amount of data that is generated in the field of molecular genetics, is of paramount importance that techniques which allow the organization and interpretation of such data be developed and widely disseminated. Initially, we carried out a composition analysis of three gene sequences of the species: ox (*Bos taurus*), goat (*Capra hircus*), and sheep (*Ovis aries*), then we applied alignment techniques for identification of similarities between them. Subsequently, we used the Markov Chain theory with hidden states, i.e. Hidden Markov Models (HMMs, hereafter), in the application of discrimination problem of homogeneous regions in DNA sequences. We used the Viterbi algorithm as an auxiliary tool to obtain homogeneous regions, and then the Baum-Eelch algorithm in order to maximize the probability of a sequence of observations. We analyzed portions of HSP70.1 and NRAMP-1 genes for three different species.

Keywords: homogeneity, HMM, gene sequences

SUMÁRIO

LISTA DE FIGURAS

CAPÍTULO 1

CAPÍTULO 2

CAPÍTULO 3

LISTA DE TABELAS

CAPÍTULO 2

CAPÍTULO 3

CAPÍTULO 1

1. Introdução

1.1 Motivação

Durante os últimos anos ocorreu um avanço acentuado no estudo da evolução em nível molecular devido ao desenvolvimento de técnicas para o estudo do DNA, como por exemplo, o sequênciamento automático de ácidos nucléicos (WATSON, 1992). O sequênciamento de DNA é um processo que determina a ordem dos nucleotídeos (blocos que constituem a molécula de DNA) e a partir dessa ordem são geradas as diversas características presentes nas espécies.

Em 25 de Abril de 1953, James Watson e Francis Crick publicaram sobre a descoberta da molécula de DNA, o que mudou a história da ciência. Nesta época, já se sabia que a molécula de DNA transporta as informações genéticas e se conhecia sua composição, porém ninguém nunca tinha sido capaz de desvendar a sua estrutura. E esse desafio foi vencido por Watson e Crick (WATSON & BERRY, 2005)

. O estudo de genomas completos teve inicio com a proposta de utilizar a tecnologia de DNA para expandir o conceito básico de mapeamento genético proposto por A. H. Sturtevant no começo do século XX. O desenvolvimento de mapas genéticos tem sido um dos objetivos dos geneticistas desde que foi percebido que estas são ferramentas de fundamental importância nos estudos de herança genética. A construção de mapas genéticos foi viabilizada após a descoberta, no início do século passado, da ligação entre genes e da localização desses nos cromossomos (GRIFFIN et al, 2005).

Em meados da década de 80 houve uma grande revolução na área da genética médica quando mais de 1.000 genes que causam doenças no homem foram mapeados. O sequênciamento genético completo de algumas espécies permitiu observar que o número de genes distintos necessários para o

desenvolvimento de um organismo complexo como o ser humano (< 40.000 genes) não é muito maior do que o de um genoma de um eucarioto como a planta arabidopsis (~25.000 genes) e, possivelmente, inferior ao de outras plantas como o arroz (> 40.000 genes) (WATTERMAN, 1995).

Os seres vivos compartilham um sistema comum de armazenamento e expressão da informação genética e demonstram homologia (semelhança com base na proximidade evolutiva) em muitas estruturas, até mesmo nos próprios genes. Com isso, percebe-se que estudos envolvendo a genética molecular são de grande valia para entender a proximidade que existe entre as espécies.

1.2 As pesquisas genômicas

A pesquisa genômica tem como principal objetivo a caracterização molecular de genomas na sua totalidade. Vários métodos vêm sendo utilizados recentemente em estudos de larga escala, os quais envolvem desde o conhecimento do genoma completo de diferentes organismos, até o estudo de um gene particular que possa ser responsável pela expressão de uma doença ou determinada característica em questão (GIBSON & MUSE, 2004). Estas pesquisas podem ser divididas em duas vertentes principais: a genômica estrutural, caracterizando a natureza física de genomas completos e a genômica funcional, caracterizando as regiões codificadoras e padrões globais de expressão de genes (GRIFFITHS et al., 1999). A caracterização de genomas completos é importante por dois fatores. O primeiro possibilita a obtenção da visão global da arquitetura genética de um organismo e, o segundo, provê todos os dados para a descoberta de novos genes, como por exemplo, os que estão relacionados ao surgimento de uma determinada doença (WATSON, 1992).

A análise genética é possível em qualquer organismo. As diferenças genéticas que são comuns entre organismos da mesma espécie são chamadas de polimorfismos genéticos, enquanto que as diferenças genéticas acumuladas entre espécies constituem as divergências genéticas. A informação hereditária de todos os organismos vivos encontra-se presente nas moléculas de DNA

(Acido desoxirribonucléico). O DNA é uma molécula que existe dentro das células de todos os seres vivos, desde as bactérias, fungos e protozoários até os animais e plantas, e contém as informações necessárias para formar um ser vivo e para que ele possa se reproduzir. O DNA é como um código de letras, que ao ser interpretado pela célula, produz os componentes que fazem parte do nosso corpo. O DNA é constituído de duas cadeias complementares, cada uma formada por quatro tipos de nucleotídeos: adenina (A), guanina (G), timina (T) e citosina (C) que são ligados por pontos de hidrogênio. A localização física do nucleotídeo na sequência de DNA é denominada sítio (WATSON, 1992; GRIFFITHS, 2000).

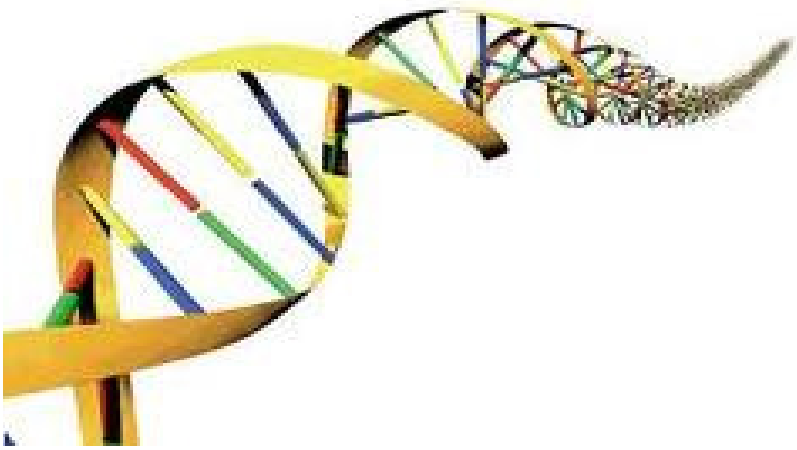

Figura 1: Representação da estrutura do DNA

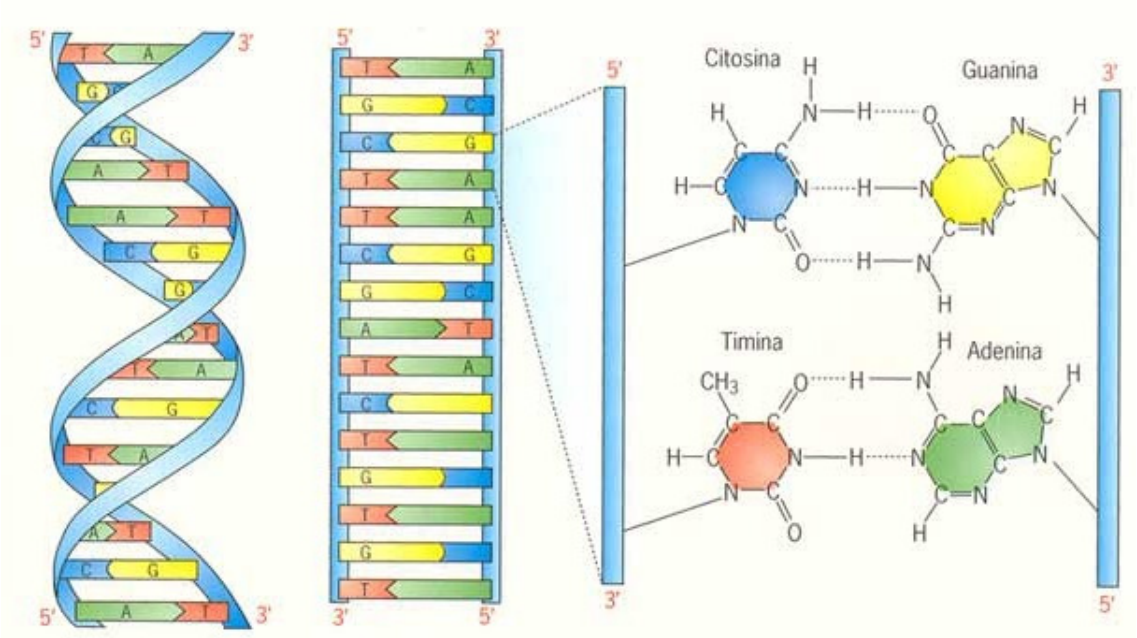

Figura 2: Representação da estrutura da ligação química das bases no DNA

O gene é uma unidade hereditária, situada no cromossomo, e que determina as características de um indivíduo. A totalidade de DNA numa célula é o genoma. Os genes são dispostos em uma ordem linear ao longo de corpúsculos filamentosos chamados cromossomos. A localização física de um gene no DNA recebe o nome de locus. As variantes de um gene em um determinado locus são chamadas de alelos. Grandes volumes de dados são gerados em pesquisas genéticas que tem, por sua vez, exigido o desenvolvimento de métodos ligados à área de estatística e de bioinformática para análise desses dados (WATTERMAN, 1995).

1.3 Mastite animal e seu impacto sobre a pecuária

A mastite é considerada a doença que acarreta os maiores prejuízos econômicos à produção leiteira, pela redução da quantidade e pelo comprometimento da qualidade do leite produzido, ou até pela perda total da capacidade secretora da glândula mamária. Esta doença é dada pela inflamação da glândula mamária que pode ser causada por muitos fatores, sendo os agentes infecciosos, principalmente as bactérias, os mais importantes. A mastite pode ser classificada como clínica ou subclínica. A mastite clínica apresenta sinais evidentes, tais como: edemas, aumento de temperatura, endurecimento, dor na glândula mamária, grumos, pus ou qualquer alteração das características do leite. Na forma subclínica não se observam alterações macroscópicas e sim alterações na composição do leite; portanto, não apresenta sinais visíveis de inflamação da glândula mamaria. (BRANT & FIGUEIREDO, 1994)

No Brasil pode-se afirmar que a mastite subclínica está presente em boa parte dos rebanhos leiteiros (MACHADO et al., 2000). A mastite subclínica pode ser detectada pela contagem direta ou indireta de células somáticas no leite. Estas são compostas basicamente por dois tipos de células de descamação do epitélio secretor e leucócitos de origem do sangue, sendo que estas se apresentam com elevadas concentrações nos casos de mastite.

No Brasil, segundo BRANT & FIGUEIREDO (1994), a mastite subclínica caracteriza-se pela alta incidência, com índices variando de 44,88% a 97,0%, e a redução da produção de leite situa-se entre 25,4% e 43,0%. A mastite é uma doença complexa. A resposta inflamatória causada por esta doença pode ser causada, também, por fatores químicos, físicos ou traumáticos. Ela é considerada uma doença multifatorial, em que os fatores de risco podem estar relacionados ao hospedeiro (condições fisiológicas e genéticas), ao microrganismo patogênico e ao ambiente, que pode contribuir para a ocorrência da doença (LEBLANK et al., 2006). As conseqüências mais sérias estão relacionadas com as perdas econômicas, acarretada pela diminuição de qualidade e quantidade do leite, aumento nos custos de tratamentos e serviços veterinários, e nos casos clínicos, no descarte de toda a produção de leite ou até mesmo em casos extremos, na morte do animal (NONNECKE e HARP, 1988).

O *California Mastitis Test* (CMT) é um dos testes mais usuais para o diagnóstico da mastite subclínica, sendo um indicador indireto da contagem de células somáticas no leite. Este consiste na coleta de leite dos quartos mamários, individualmente, em uma bandeja apropriada, adicionando-se um detergente aniônico neutro, que atua rompendo a membrana das células e liberando o material nucléico (DNA), que apresenta alta viscosidade. De acordo com a intensidade da reação classifica-se em: negativa (0), reação leve (+), moderada (++) e intensa (+++) (FONSECA & SANTOS, 2000).

Figura 3: Animais com mastite em vários estágios da doença

Imagens gentilmente cedidas pelo professor Rinaldo Aparecido Mota (Professor do Departamento de Medicina Veterinária - UFRPE)

1.4 Melhoramento genético animal

O melhoramento animal tem por finalidade a utilização da variação genética entre os indivíduos, para aumentar qualitativa e quantitativamente a produção dos animais domésticos. O principal efeito do melhoramento genético é o aumento da herança desejável na população, ou seja, escolhendo-se sempre os animais mais produtivos, o rebanho, como um todo, será beneficiado com um ganho genético, ou seja, a seleção de modo geral, tem o objetivo de progresso e/ou fixação de alguma característica de importância. Isso quer dizer que ela tem por finalidade, aumentar na população, a freqüência de alelos favoráveis. Todo programa de melhoramento genético deve visar à maior qualidade dos animais em termos de produtividade, como também, maior qualidade do produto oferecido (TONHATI et al.,1996).

O fenótipo de um indivíduo nada mais é que o produto da interação genótipo e meio ambiente, e apesar de ter a fundamentação teórica desenvolvida há muitos anos, tem recentemente, recebido grandes contribuições que são, com a necessidade de melhoria genética imposta pelo mercado, as principais responsáveis tanto pela expansão quanto pelos progressos genéticos que têm sido observados nas mais diferentes espécies de animais domésticos explorados comercialmente. Contudo, a melhoria genética se processa com base na escolha correta daqueles que participam, ou melhor, daqueles aos quais é dada a possibilidade de participar, do processo de constituição da geração seguinte. Isso vale para a escolha dos indivíduos que produzirão filhos, ou mesmo, para escolha de raças, sendo assim chamado à escolha dos indivíduos, de processo de seleção, que é importante para melhoria de raças puras ou para cruzamentos; e a escolha de raças propriamente dita que orienta e sinaliza o sucesso dos cruzamentos. (CRUZ & REGAZZI, 2001)

1.5 Genéticas da resistência à mastite e os genes candidatos

Devido ao grande número de vias metabólicas, moleculares e células envolvidas na característica de resistência. De forma mais geral, define-se a resistência como sendo a habilidade que o hospedeiro tem de evitar a infecção ou de rapidamente recuperar-se dela (DETILEUX et al., 2002). Num sentido mais restrito, a resistência pode ser definida com sendo a capacidade do animal em impedir o estabelecimento, sobrevivência e/ou desenvolvimento de uma bactéria em particular (Richard & Riollet, 2006).

Existem, no entanto, recursos genéticos que, se forem identificados adequadamente podem contribuir para os sistemas de produção animal. Muitos aspectos da forma do corpo, do funcionamento dos órgãos e dos comportamentos dos animais e dos seres humanos são transmitidos por hereditariedade. Cada novo indivíduo recebe ao se formar um conjunto de cromossomos do pai e outro da mãe sendo restabelecido o número de cromossomos da espécie. Na reconstituição dos cromossomos, a predominância ou não dos genes para uma mesma característica determinará se ela será expressa ou não (VIDAL et al, 1993).

O estudo da genética molecular permite identificar genes com responsabilidade de conferir determinadas características. A resistência à mastite, que é o principal problema de saúde relacionado à produção de leite, tanto no Brasil quanto no mundo, visando o aumento da saúde da glândula mamária.

Até o momento, apesar das constantes pesquisas em genética molecular, não foi encontrado um gene que determine à resistência à mastite. O que têm-se encontrado é um conjunto de genes, alguns altamente polimórficos que conjuntamente são responsáveis pela resposta imunológica dos animais (GRIFFIN et al., 2005; LEYVA-BACA et al., 2008).

Nas últimas décadas vários estudos vêm sendo realizados na espécie bovina a fim de identificar regiões cromossômicas ou genes associados à resistência a essa doença. Na reconstituição dos cromossomos, a predominância ou não dos genes para uma mesma característica determinará se ela será expressa ou não (LEYVA-BACA et al, 2008).

A proteína do complexo de choque térmico (*Heat Shock Protein* - HSP), HSP70-1 tem demonstrado sua ação de proteção à célula pela termo-tolerância e habilidade coadjuvante em apoiar a sobrevivência da célula ao estresse oxidativo, estando envolvidas em vários processos essenciais para a função celular e em diferentes aspectos da reprodução em muitas espécies (ADAMOWICZ et al., 2005). Estudos já demonstraram que a variabilidade existente entre as raças de determinadas espécies pode ser explorada com o objetivo de obter animais adaptados ao estresse térmico, baseando-se nos mecanismos celulares que os detém (HANSEN, 2004).

Considera-se também como um fator de resistência natural associado a macrófagos (*Natural resistance – associated macrhophage protein* - NRAMP), o qual possui ação de resistência contra organismos intracelulares. A partir de alguns estudos, foi sugerido que a proteína formada teria função na membrana fagolisossomal como uma concentradora de produtos oxidativos, mediando atividades contra os parasitas ingeridos dos macrófagos infectados (VIDAL et al., 1993). Experimentos demonstraram que vacas resistentes produziram mais NRAMP-1 que as susceptíveis, e que a expressão da razão dessa proteína foi alta em vacas resistentes sendo possível a aplicação da seleção para resistência à mastite em vacas antes de serem infectadas baseando-se na expressão do NRAMP-1 (JOO et al., 2003).

O gene BoLA (*Bovine Lymphocyte Antigen*), que se localizam no Complexo Principal de Histocompatibilidade (MHC – *Major Histocompatibility Complex*) do genoma bovino e que são altamente polimórficos, estão envolvidos nos processos celulares relacionados ao sistema imunológico dos animais. Eles são responsáveis por codificar as proteínas presentes na superfície das células e envolvidas na relação entre antígenos e anticorpos. O

extenso polimorfismo dos genes BoLA dificulta uma genotipagem correta dos animais. O gene BoLA-DRB3 desta família está envolvido no processo molecular de resistência à mastite (SHARIF et al., 2003).

O loco BoLA tem sido amplamente estudado nos últimos 20 anos em razão de sua influência sobre as características produtivas e às relacionadas à saúde animal. O efeito sobre a saúde animal pode ser resultante da ação direta dos alelos BoLA sobre as funções imunológicas. O produto do gene BoLA-DRB3 é uma proteína relacionada com a formação do complexo antígeno-anticorpo associada com a resposta imunológica específica (STEAR, et al,. 1989).

1.6 Comparação de sequências genéticas

O objetivo principal em se comparar os genes de várias espécies é avaliar o grau de similaridade (ou diferença) entre eles. Quanto mais parecidas forem duas espécies, mais semelhantes serão seus genes.

O gene escolhido para suporte das análises filogenéticas deve estar desempenhando a mesma função nos diferentes organismos avaliados. Genes que adquirem novas funções em uma determinada espécie estão sujeitos a novas pressões evolutivas e por isto acumulam diferenças numa velocidade diferente. Desta maneira é possível construir a história evolutiva dos genes cuja função foi conservada e em seguida correlacionar estes resultados com a história evolutiva das espécies.

A análise de seqüências genéticas revolucionou o estudo de várias áreas de pesquisa. É comum hoje a estimativa de árvore filogenéticas relacionando organismos com base na similaridade dos seus genes e não da sua morfologia. As informações até o momento apreendidas dos projetos genoma são inúmeras. Por exemplo, comparações de seqüências gênicas revelaram que proteínas altamente similares são codificadas nos genomas de organismos tão distantes evolucionariamente quanto a levedura e os seres humanos.

1.7 Alinhamentos

Com o objetivo de entender o quanto duas seqüências são semelhantes, utilizam-se métodos de comparação de cadeias de símbolos, e que genericamente são conhecidos pela designação de alinhamento de seqüência.

Um problema de importância em Biologia Molecular é realizar um estudo comparativo de um conjunto de sequências de bases nitrogenadas ou de aminoácidos. Uma maneira de efetuar a comparação destas sequências é calcular um alinhamento entre elas. Intuitivamente falando, um alinhamento é uma maneira de inserir espaços nas sequências de forma que elas fiquem com o mesmo comprimento e, possam, desta maneira, ser facilmente comparadas. (BRITO, 2001)

Em geral, as moléculas que se consideram em alinhamentos são moléculas de DNA, de RNA e de proteínas. Como tais moléculas são polímeros que podem ser representadas de maneira fácil por uma sequência de caracteres, comparar moléculas resume-se, na prática, a fazer uma comparação das sequências correspondentes.

O problema de encontrar alinhamentos entre sequências ocupa uma posição de destaque em Biologia Computacional: alinhamentos são usados para comparações de sequências, para construção de árvores evolutivas (“árvores filogenéticas”) e para predição de estrutura secundária de moléculas de RNA e de proteínas (BRITO, 2003).

Uma vez realizado o sequenciamento, é preciso identificar e agregar informação a sequência. A melhor maneira de fazer isto é comparar a sequência com sequências já conhecidas. Para isso, é feito o alinhamento par a par com um banco de dados. Uma vez que um conjunto se sequências foram alinhadas, será possível encontrar as seguintes características:

- ✓ **Identidade:** número que indica a quantidade de nucleotídeos alinhados
- ✓ **Similaridade**: considera a probabilidade do alinhamento ter ocorrido por acaso (*e-value*). Considera todos os outros possíveis alinhamentos
- ✓ **Homologia**: dividem a mesma ancestralidade com significado evolutivo

No contexto de evolução as sequências de DNA sofrem mutações. Estas modificações locais entre os nucleotídeos podem ser:

- ✓ **Inserções**: inserção de uma base ou várias bases na sequência
- ✓ **Deleções**: deleção de uma base ou mais bases na sequência
- ✓ **Substituições**: substituição de uma base por outra

Considerando $\Sigma = \{A, C, T, G\}$ o alfabeto correspondente as bases nitrogenadas presentes em sequências de DNA e que sobre Σ são fragmentos de DNA de alguma espécie em análise ou que Σ é o alfabeto correspondente ao conjunto de aminoácidos e que uma sequência sobre Σ e uma proteína (ou fragmento de uma proteína).

Com o intuito de entender como se dá a realização de um alinhamento consideremos o seguinte: Seja k um inteiro positivo e $s_1, s_2, \ldots, s_k$ várias sequências sobre Σ. Um alinhamento A de $s_1, s_2, \ldots, s_k$ é uma matriz $A = (A_{ij})$ de dimensões $k \times n$.

Considerando como um alinhamento A de $s_1, s_2, \ldots, s_k$ a matriz $A = (A_{ij})$ de dimensões $k \times n$ tal que, para cada i, existe um conjunto $J_i = \{j_1, j_2, \ldots j_{n_i}\} \subseteq \{1, \ldots, n\}$, como $j_1 < j_2 < \cdots < j_{n_i}$ e tal que $A_{ij1} A_{ij2} \cdots A_{ijn_i} = s_i$ e tal que para todo $j \in \{1, \ldots, n\} - J_i$, temos $A_{ij} = \lfloor \Theta \rfloor$. Dizemos que dois caracteres $s_i[j] = s_{i'}[j']$ estão alinhados em A se $s_i[j] = s_{i'}[j']$ estão na mesma coluna de A.

Uma vez que *n* sequências são alinhadas é possível obter uma propriedade de fundamental importância para realizar uma indicação inicial de determinada característica que esta sendo analisada, no caso, a mastite.

A seqüência consenso refere-se à região de consenso em seqüências alinhadas de forma a maximizar suas homologias. Para que uma seqüência seja aceita como consenso, cada base individual deve ser razoavelmente

predominante na respectiva posição. Estando alinhadas *n* sequências, ao final desse alinhamento, será identificada uma nova sequência com as similaridades, polimorfismo e os *gap*s (espaços inseridos na sequência consenso quando não é possível identificar a combinação entre o alinhamento das sequências que estão sendo alinhadas). A sequência consenso é de fundamental importância para identificar que indivíduos que sofrem mutações nessas sequências normalmente expressam menos uma dada característica num determinado gene.

1.8 Alinhamento global e local

O alinhamento global tem como objetivo tentar alinhar as sequências completamente, usando o máximo de caracteres possíveis em toda a extensão. Seqüências que possuem alguma similaridade e aproximadamente o mesmo tamanho são candidatas convenientes para alinhamento global (Figura 4).

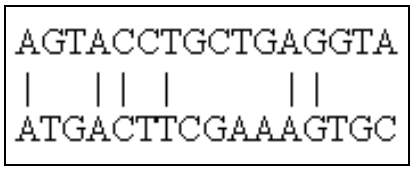

Figura 4: Exemplo de alinhamento global entre duas seqüências

Em alinhamentos locais, segmentos das seqüências com as mais altas densidades de coincidências são alinhados, gerando uma ou mais ilhas de sub-alinhamentos nas seqüências. Alinhamentos locais são mais apropriados para alinhar seqüências que são similares ao longo de um trecho de suas seqüências e mas dissimilares em outros trechos, ou então, seqüências que diferem muito no tamanho ou que compartilham uma região ou domínio conservado (Figura 5).

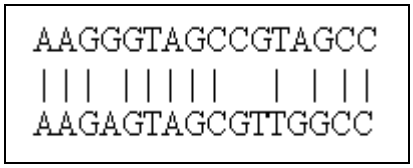

Figura 5: Exemplo de alinhamento local entre duas seqüências

1.7 Medidas de distâncias entre sequências genéticas

A distância genética é uma medida da diferença de material genético entre diferentes espécies ou indivíduos da mesma espécie ou não. Uma das formas de conhecer a relação que existe entre determinadas espécies é através do cálculo da distância entre sequências de DNA. Ao comparar o percentual da diferença entre genes ou sequências de DNA de função desconhecida de diferentes espécies, um valor pode ser obtido, tal medida é a medida da distância genética. Dependendo da diferença, a distância genética pode ser usada como uma ferramenta para construção de dendogramas mostrando a árvore filogenética das espécies em estudo. A utilização da distância genética é uma técnica de grande importância nos programas de melhoramento genético, pois fornece informações úteis na caracterização, conservação e utilização dos recursos genéticos disponíveis.

A análise de divergência genética entre qualquer espécie pode ser dividida de forma simplificada nas seguintes etapas:

a) escolha dos genótipos a serem analisados;
b) obtenção e sistematização dos dados;
c) definição da medida de similaridade ou dissimilaridade a ser estimada;
d) escolha do método de agrupamento
e) interpretação dos resultados a partir da definição física e /ou biológicas.

A seguir, serão abordados dois tipos de medidas de distância entre sequencias de DNA, que são bastante utilizadas: Jukes-Cantor 69 e Kimura 80.

Distância de Jukes-Cantor - JC69

Dentre os modelos desenvolvidos com o objetivo de determinar a distância evolucionária entre sequências de DNA, o primeiro e mais simples foi proposto em 1969 por Thomas H. Jukes e Charles R. Cantor. Neste modelo, assume-se que as transições (trocas entre bases do tipo purinas: adenina e guanina) ocorrem com a mesma probabilidade que as demais substituições de

nucleotídeos (transversões). Através deste modelo, obtém-se uma fórmula que determina o número *d* estimado de substituições de nucleotídeos por sítio ou distância evolucionária:

$$d = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right), \qquad (1)$$

Em que *p* = proporção de diferenças de nucleotídeos observada entre as duas sequências (número de substituições observadas dividido pelo número total de sítios de nucleotídeos comparados).

Esta fórmula leva em consideração as substituições que não são evidentes (não observadas, isto é, substituições múltiplas), convertendo a proporção p em uma distância evolucionária *d*.

Distância de Kimura 80

A distância de Kimura (1980) é baseada em seu modelo de dois parâmetros. É composto por algumas fórmulas simples que permite estimar a distância em termos do número de substituições de nucleotídeos (e, também, as taxas de evolução, quando as divergências são conhecidas). Ao comparar um par de sequências de nucleotídeos, distinguimos as diferenças entre as sequências das espécies em estudo.

$$K80 = -\frac{1}{2}\ln\left(1-2\cdot p-2\cdot q\right)-\frac{1}{4}\ln\left(1-2\cdot q\right) \qquad (2)$$

Em que *p* = proporção de sítios[1] na sequência de DNA;
q = proporção de sítios que mostram diferenças transversais

[1] Sito é definido como a localização física dos nucleotídeos na sequência de DNA.

As duas distâncias mostradas anteriormente serão consideradas para as análises posteriores, pois as mesmas são indicadas quando tem-se sequências de diferentes tamanhos á serem analisadas.

1.9 Análises descritivas

A análise descritiva de uma sequência genética ou análise da composição é uma forma de representar, através de tabelas as freqüências das bases de nucleotídeos em cada organismo que esta sendo comparado. A sequência consenso também é representada através de uma tabela e assim é possível comparar dentre as espécies analisadas, qual apresenta maior semelhança em termos de freqüência com a sequência consenso que como já foi discutido anteriormente é uma característica utilizada para indicar espécies que expressam determinadas características com maior probabilidade que outras.

1.9 Métodos de agrupamentos

Os métodos de agrupamento têm por finalidade separar um grupo original de observações em vários subgrupos, de forma a obter homogeneidade dentro e heterogeneidade entre os subgrupos (MINGOTI, 2005). Dentre estes métodos, os hierárquicos e os de otimização são empregados em grande escala para na área de melhoramento genético.

Nos métodos hierárquicos, os genótipos são agrupados através de um processo que se repete em vários níveis, sendo estabelecido um dendrograma, Uma forma de representar a estrutura de agrupamento com base na distância entre os pares de genótipos é definida por CRUZ & REGAZZI (2001).

Existem várias formas de representar esta estrutura de agrupamento, tais como: o método do vizinho mais próximo, o método do vizinho mais distante, método de agrupamento pareado não ponderado baseado na média aritmética - UPGMA (*unweighted pair-group method using arithmetic averages*), método de Ward, dentre outros. O método UPGMA é a técnica mais simples

para construção de árvores filogenéticas. Ele foi desenvolvido para a construção de árvores que apresentem similaridade nas unidades taxonômicas que se deseja comparar (Graur & Li, 1999).

O método UPGMA apresenta vantagem sobre os demais métodos por considerar médias aritméticas das medidas de dissimilaridade, o que evita caracterizar a dissimilaridade por valores extremos entre os indivíduos considerados (CRUZ; CARNEIRO, 2003). Esse método foi utilizado por ser um dos mais usados na prática e pela facilidade de ser encontrado nos mais diversos programas computacionais.

Existem diversas técnicas estatísticas que são utilizadas com a finalidade de separar determinadas características em uma dada população. A análise de agrupamento foi usada primeiramente por Tyron em 1939 e tem como principal objetivo dividir os elementos de uma amostra ou população, em grupos de forma que os elementos pertencentes a um mesmo grupo sejam similares entre si.

1.10 Bioinformática

A bioinformática é uma área interdisciplinar que envolve a biologia, informática, matemática e estatística. A bioinformática tem como principal objetivo a organização de dados de uma maneira que permita aos pesquisadores acessar as informações existentes e submeter novos dados assim que são produzidos. Entretanto, as informações armazenadas em bancos de dados por si só não são muito úteis se não forem analisadas de forma correta. Portanto, um segundo objetivo da bioinformática consiste em desenvolver ferramentas e recursos que ajudem na análise dos dados.

Os bancos de dados biológicos contêm diversas informações sobre sequências e estruturas de várias espécies. Essas informações são geralmente armazenadas em bancos de dados públicos e vários pesquisadores podem ter acesso. O *GenBank* é um banco de dados mundial e nele contém seqüências de DNA disponíveis publicamente de mais de 140.000 organismos diferentes.

Essas seqüências são obtidas principalmente através de submissões de dados de seqüência de laboratórios individuais e submissões em lotes de projetos de seqüenciamento de grande escala. O *GenBank* é mantido pelo "*National Center for Biotechnology Information*" (NCBI). O último registro disponível, mostra que até o ano de 2008 existiam mais de 99 bilhões de pares de bases (99.116.431.942) e aproximadamente 99 milhões de sequências (98.868.465) presentes no site do NCB (http://www.ncbi.nlm.nih.gov).

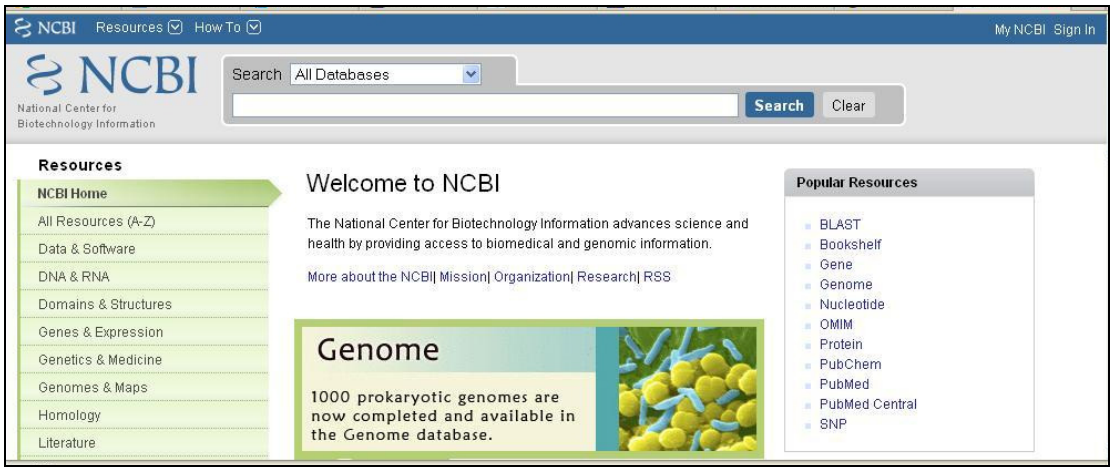

Figura 6: *Home Page* do NCBI - http://www.ncbi.nlm.nih.gov

O BLAST - "*Basic Local Alignment Search Tool*" é a ferramenta computacional de maior referência para a busca de similaridade em bancos de dados de seqüências genéticas. FASTA e BLAST também são algoritmos de alinhamento. Pelo fato de usarem heurísticas (quantificação de proximidade a um determinado objetivo), esses algoritmos, principalmente o BLAST, são os mais usados atualmente (ALTSCHUL et al., 1990).

O DAMBE (*Data Analysis in Molecular Biology and Evolution*) é um software para análise de dados de biologia molecular e tem como objetivo a manipulação, alinhamento e a aplicação de análises estatísticas em sequências de DNA (Xia, 2002).

O Clustal W é um algoritmos muito utilizado para alinhamentos múltiplos de sequências de DNA (Thompson, 1994). O algoritmo consiste em três etapas:

- ✓ Alinhamentos par-a-par realizado entre todas as sequências no grupo em estudo. Pontuações são utilizadas para se construir uma matriz de distâncias. Ao calcular essa matriz, o programa leva em consideração a divergência entre as sequências.

- ✓ O alinhamento progressivo das sequências é feito, seguindo a ordem dos ramos na árvore-guia. As sequências são alinhadas das extremidades até a raiz. Este alinhamento é feito de acordo com as relações filogenéticas encontradas na árvore filogenética(Figura 7).

- ✓ Uma árvore filogenética é construída a partir da matriz de distâncias utilizando o método de neighbour-joining que é o mais popular na construção de árvores a partir de medidas de distâncias genéticas (SAITOU & NEI, 1987). Essa árvore tem ramos de diferentes tamanhos (Figura 8).

```
      1620      1630      1640      1650      1660      1670      1680      1690
      ----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
Bos   AGAACGCGCTGGAGTCATACGCCTTCAACATGAAGAGCGCCGTGGAGGATGAGGGGCTGAAGGGCAAGATCAGCGAGGCC
Capra AGAACGCGCTGGAGTCGTACGCCTTCAACATGAAGAGCGCCGTGGAGGATGAGGGGCTGAAGGGCAAGATCAGCGAGGCG
Ovis  AGAACGCGCTGGAGTCGTACGCCTTCAACATGAAGAGCGCCGTGGAGGATGAGGGGCTGAAGGGCAAGAT----------
      ********** ***************************************************
```

Figura 7: Alinhamento realizado entre três espécies no Clustal W

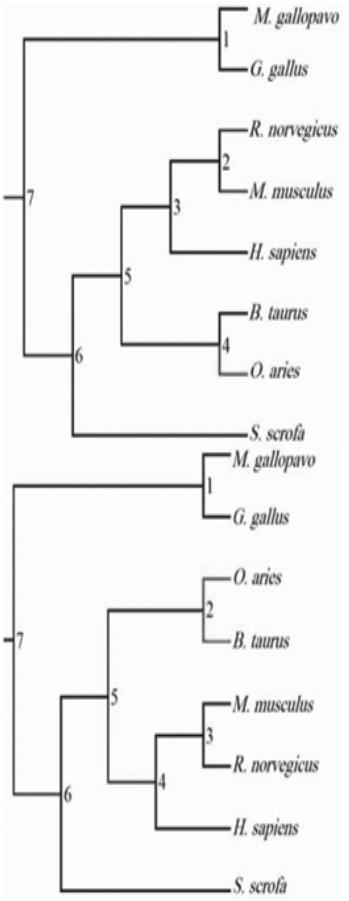

Figura 8: Ilustração de uma árvore filogenética construída no Clustal W

1.11 Objetivos da dissertação

A partir de uma necessidade de entender os fundamentos de genética, da biologia molecular e de algumas análises computacionais de sequências genéticas e, com eles, aplicar métodos estatísticos e esquemas de análise para buscar as interpretações que possam responder por vários fenômenos presentes no nosso cotidiano. Esta dissertação tem por objetivo investigar os polimorfismos de genes HSP70.1 e o NRAMP-1 associados à resistência à mastite em três espécies de mamíferos (bovinos (*Bos taurus*), Ovinos (*Ovis aries*) e Caprinos (*Capra hircus*) e os fatores genéticos pertinentes por meio de alinhamentos e métodos de distâncias em sequências genéticas. Também é de interesse realizarr uma aplicação do método de Cadeias de Markov Ocultas,

com o objetivo de mensurar as probabilidades de transição entre bases de nucleotídeos numa dada sequência genética e consecutivamente encontrar regiões homogêneas em tais sequências (sequências mais prováveis). Por fim, com uso do algoritmo de *Baum-Welch* buscaremos pelos melhores parâmetros otimizando a probabilidade de observações de uma dada sequência.

Referências Bibliográficas

ADAMOWICZ, T, PERS, E, LECHNIAK, D 2005. A new SNP in the 3-UTR of the Hsp70-1 gene in Bos taurus and Bos indicus. Biochemical Genetics 43, 623–627.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS. E. W.; LIPMAN, D. J. Basic Local Alignment Search Tool. J. Mol. Biol. (1990). V. 215. P. 403-415

BRANT, M.C.; FIGUEIREDO, J.B. Prevalência da mastite subclínica e perdas de produção em vacas leiteiras. Arquivo Brasileiro de Medicina Veterinária e Zootecnia, v. 46, n. 6, p. 595-606, 1994.

BRITO, R. M., CAIXEIRO, A. P. A., POMPOLO, S. G., & AZEVEDO, G. G. (2003) Cytogenetic data of Partamona peckolti (Hymenoptera, Apidae, Meliponini) byy C banding and fluorochrome staining with DA/CMA3, And Da/DAPI. Gen. Mol. Biol. 26: 53-57.

BRITO, Rogério Theodoro. Alinhamentos de Múltiplas Sequências. 2001. 206 f. Dissertação (Mestrado em Ciências da computação) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2001.

CRUZ, C. D.; CARNEIRO, P. C. S.; Modelos Biométricos aplicados ao melhoramento genético- Volume 2. Viçosa: UFV, 2003. 585p.

CRUZ, C. D.; REGAZZI, A. J. Modelos biométricos aplicados ao melhoramento genético. Viçosa, MG: Ed. UFV, 2ª ed. rev., v.1, 390 p, 2001.

DETILEUX, J; PAAPE, MJ; MEHRZAD, J; ZHAO X and BURVENICH, C (2002). Defense of the bovine mammary gland by polymorphonuclear neutrophil leukocytes. J. Mammary Gland Biology and Neoplasia. 7: 109-121.

FONSECA, L. F. L.; SANTOS, M. V. Qualidade do Leite e Controle de Mastite. São Paulo: Lemos Editorial, 2000.

GIBSON, G.; MUSE, S. V. A Primer of Genome Science. 2nd ed. [S.l.]: Sinauer, 2004.

GRAUR, D. and Li, W.-H. 1999. Fundamentals of Molecular Evolution. Second Edition, Sinauer Associates.

GRIFFIN, K. B.; MICHAEL, J. J.; FOX, L. K.; GASKINS, C. T.; JIANG, Z. Fine mapping of the bovine chromosome 22q24 region that harbours antimicrobial genes and a QTL for samtic cell scores. Animal Genetics, v. 36, p. 435-462. 2005.

GRIFFITHS, A. J. F. et al. An Introduction to Genetic Analysis. New York: W. H. Freeman, 2000.

HANSEN P. J. 2004 Physiological and cellular adaptations of zebu cattle to thermal stress. Anim. Reprod. Sci. 82–83, 349–360.

JOO, Y. S.; MOON, J. S.; FOX, L. K.; SUH, G. H.; KWON, N. H.; KIM, S. H.; PARK, Y. H. Comparison of natural resistance-associated machrophage protein (NRAMP)1 espression between coes with high and low milk somatic cells counts. Asian-Australasian Journal of Animal Science, v. 16, n. 12, p.1830-1836. 2003.

LEBLANC, S. J.; LISSEMORE, K. D.; KELTON, D. F.; DUFFIELD, T. F.; LESLIE, K. E. Major advances in disease prevention in dairy cattle. Journal of Animal Science, Champaign, v. 89, p. 1267-1279, 2006.

LEYVA-BACA, I.; PIGHETTI, G.; KARROW N. A. Genotype-specific IL8RA gene expression in bovine neutrophils in response to Escherichia coli lipopolysaccharide challenge. Animal Genetics, v.39, p. 298-300. 2008.

MACHADO, P.F.; PEREIRA, A.R.; SILVA, L.F.P. Células somáticas no leite em rebanhos brasileiros. Scientia Agrícola, São Paulo, v.57, n.2, p.359-361,2000.

MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2005.

NONNECKE, B.J. e HARP, J.A. Function and regulation of Lymphocyte-Mediated responses: Relevance to bovine mastitis. Journal of Dairy Science, v.72, n.5, p.1313-1327, 1988.

RAINARD, P. e RIOLLET, C. Innate immunity of the bovine mammary gland. Veterinary Research, v. 37, p. 369-400, 2006.

SAITOU, N. and NEI, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4, 406–425.

SHARIF, S.; MALLARD, B. A.; WILKIE, B, N. Charazterization of naturally processed and presented peptides associated with bovine major histocompatibility complex (BoLA) class II DR molecules. Animal Genetics, v. 34, p. 116-123. 2003

STEAR, M. J.; POKORNY, T. S.; ECHTERNLAMP, S. E.; LUNSTRA, D. D. The influence of the BoLA- A locus on reproductive traits in cattle. Journal of Immunogenetics, v. 16, p. 77-88. 1989.

THOMPSON, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice. Nucl. Acids Res. 22:4673-4680.

TONHATI, H., ALBUQUERQUE, L.G., OLIVEIRA, J.F.S. et al. Melhoramento genético em bubalinos. Programa Vale do Ribeira, SP.

VIDAL, S.M., MALO, D., VOGAN, K., SKAMENE, E., GROS, P. (1993). Natural resistance to infection with intracellular parasites: isolation of a candidate for Bcg. Cell 73,469-485.

XIA, X., 2002, Data Analysis in Molecular Biology and Evolution. New York. Kluwer Academic Publishers, 284p.

WATSON, J. et al. Recombinant DNA. 2nd ed. New York: Scientific American Books, 1992.

WATSON, James D. ; BERRY, Andrew. DNA: O segredo da Vida. (Tradução de Carlos Afonso Malferrari) São Paulo : Companhia das Letras, 2005.

WATTERMAN, M. S. Introduction to Computational Biology: Maps, Sequences and Genomes, 1995.

CAPÍTULO 2

Identificação dos polimorfismos e agrupamento de sequências dos genes HSP70.1 e NRAMP-1 em três espécies candidatas ao acometimento de Mastite[2]

Rita de Cássia de Lima Idalino[3]

Orientador: Prof. Dr. Kleber Régis Santoro[2]

Resumo: O estudo da genética molecular permite identificar genes com responsabilidade de conferir resistência ou propensão a algumas doenças como a Mastite que é uma inflamação da glândula mamária que acomete os mamíferos. Os genes HSP70.1 e o NRAMP-1 estão envolvidos no processo molecular de estresse à Mastite. O objetivo deste estudo foi realizar uma análise dos polimorfismo em sequências de DNA de três espécies de animais (*Bos Taurus, Ovis Aires e Capra Hircus*) afim de conhecer o grau de similaridade entre as espécies em estudo. Com uso de técnicas de alinhamento de sequências genéticas foi possível comparar as frequências de polimorfismos bem como através do cálculo das distâncias genéticas JC69 e KM80 foi possível quantificar o distanciamento entre as espécies.

Palavras-Chave: alinhamento; leite; sequências genéticas.

Abstract: The study of molecular genetics can identify genes responsible for conferring resistance or propensity for certain diseases such as mastitis which is an inflammation of the mammary gland that occurs in mammals. The genes HSP70.1 and Nramp-1 are involved in molecular stress to mastitis. The aim of this study was an analysis of polymorphism in DNA sequences of three species of animals (Bos Taurus, Ovies aires and Capra hircus) in order to know the degree of similarity between these species. With use of techniques of genetic sequences alignment was possible to compare the frequencies of polymorphisms, and by calculating the genetic distances KM80 and JC69 was possible to quantify the distance between the species.

Keywords: alignment; milk; gene sequences.

2 Projeto financiado pelo CNPq (505912/2008-2)

3 Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - UFRPE, CEP: 52.171-900, Recife, Pernambuco, Brasil, E-mail para contato: ritalimex@yahoo.com.br

1. Introdução

A pesquisa genômica é de grande interesse para diversas áreas que compõem uma população. Por isso, o entendimento de como os genes influenciam na aparição de determinadas características é de grande importância para a criação de métodos de diagnósticos e drogas apropriadas. A maioria dos genes apresenta uma grande frequência de variações alélicas, conhecidas como polimorfismos. Estas variações podem ser a chave para a resistência que algumas espécies de animais apresentam a certas doenças (BARBOSA, 2006).

Na produção animal, dentre os problemas de sanidade, as doenças infecto-contagiosas são as que mais se destacam. Uma das opções mais promissora para a redução dos problemas causados pelas doenças infecto-contagiosas, além dos cuidados sanitários, é a seleção de animais resistentes. Como ferramenta para a seleção de animais melhores adaptados e mais produtivos pode ser utilizada a estratégia de genes candidatos, e para sua escolha é importante caracterizar a expressão gênica em animais resistentes e susceptíveis às doenças (FONSECA 2008).

A genômica animal é uma realidade presente nos programas de melhoramento e os impactos das aplicações desses métodos podem ser notados em várias áreas da produção animal. O estudo da biologia molecular representa hoje uma das áreas de maior potencial para a realização de pesquisas, considerando-se não apenas sua grande relevância clínica e epidemiológica, mas também pela possibilidade de aplicação de ferramentas estatísticas (COLLINGS et al, 1998).

No Brasil, a produção de leite é uma atividade cada vez mais competitiva. Portanto é importante quantificar e qualificar os fatores que podem influenciar nesta produção, buscando maior ganho, na tentativa de suprir a demanda nacional. A seleção de indivíduos melhorados para acasalamento e multiplicação repercute nos índices de produtividade e produção (HOLT & CARVALHO, 2007)

.

O leite é considerado o mais nobre dos alimentos, por sua composição rica em proteína, gordura, carboidratos, sais minerais e vitaminas, proporciona nutrientes e proteção imunológica para o recém-nascido. O Brasil é o quarto maior produtor de leite do mundo com cerca de 29 mil toneladas no ano de 2009, segundo dados do Departamento de Agricultura dos Estados Unidos (***U**nited **S**tates **D**epartment of **A**griculture* - USDA).

Um dos fatores de maior influência sobre a produção leiteira, tanto sob o caráter de sanidade animal quanto de resultados econômicos, são as doenças que acometem os animais destinados a produção de leite, seus constituintes e a sua qualidade. Dentre eles a mastite é a que possui a maior importância para os rebanhos, pois é tida como a doença que produz os maiores prejuízos econômicos (Fonseca e Santos, 2000).

A Mastite é uma inflamação na glândula mamária causada pelos mais diversos agentes ambientais, no caso da mastite clínica, diversas bactérias. É considerada uma das doenças mais comuns na pecuária leiteira e que acarreta diversos prejuízos aos criadores, pois ela diminui a produção de leite, compromete a qualidade deste, desvaloriza comercialmente o animal e pode até causar a morte dele por infecção irreversível (RADOSTITS et al,. 2000).

A mastite provoca alterações nos três principais componentes do leite, gordura, proteína e carboidratos. Enzimas e minerais também são afetados. O aumento da contagem de células somáticas (CCS) e as mudanças na composição do leite estão diretamente relacionadas com a superfície do tecido mamário atingido pela reação inflamatória (SCHÄELLIBAUM, 2000).

A Mastite é uma doença infecciosa que tem o maior impacto negativo na pecuária leiteira. Segundo dados da Pesquisa de Orçamento Familiar, realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), o consumo per capita de leite de vaca no Brasil é de 27,05 litros por ano. De acordo com informações da Organização das Nações Unidas para Agricultura e Alimentação (FAO), a produção de leite para diferentes espécies, mostra o leite

de vaca como a maior produção mundial. Este é o leite mais utilizado na produção de laticínios devido às propriedades que possui, às quantidades que se obtém, agradável sabor, fácil digestão.

O controle preventivo de doenças como a Mastite proporciona um aumento da produtividade, maior qualidade dos produtos e redução os custos para o criador. O melhoramento genético animal auxilia no controle e até na eliminação de algumas doenças quando altera geneticamente populações de animais aumentando a frequência de genes (ou alelos) e de genótipos desejáveis refletindo favoravelmente no mérito fenotípico médio de características destas populações que sejam importantes economicamente (KOIYAMA, 2009).

Diferenças genéticas que são comuns entre organismos da mesma espécie são chamadas de polimorfismo genético. A identificação bem como a análise dos polimorfismos em determinados genes pode ter um impacto positivo no melhoramento animal, principalmente em características de mensuração tardia e dificultosa para o pecuarista (LEWIN, 2001)

.

O objetivo deste estudo consiste em analisar as sequências completas dos genes HSP70.1 (*Heat Shock Protein*) e o NRAMP-1 (*Natural Resistance – Associated Macrhophage Protein*), sendo que estes possuem relação com a expressão da mastite. Especificamente, essa análise será realizada em sequências genéticas de vacas, ovelhas e cabras e também será feita uma análise de agrupamento para mostrar a proximidade existente entre as espécies em estudo através de métodos de distância genética e agrupamento. Tanto o método de agrupamento bem como as distâncias utilizadas são os mesmos já mencionados no capítulo introdutório desta dissertação.

2 Materiais e Métodos

A resposta de resistência é uma característica complexa e vários genes podem estar envolvidos na determinação desta função. Assim, os genes envolvidos na resposta imune têm sido apontados como fortes candidatos para o fenótipo de resistência, sendo que os principais genes envolvidos com a questão da Mastite, analisados neste trabalho foram o HSP70.1 (*Heat Shock Protein*) e o NRAMP-1 (*Natural Resistance – Associated Macrhophage Protein*).

Foram analisadas seis sequências de DNA das espécias *Bos taurus, Ovis aries* e *Capra hircus* relacionadas aos genes em estudo. Estas sequências foram retiradas do *Gen Bank* no site do NCBI (*Nactional Center for Biotechnology Information* - www.ncbi.nlm.nih.gov). Em media cada sequência possue 2000 pares de bases. Os alinhamentos para análise de composição das sequências bem como a análise de agrupamento foram realizados utilizando a ferramenta ClustalW do software DAMBE (*Data Analysis in Molecular Biology and Evolution*).

Com o alinhamento das sequências é possível observar as regiões onde os nucleotídeos se repetem, e desta forma é gerada uma sequência padrão para comparação e análise dos polimorfismos, denominada sequência consenso. Após a construção e análise dos alinhamentos foi feito um estudo descritivo com a composição de cada sequência e com a similaridade entre as raças mencionadas e com a sequência consenso, esta obtida a partir dos pontos de similaridade no alinhamento.

3 Resultados e Discussões

A partir dos alinhamentos das seqüências dos genes HSP70.1 e NRAMP-1 das espécies *Bos taurus, Ovis aries e Capra hircus*, foi possível observar a presença de pontos polimórficos com troca de nucleotídeos, sítios de deleções, sítios de adição de nucleotídeos e similaridades.

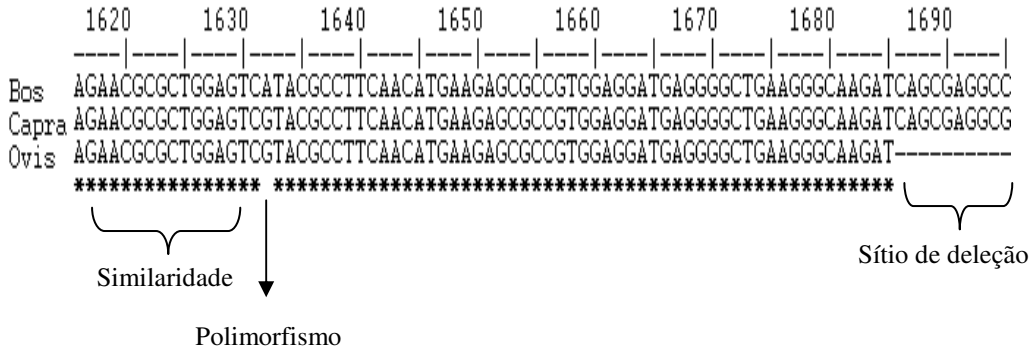

Figura 1: Alinhamento das sequências genéticas da posição 1620 até a posição 1695 do gene HSP70-1

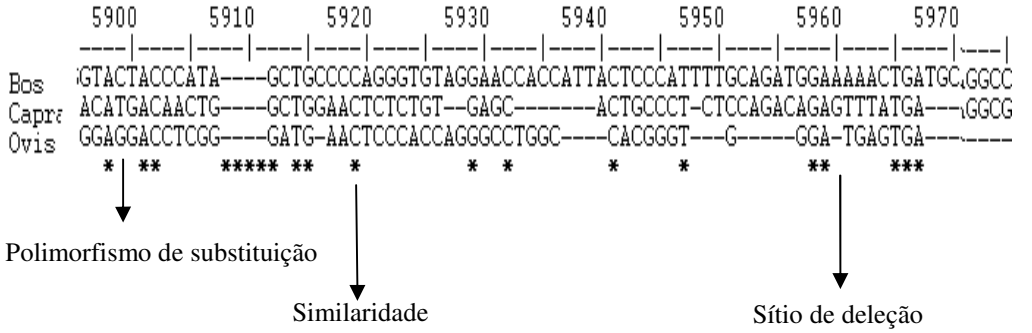

Figura 2: Alinhamento das sequências genéticas da posição 5895 até a posição 5975 do gene NRAMP.1

Uma vez realizado o alinhamento entre as sequências é possível observar a relação de proximidade entre as espécies em estudo através de uma análise gráfica. O dendograma é um meio prático de sumarizar um padrão de agrupamento. Este começa com todos os indivíduos separados, fundindo-se progressivamente em pares até chegar a uma única raiz. A ordem dos indivíduos mostrada no dendograma e a ordem na qual os grupos entram no agrupamento. As figuras abaixo ilustram os dendogramas formados a partir dos alinhamentos mostrados anteriormente entre as três espécies considerando os dois genes.

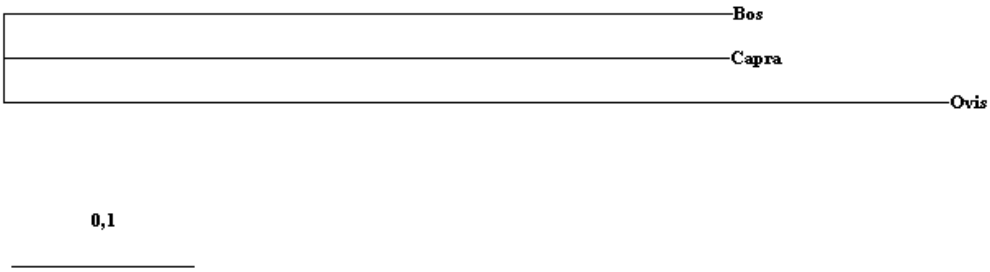

Figura 3: Estrutura filogenética das distâncias entre as espécies relacionadas ao gene HSP70.1

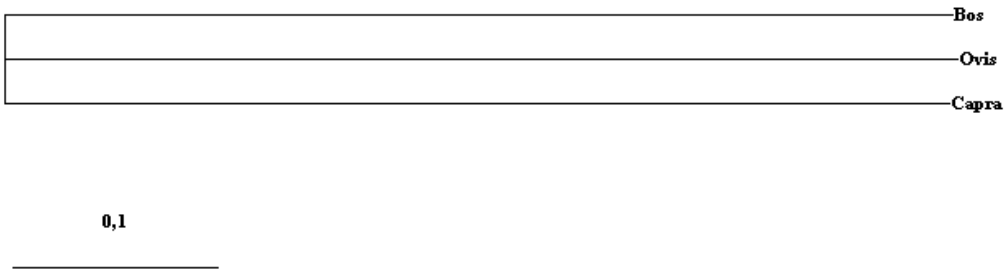

Figura 4: Estrutura filogenética das distâncias entre as espécies relacionadas ao gene NRAMP-1

Uma árvore filogenética é uma descrição, através de diagramas, das relações entre entidades biológicas interligadas por ancestrais comuns. A necessidade e o instinto humano em classificar os objetos em seus meios levaram-no a desenvolver ferramentas cada vez mais úteis e práticas. O surgimento da sistemática biológica faz uso intenso de árvores filogenéticas.

Observando o dendograma relacionado às espécies do gene HSP70.1, observa-se que a espécie *Bos Taurus* apresentou maior distanciamento com relação as demais espécies relacionadas ao mesmo gene. Observando a figura 2, em que esta demonstrado a semelhança entre as três espécies consideradas para o gene NRAMP-1.

Os locais onde os nucleotídeos se repetiam entre si gerou uma sequência padrão, denominada sequência consenso e será com base nesta que será fornecida uma indicação do potencial que cada espécie pode ter de resistir ao desenvolvimento a Mastite, uma vez que os genes em estudo são classificados como fatores de resistência ao estresse associados a esta doença.

O alinhamento desta sequência com a sequência de cada em análise forneceu o número de similaridades e de polimorfismo existentes em cada estrutura bem como a média de cada informação. A análise descritiva da composição das sequências dos genes estudados e das sequência consenso foram realizadas no *software* DAMBE (Xia, 2002).

Nas tabelas 1 e 2 estão apresentas as distribuições de frequência das similaridades e de polimorfismos em cada espécie por bases de nucleotídeos e as respectivas porcentagens com relação aos dois genes, HSP70.1 e o NRAMP-1. As tabelas 3 e 4 mostram a composição em termos de freqüência de bases e probabilidades. Estas composições são importantes para que no momento da comparação, seja possível identificar quais as espécies são mais semelhantes à sequência consenso em termos de freqüência.

Com base nas tabelas 5 e 6, relacionadas ao consenso entre as espécies dos genes é possível observar que considerando o gene HSP70.1, a espécie que apresentou maior similaridade foi a *Capra hircus,* ou seja, 46,47%. Com relação a análise de polimorfismo, observa-se que a espécie *Ovies aries* apresentou maior diferença entre as outras espécies, 64%. Considerando a composição do gene NRAMP-1, a espécie que apresentou maior similaridade foi a *Bos taurus* com 84,15%. Já para a análise dos polimorfismos, a espécie *Capra hircus* apresentou um percentual de 51,92%.

Analisando a composição de similaridade gerada a partir da sequência consenso, ou seja, a sequência padrão entre as espécies em análise, tabelas 5 e 6, observa-se que para o gene HSP70.1, a espécie que apresentou maior percentual de similaridade ao consenso é a *Capra hircus* com 84,48%. Quando parte-se para o gene NRAMP-1, a espécie mais similar, com 93,71%, é a espécie *Bos taurus.* Sendo

assim, para as sequências consideradas, estas são as espécies que apresentam maior resistência ao desenvolvimento da mastite quando comparada as demais espécies que foram considerados. Considerando os polimorfismos entre os genes, a espécie *Ovis Aires,* foi a que se mostrou com maior percentual de diferença com relação as demais espécies consideradas no gene HSP70.1, 74,01%. Já para o gene NRAMP-1, o maior percentual de polimorfismo esta associado à espécie *Capra hircus*, 96,09%.

Tabela 1: Frequência da similaridade e dos polimorfismos do gene HSP70.1

Raça	Similaridade	Polimorfismo	Percentual de Similaridade	Percentual de Polimorfismo
Bos taurus	1619	649	0,3925*	0,2334
Ovis aries	589	1780	0,1428	0,6400
Capra hircus	1917	352	0,4647*	0,1266
TOTAL	**4125**	**2781**	**1,0000**	**1,0000**
MÉDIA	**1375**	**927**	**-**	**-**

Tabela 2: Frequência da similaridade e dos polimorfismos do gene NRAMP-1

Raça	Similaridade	Polimorfismo	Percentual de Similaridade	Percentual de Polimorfismo
Bos taurus	6875	461	0,8415	0,0333
Ovis aries	1008	6328	0,1234*	0,4573
Capra hircus	287	7049	0,0352*	0,5192
TOTAL	**8170**	**13838**	**1,0000**	**1,0000**
MÉDIA	**2723**	**4613**	**-**	**-**

*Espécies com maior percentual de similaridade.

Tabela 3: Frequências e probabilidades de cada base (A, C, T, G) em cada raça no gene HSP70.1

Sequência	A	C	T	G	Soma	P(A)	P(C)	P(T)	P(G)
Bos taurus	374	433	310	521	1638	0,2283	0,2643	0,3181	0,1893
Ovis aries	136	169	73	211	589	0,2309	0,2869	0,1239	0,3582
Capra hircus	434	562	268	662	1926	0,2253	0,2918	0,3437	0,1391

Tabela 4: Frequências e probabilidades de cada base (A, C, T, G) em cada raça no gene NRAMP-1

Sequência	A	C	T	G	Soma	P(A)	P(C)	P(T)	P(G)
Bos taurus	1807	1793	1632	2059	7291	0,2478	0,2459	0,2238	0,2824
Ovis aries	378	367	342	407	1494	0,2530	0,2456	0,2289	0,2724
Capra hircus	85	80	132	93	390	0,2179	0,2051	0,3385	0,3385

Tabela 5: Frequência e percentual da similaridade e dos polimorfismos na sequência consenso relacionada ao gene HSP70.1

Raça	Similaridade	Polimorfismo	Percentual de Similaridade	Percentual de Polimorfismo
Bos taurus	1620	650	0,7136*	0,2863
Ovis aries	590	1680	0,2599	0,7401
Capra hircus	1917	352	0,8448*	0,1551
TOTAL	**4127**	**2682**	**-**	**-**
MÉDIA	**1375**	**849**	**-**	**-**

Tabela 6: Frequência e percentual da similaridade e dos polimorfismos na sequência consenso relacionada ao gene NRAMP-1

Raça	Similaridade	Polimorfismo	Percentual de Similaridade	Percentual de Polimorfismo
Bos taurus	6875	461	0,9371	0,0629
Ovis aries	1008	6328	0,1458*	0,8542
Capra hircus	287	7050	0,0391*	0,9609
TOTAL	**8170**	**13839**	**-**	**-**
MÉDIA	**2723,33**	**4613**	**-**	**-**

*Espécies com maior percentual de similaridade.

Uma forma de mensurar esse distanciamento entre as espécies é através do calculo de algumas distâncias genéticas. Foram consideradas duas distâncias para observar a separação entre as espécies. A distância JC69 (Jukes e Cantor 1969) considera a taxa de substituição de nucleotídeos é a mesma para todos os pares dos quatro nucleotídeos A, T, C e G. Ela pressupõe uma igualdade de taxas de substituição entre os locais das sequências em consideração. Já a distância K80 (Kimura 1980) é utilizada, com exclusão de pares. Tais distâncias foram consideradas, pois as mesmas devem ser usadas quando as sequências sob

análise não possuem o mesmo dimensionamento. O método de agrupamento utilizado foi o UPGMA.

A distância genética JC69 e K80, encontradas a partir do programa DAMBE, das três espécies relacionadas ao gene HSP70.1 forneceu a seguintes matrizes de distâncias:

Tabela 7: Matriz de distância genética – JC69 – HSP70.1

Espécie	*Bos taurus*	*Ovis Aires*	*Capra Hircus*
Bos taurus	0	0,02194	0,01544
Ovis Aires	0,02194	0	0,00170
Capra Hircus	0,01544	0,00170	0

Tabela 8: Matriz de distância genética – K80 – HSP70.1

Espécie	*Bos taurus*	*Ovis Aires*	*Capra Hircus*
Bos taurus	0	0,02195	0,01545
Ovis Aires	0,02195	0	0,00170
Capra Hircus	0,01545	0,00170	0

Os dendogramas construídos com os resultados das matrizes de distância dos modelos JC69 K80 apresentam de forma mais clara a relação da intensidade do distanciamento entre as espécies.

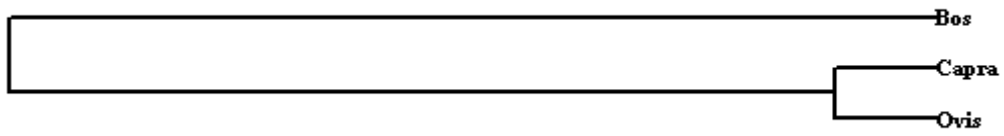

Figura 5: Estrutura filogenética das distâncias entre as espécies relacionadas ao gene HSP70.1 com base na distância JC69

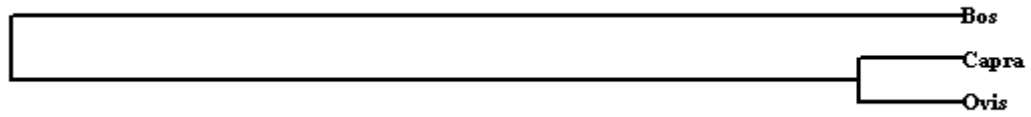

Figura 6: Estrutura filogenética das distâncias entre as espécies relacionadas ao gene HSP70.1 com base na distância K80

Observando as matrizes de distâncias bem como os dendogramas gerados é possível perceber que os resultados obtidos são muito semelhantes entre as distâncias consideradas. A espécie *Bos taurus* é a que apresenta um maior distanciamento com relação às demais espécies (Figura 5 e 6).

Considerando as sequências genéticas relacionadas ao gene NRAMP-1, foi realizada uma análise semelhante à aplicação das distâncias anteriores para observar o comportamento das espécies em relação às distâncias genéticas JC69 e K80. As matrizes geradas são mostrados a seguir:

Tabela 9: Matriz de distância genética – JC69 - NRAMP-1

Espécie	*Bos taurus*	*Ovis Aires*	*Capra Hircus*
Bos taurus	0	1,21750	0,77064*
Ovis Aires	1,21750	0	1,02607
Capra Hircus	0,77064*	1,02607	0

Tabela 10: Matriz de distância genética – JC69 - NRAMP-1

Espécie	*Bos taurus*	*Ovis Aires*	*Capra Hircus*
Bos taurus	0	1,21822	0,77101*
Ovis Aires	1,21822	0	1,03515
Capra Hircus	0,77101*	1,03515	0

*Espécies com maior distanciamento

Os dendogramas construídos com os resultados das matrizes de distância dos modelos JC69 K80 apresentam de forma mais clara a relação da intensidade do distanciamento entre as espécies.

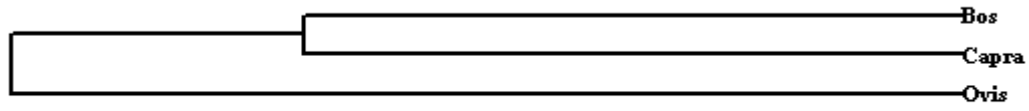

Figura 7: Estrutura filogenética das distâncias entre as espécies relacionadas ao gene NRAMP-1 com base na distância JC69

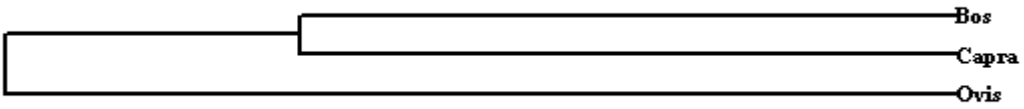

Figura 8: Estrutura filogenética das distâncias entre as espécies relacionadas ao gene NRAMP-1 com base na distância K80

As matrizes de distâncias bem como os dendogramas gerados mostram resultados semelhantes entre as distâncias consideradas. Considerando o gene NRAMP-1, a espécie *Ovis Aries* apresentou um maior distanciamento em relação as demais espécies, considerando as duas distâncias. (Figuras 7 e 8).

4 Conclusões

Os resultados apresentados neste capítulo mostram o procedimento descritivo de uma sequência genética, desde a composição até o uso de técnicas que mostram a formação de grupos e suas interrelações. Descrever a composição, em termos de freqüência é uma forma de conhecer do ponto de vista genético, a formação de um determinado organismo. É com base na composição filogenética que é possível observar a relação que existe entre organismos diferentes.

Com uso das técnicas de distância genéticas e do método de agrupamento foi possível quantificara relação de proximidade entre os organismos em análise. Uma vez identificado o percentual de proximidade, com base nas distâncias e o método de agrupamento utilizado, foi possível identificar entre os grupos as relações genéticas entre os mesmos.

Os métodos de distâncias são empregados com o intuito de mostrar possíveis divergências que existem entre as espécies, e isso é importante para entender o mecanismo da expressão gênica entre diversos organismos.

Entre as duas distâncias utilizadas, JC69 e K80 não percebeu-se distinções entre as mesmas, fato que é aceitável devido a formação como tais distâncias agrupam os elementos e também os pressupostos que cada uma possui.

Estudos incluindo outros genes envolvidos com a resistência da mastite, como também um maior número de sequências e espécies que são susceptíveis ao acometimento da mastite devem ser iniciadas, não apenas para avaliação do potencial dos genes aqui estudados, de modo que se adquira um maior entendimento sobre o mecanismo fisiopatológico da Mastite.

Referências Bibliográficas

ABLES, G.P., NISHIBORI, M.; KANEMAKI, M.; WATANABE, T. Sequence analysis of the NRAMP1 genes from different bovine and buffalo breeds. J. Vet. Med. Sci., v.64, n.11, p.1081-83, 2002.

BARBALHO, T.C.F.; MOTA, R. A. Isolamento de agentes bacterianos envolvidos em Mastite subclínica bovina no estado de Pernambuco. Revista Brasileira de Produção Animal, v. 2, n. 2, p. 31-36. 2001.

BARBOSA NETO J.F.; SORRELLS M.E.; CISAR, G. Prediction of heterosis in wheat using coefficient of parentage and RFLP – based estimates of genetic relationship. Genome, Ottawa, v.39, p. 1142-1149, 1996.

BAXEVANIS, A. D., OUELLETTE, B. F. F.,2001, Bioinformatics - A Practical Guide to the Analysis of Genes and Proteins. New York. A John Wiley & Sons, Inc., Publication, 489p. Belo Horizonte: Editora UFMG, 2005, 295p.

Benjamin Lewin. Genes VII. Editora ARTMED. São Paulo, SP. 2001. Pág. 39-42.

BRAZ, A. C. V. S; MORAIS F.V. Análise in silico de genes relacionados à autofagia de três isolados do Paracoccidioides brasiliensis: Avaliação do polimorfismo, trabalho de conclusão de curso (Graduação em Biomedicina), UNIVAP, 2009.

CARDOSO, V.L., MONSALVES, F.M., EL FARO, L. et al. Valores econômicos para ocorrência de Mastite clínica e contagem de células somáticas em um sistema intensivo de Produção de Leite. 42ºReunião da Sociedade Brasileira de Zootecnia . Goiânia, Goias. CDROm. 2005.

COLLINGS, F.S.; BROOKS, L.D.; CHAKRAVARTI, A. A DNA polymorphism discovery resource for research on human genetics variation. Genome Research, v.8, p.1229-1231, 1998.

DILLON, W. R. Multivariate analysis. Canadá. John Wiley e Sons, 1984.

FONSECA, I. Perfil da expressão de genes relacionados à resistência à Mastite em bovinos leiteiros. 2008 56f. Dissertação (Mestrado em Zootecnia) – Universidade Federal de Viçosa, Minas Gerais.

HOLT, M. C.; CARVALHO, G. R. Análise espacial da concentração da produção de leite no Brasil e potencialidades geotecnológicas para o setor. Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, Brasil, 21-26 abril 2007, INPE, p. 2729-2736.

IBGE, Pesquisa da Pecuária Municipal 2003, Disponível em: http.\\www.ibge.gov.br/home/estatística/economia/agropecuária/censoagro Acesso em: 10 setembro 2010.

LANGONI, H. Complexidade etiológica na Mastite bovina. In: III Encontro de pesquisadores em Mastite, 1999, Botucatu. Anais Botucatu: FMVZ/UNESP, 1999, 172P. P. 3-14.

LEITE, R.C., BRITO, J.R.F., e FIGUEIREDO, J.B. Alterações da glândula mamária de vacas ratadas intensivamente via mamária, com penicilina em veículo aquoso. Arq. Esc. Vet., UFMG, v.28, p.27-31. 1976.

MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2005.

NCBI, Sequências de DNA de diversos organismos. Disponível em: www.ncbi.nlm.nih.gov. Acesso em: 16 setembro 2010.

Produção de leite mundial. *Department of Agriculture United States* (USDA). Disponível em: http://www.milkpoint.com.br/estatisticas/producao_mundial.htm. Acesso em: 30 setembro 2010.

RADOSTITS O.M, GAY C.C., BLOOD D.C., HINCHCLIFF K.W. Um tratado de doenças dos bovinos, ovinos, suínos, caprinos e eqüinos. Rio de Janeiro: Guanabara Koogan S. A, 2000, 1737p.

KOIYAMA, N. T. G.. Ação do Melhoramento Genético na Contagem de Células Somáticas de Vacas Holandesas. Rev. Form. Inf. Zootec., v.1, n.1, maio 2009.

SCHÄELLIBAUM, M. Efeitos de altas contagens de células somáticas sobre a produção e qualidade de queijos. In: Simpósio Internacional sobre Qualidade do Leite, 2, 2000, Curitiba. Anais... Curitiba: CIETEP/FIEP, 2000. p.21-26.

TONHATI, H. Critérios de seleção para produção total de leite em bubalinos criados no estado de São Paulo, Brasil. Jaboticabal: Universidade Estadual Paulista, 2002. 68p. Tese (Livre Docência) - Universidade Estadual Paulista, 2002.

VIANNI, M.C.E., LÁZARO, N.S. Perfil de suscetibilidade a antimicrobianos em amostras de cocos Gram-positivos, catalase negativos, isolados de Mastite subclínica bubalina. Pesq. Veterin. Bras. n.23, p.47-51. 2003.

VIANNI, M.C.E., LÁZARO, N.S. Suscetibilidade a antimicrobianos em amostras de cocos Gram-positivos, catalase negativos, isolados de Mastite subclínica bubalina. Pesq. Veterin. Bras. n.23, p.47-51. 2003.

XIA, X., 2002, Data Analysis in Molecular Biology and Evolution. New York. Kluwer Academic Publishers, 284p.

XIA, X., 2007, Bioinformatics and the cell - Modern Computacional Approaches in Genomics, Proteomics and Transcriptomics. New York. Springer, 363p.

CAPÍTULO 3

Uso de Cadeias de Markov com Estados Ocultos para Identificação de Regiões Homogêneas em Sequências Genéticas [1]

Rita de Cássia de Lima Idalino[2]

Orientador: Prof. Dr. Kleber Régis Santoro[2]

Resumo: Apesar de milhares seqüências biológicas que são armazenadas em bancos de dados públicos ao redor do mundo, apenas uma pequena parte das informações têm sido sistematicamente explorada. Nesta perspectiva, muitos esforços têm sido realizados através do uso de técnicas que possam extrair informações úteis desses dados com menor custo computacional e maior acurácea. A teoria das Cadeias de Markov com estados ocultos, (HMM - *Hidden Markov Models*), é aplicada ao problema de discriminação de regiões homogêneas em seqüências de DNA, pois através desse modelo matemático é possível tratar de forma probabilística a variação estrutural dos elementos de uma mesma classe biológica. Neste trabalho, objetivou-se identificar regiões conservadas em seqüências associadas a sequências geneticas envolvidos com a resposta imunológica animal frente a agentes causadores da mastite, através de HMMs.

Palavras Chave: Biologia Molecular, HMM, Regiões homogêneas

Abstract: Despite thousands of biological sequences which are stored in public databases around the world, only a small part of these informations have been systematically explored. In this context, many efforts have been made through the use of techniques which can extract profitable informations from such data with more speed and sensitivity. The theory of Markov chains with hidden states, i.e. Hidden Markov Models (HMMs, hereafter), is applied to the problem of discrimination of homogeneous regions in DNA sequences, since this technique permits to treat, in a probabilistic way, the structural variation of elements of the same biological class. In this work, we pursue to identify conserved regions in sequences associated with genes involved in the inflammatory response in animal mastitis, using HMMs.

Keywords: Molecular Biology, HMM, Homogeneous regions

[1] Projeto financiado pelo CNPq (505912/2008-2)

[2] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - UFRPE, CEP: 52.171-900, Recife, Pernambuco, Brasil, E-mail para contato: ritalimex@yahoo.com.br

1. Introdução

De posse da crescente massa de dados biológicos, em particular na biologia molecular, torna-se necessário, o uso técnicas computacionais, bem como modelos teóricos para obtenção de informações pertinentes sobre estes dados. A automatização do processo de seqüenciamento de DNA tornou possível a obtenção de sequências de bases associadas a uma dada molécula de DNA. Houve então a necessidade de armazenar estas seqüências e analisá-las. A natureza dos dados (seqüências de caracteres) permite a aplicação de métodos de análise em texto (Busca e Comparação) (DURBIN et al, 1998). Métodos associados a reconhecimento de padrões também são amplamente utilizados, possibilitando a aplicação de métodos probabilísticos. As técnicas de sequenciamento de DNA possibilitam a obtenção do código de DNA. Porém, a informação de que determinado trecho da seqüência é expresso (responsável pela codificação de uma proteína) não é conhecida. Algumas proteínas codificadas pela seqüência de DNA são, isoladamente ou em conjunto, responsáveis por uma determinada funcionalidade de um organismo. Localizar estas regiões funcionais é um problema de análise de DNA. Alguns autores denominam este tipo de análise como Segmentação de DNA (BOYS, 2004). Outros denominam este problema como análise de sequências de DNA heterogêneas (CHURCHILL, 1992).

As Cadeias de Markov Ocultas são modelos probabilísticos baseados em Cadeias de Markov, e fazem parte da teoria de Processos Estocásticos. As Cadeias de Markov Ocultas foram propostas por Baum (1966), tendo sua formalização complementada por uma série de artigos posteriores, publicados até meados da década de 70. Estes modelos são referenciados na literatura como *HMM* (*Hidden Markov Models*) e, por simplicidade, utilizaremos esta denominação.

O objetivo da modelagem HMM consiste em encontrar o ajuste dos parâmetros do modelo que maximize a verossimilhança, ou seja, localizar a melhor sequência de estados, que maximize a probabilidade. Para tal, faz-se uso de algoritmos dinâmicos, ou simplesmente métodos interativos que consiste em métodos de busca do estado mais provável a cada unidade de tempo.

1.1 Processos Estocásticos

Um processo estocástico *{X(t), t ∈ T}* é uma coleção de variáveis aleatórias onde *t* representa, na maioria das vezes, o tempo. E *X(t)* representa o estado do processo no tempo *t*. O Conjunto *T* é chamado o conjunto índice do processo (Hoel et al., 1972).

- Se *T* é um conjunto enumerável, então *{X(t), t ∈ T}* é um processo estocástico discreto no tempo.
- Se *T* é um conjunto não enumerável ou *T* é um intervalo aberto ou fechado da reta, então *{X(t), t ∈ T}*, é um processo estocástico contínuo no tempo.

1.2 Cadeias de Markov

Uma cadeia de Markov é um processo estocástiico, ou seja, é uma seqüencia $X_1, X_2, X_3, \ldots X_n$ de variáveis aleatórias. O conjunto de valores que tais variáveis podem assumir, é chamado de espaço de estados, onde X_n denota o estado do processo no tempo *n*. Se a distribuição de probabilidade condicional de X_{n+1} nos estados passados é uma função apenas de X_n, então:

$$\Pr\left(X_{n+1} = x \mid X_0, X_1, X_2, ..., X_n\right) = \Pr\left(X_{n+1} = x \mid X_n\right) \qquad (1)$$

Em termos gerais, quando a probabilidade de algo acontecer no tempo *t*+1 depender somente do que ocorreu no tempo *t*, temos uma cadeia de Markov (ROSS, 2003). A análise de uma Cadeia de Markov caracteriza-se principalmente pelo cálculo das probabilidades de transições em *n* passos. Fundamentais, portanto, são as matrizes de probabilidades de transição em *n* passos,

$$\mathrm{P}(n) = \left\| \mathrm{P}_{ij}^{(n)} \right\|, \qquad (2)$$

onde $P_{ij}^{(n)}$ denota a probabilidade que o processo vá do estado *i* para o estado *j* em *n* transições. As probabilidades P_{ij} satisfazem as condições:

$$\mathrm{P}_{ij} \geq 0, \; para \quad i = 0, 1, 2, ... \tag{3}$$

$$\sum_{i=0}^{\infty} \mathrm{P}_{ij} = 1 \; para \quad i = 0, 1, 2, ... \tag{4}$$

Um Processo de Markov está completamente definido quando sua matriz de probabilidades de transição e seu estado inicial X_0 (ou, mais especificamente, a distribuição de probabilidade de X_0) estão especificados. Essa especificação é demonstrada a seguir:

Prova

Seja $\Pr(X_0 = i) = p_{i0}$. É suficiente mostrar como calcular as quantidades

$$\Pr\{X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n\}, \tag{5}$$

uma vez que qualquer probabilidade envolvendo $X_{j1}, X_{j2}, ..., X_{jk}$, para, $j_1, j_2, ..., j_k$ pode ser obtida, de acordo com a lei da probabilidade total, somando os termos da forma (5).

Pela definição de probabilidade condicional, temos:

$$\Pr\{X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n\} = \tag{6}$$
$$\Pr\{X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_{n-1} = i_{n-1}\} \cdot \Pr\{X_n = i_n \mid X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_{n-1} = i_{n-1}\}.$$

Agora, pela definição de Processos de Markov:

$$\Pr\{X_n = i_n \mid X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_{n-1} = i_{n-1}\} \tag{7}$$
$$= \Pr\{X_n = i_n \mid X_{n-1} = i_{n-1}\} = P_{in-1,in}$$

Substituindo (7) em (6), temos:

$$\Pr\{X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n\}$$

$$= \Pr\{X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_{n-1} = i_{n-1}\} \cdot P_{in-1,in}$$

Assim, por indução, (5) torna-se

$$\Pr\{X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n\} = p_{i0} P_{i0,i1} P_{i1,i2} \cdots P_{in-1,in-2} P_{in-1,in} \quad (8)$$

Observa-se, então, que todas as probabilidades de dimensão finita podem ser obtidas a partir das probabilidades de transição e da distribuição inicial. O processo é, portanto, definido por essas quantidades.

1.3 Cadeias de Markov Ocultas

Uma vez especificados as bases nas quais os HMMs estão fundamentados, pode-se definir de forma mais segura o método, cuja aplicação é o objetivo do presente estudo. Os Modelos de Markov Ocultos (HMM) é uma importante ferramenta na modelagem de sequências de variáveis aleatórias. Estes modelos estocásticos estão sendo aplicados em várias áreas do conhecimento. As Cadeias de Markov Ocultas foram propostas por BAUM (1966). É uma técnica utilizada em diversas áreas do conhecimento, tais como, Climatologia (HUGHES et al, 1999), Econometria (RYDEN et al, 1998), Reconhecimento de Texto (VLONTZOS et al, 1992), Processamento de Imagens (AAS et al, 1999) e Reconhecimento de Fala (RABINER, 1989).

Os HMM são definidos como modelos onde a observação da formação do sistema se dá de forma indireta, como função probabilística da transição entre os estados definidos num espaço de estados discreto e finito. Por mais que todos os parâmetros do modelo* sejam conhecidos, a evolução que demonstre a formação de tal sistema que governa esse processo, contínua oculta. Em outras palavras, não se

---

* São chamados parâmeros do modelo o conjunto de valores $\lambda = (A, B, \Pi)$ que definem o modelo, onde Π é o vetor de probabilidade inicial de cada estado da cadeia de Markov Oculta, A é a matriz que define a probabilidade de transição entre esses estados e B é a matriz de probabilidade de emissão.

sabe qual o caminho ou sequência de passos exatos que levaram a uma determinada observação.

Sejam

1. Um espaço de estados $S = \{s_1, s_2, ..., s_N\}$
2. Um conjunto de observáveis $Y = \{y_1, y_2, ..., y_M\}$
3. Uma variável estocástica Q a assumir valores do espaço de estados S em diferentes instantes de tempo
4. Uma variável estocástica O a assumir valores do conjunto de observáveis Y em diferentes instantes de tempo
5. Uma distribuição de probabilidade inicial para cada estado $\Pi = \pi_i$, tal que, $\Pi = \pi_i$, $\pi_i = P(q_0 = s_i)$
6. Uma distribuição de probabilidade de transição entre estados $A = a_{ij}$, tal que, $a_{ij} = P(q_t = s_j \mid q_{t-1} = s_i)$
7. Uma distribuição de probabilidade de observação $B = \{b_{ij}(k)\}$, tal que, $b_{ij}(k) = P(O_t = y_k \mid q_{t-1} = s_i, q_t = s_j)$ associada às transições do estado s_i para o estado s_j.

A detecção de regiões homogêneas em sequências de DNA tem como objetivo encontrar segmentos distintos responsáveis por funções de regulação celulares distintas dentro de um determinado organismo. Uma metodologia proposta por HINKLEY et al, (1970) e SMITH (1975) sugere que estas regiões de regulação possam ser encontradas a partir da estimação de múltiplos pontos de mudança em sequências de variáveis aleatórias, no entanto, essa técnica se mostra insatisfatória para o problema associado a busca de homogeneidade em sequências de DNA, devido a dependência que existe entre as bases de nucleotídeos.

No que tange a bioinformática, as Cadeias de Markov Ocultas são aplicadas em uma série de problemas, tais como, o problema do Alinhamento Múltiplo (GUSFIELD, 1997) e o problema de detecção de regiões homogêneas em seqüências de DNA (CHURCHILL, 1989), sendo este o alvo principal desse capítulo.

A Grosso modo, pode-se definir as Cadeias de Markov Ocultas como uma classe de processos estocásticos, baseada em uma cadeia de Markov *{X(t) t ∈ /N}* não observável que descreve a evolução dos estados do processo. Considerando, então, a importância e utilidade dos HMMs em várias áreas de pesquisa, nosso principal objetivo é estudar esses modelos numa aplicação em um conjunto de sequências genéticas relacionados à três espécies com propensão ao desenvolvimento da mastite.

1.4 Aplicação do método HMM em sequências genéticas

As características associadas a um determinado organismo estão codificadas em seqüências de DNA que, por sua vez, compõem os genes. Um determinado gene está associado à síntese de uma determinada proteína. Em resumo, cada segmento de DNA é responsável pela codificação de uma determinada proteína, que é responsável por um tipo de funcionalidade presente no organismo.

No processo de sequenciamento de DNA de um genoma, não é possível determinar a localização exata de um gene ou mesmo se um segmento composto de vários genes é responsável pela codificação de alguma funcionalidade no organismo (OLIVEIRA 2004). É necessário então, fazer uso de algum método que possa discriminar os segmentos que codificam determinadas proteínas com distintas funcionalidades.

Alguns estudos sobre os segmentos distintos de DNA indicaram que as freqüências de bases (nucleotídeos) podem auxiliar na identificação e distinção entre os segmentos com diferentes funcionalidades. Alguns segmentos apresentam a proporção da ocorrência de uma determinada base, ou de um determinado grupo de bases similar. Estes segmentos são denotados como regiões homogêneas. Uma característica importante é que variações nas proporções das freqüências de C+G normalmente refletem distinções funcionais ou estruturais entre os segmentos.

O método HMM é utilizado na localização de regiões homogêneas com relação à emissão de C+G, de modo a se encontrar regiões que são funcionais ou

estruturalmente distintas. CHURCHILL (1989) foi pioneiro neste tipo de aplicação e em problemas ligados a biologia computacional.

Sequências de DNA de genes associados á resistência inflamatória da mastite foram utilizadas para construção da idéia de segmentação com o intuito de observar regiões que possam estar associadas à expressão de uma determinada característica em espécies distintas, bem como a realização de um comparativo visual entre as espécies. A ligação de sinas em dinucleotídeos C+G geralmente resulta na repressão da expressão gênica e está envolvida em vários processos biológicos normais, tais como o controle de desenvolvimento de uma determinada característica.O número de estados (regiões de homogeneidade na expressão de C+G) é definido utilizando-se critérios para adequação de modelos HMM. Cada região com um distinto padrão de emissão de C+G é considerada como um estado distinto (OLIVEIRA, 2004).

Com base na proposta das Cadeias de Markov Ocultas, alguns procedimentos deverão ser elaborados para construção do método. O início se da a partir do cálculo das probabilidades de transição levando em consideração uma determinada sequência genética, ou seja, o intuito consiste em encontrar a probabilidade de dado que apareceu uma base Citosina num determinado ponto da sequência, qual será a probabilidade da próxima base ser uma Guanina?. Essa é a intuição básica da modelagem presente no uso de Cadeias de Markov.

1.5 Algoritmos *Forward-Backward*

Esses dois algoritmos se preocupam em verificar, dado um conjunto de observações $X_1, X_2, X_3, \ldots X_n$, qual a probabilidade desse conjunto ter sido gerado pelo modelo de probabilidades $\lambda(A,B,\pi)$, onde $A = a_{ij} = P(s_{t+1} = j \mid s_t = i)$; $B = b_j(O_t) = P(O_t \mid s_t = j)$ e; $\pi = \pi_i = P(s_1 = i)$.

A aplicação dos algoritmos resulta, para o algoritmo forward e backward, respectivamente:

$$P(O \mid \lambda) = \sum_{i \in S_F} \alpha_T(i) \tag{9}$$

$$P(O \mid \lambda) = \sum_{i \in S_F} \pi_i b_i O_1 \beta_1(i) \tag{10}$$

Onde:

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, ..., O_T, s_t = i \mid \lambda) \tag{11}$$

$$\alpha_t(i) = P(O_1, O_2, ..., O_t, s_t = i \mid \lambda) \tag{12}$$

1.6 O algoritmo Viterbi

O algoritmo Viterbi foi concebido por Andrew Viterbi em 1967 como um algoritmo de decodificação de códigos convolucionais (FORNEY, 1973). O algoritmo tem aplicação universal na decifração dos códigos convolucionais utilizado em CDMA (Code Division Multiple Access, ou Acesso Múltiplo por Divisão de Código) e celular digital GSM, modem dial-up, satélite de comunicações no espaço profundo, e LANs sem fio. Também é comumente usado em reconhecimento de fala, linguística computacional e bioinformática. Por exemplo, em discurso para texto (reconhecimento de voz), o sinal acústico é tratado como a seqüência de eventos observados, e uma seqüência de texto é considerada a "causa secreta" de que o sinal acústico.

1.7 O algoritmo Baum-Welch

O algoritmo Baum-Welch tem como meta a busca pelos melhores parâmetros que são encontrados a partir da otimização da probabilidade de observação de uma dada sequência. Considerando a definição do algoritmo de Baum-Welch, como apresentada no artigo RABINER e JUANG (1986), temos que

$$\xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j \mid O, \lambda); \tag{13}$$

ou seja, $\xi_t(i,j)$ é a probabilidade conjunta de estar no estado S_i no instante t e no estado S_j no instante $t+1$, dado o modelo inicial $\lambda=(A,B,\pi)$ e a sequência de treinamento O. Essa variável pode ser expressa em termos das variáveis *forward* e *bacward*, tomando a seguinte forma:

$$\xi_t(i,j)=P(q_t=S_i,q_{t+1}=S_j \mid O,\lambda)=\frac{P(q_t=S_i,q_{t+1}=S_j \mid O,\lambda)}{P(O \mid \lambda)}= \tag{14}$$

$$=\frac{\alpha_t(i)a_{ij}b_{ij}(O_{t+1})\beta_{t+1}(j)}{P(O \mid \lambda)}=\frac{\alpha_t(i)a_{ij}b_{ij}(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_{ij}(O_{t+1})\beta_{t+1}(j)}$$

Agora, façamos o somatório da equação (1) sobre o índice j, $1\le j\le N$:

$$\sum_{j=1}^{N}\xi_t(i,j)=\sum_{j=1}^{N}\frac{\alpha_t(i)a_{ij}b_{ij}(O_{t+1})\beta_{t+1}(j)}{P(O \mid \lambda)}=\frac{\alpha_t(i)\left[\sum_{j=1}^{N}a_{ij}b_{ij}(O_{t+1})\beta_{t+1}(j)\right]}{P(O \mid \lambda)} \tag{15}$$

$$=\frac{\alpha_t(i)\beta_t(i)}{P(O \mid \lambda)}$$

Não existe uma solução ótima para o problema da determinação de parâmetros de um HMM. A solução mais recomendada é o uso de uma probabilidade inicial aleatória, sobre a qual é aplicado um método de reestimação iterativa, gerando assim as sequências de observações com mais alta probabilidade. Com base nas freqüências de ocorrências, o novo modelo, $\lambda'=(A',B',\Pi')$ é calculado com o algoritmo de reestimação de *Baum-Welch.* Com uso desse algoritmo é possível melhorar a probabilidade de uma sequência observada.

2. Materiais e Métodos

Foi feito uso do *software* estatístico *R* versão 2.11.1 para realizar o cálculo das matrizes de transição, levando-se em consideração todos os estados das sequências genéticas (A, C, T e G), assim como a construção dos gráficos que mostram as regiões conhecidas como segmentação de DNA.

Utilizou-se o pacote HMM (*Hidden Markov Models*) em que este possui todos os algoritmos necessários implementados para alcançar os resultados envolvidos com as estimativas e as re-estimativas das matrizes de transição, como também para busca das probabilidades das regiões homogêneas e realizar inferências sobre os estados de transição de probabilidade entre as sequências.

As sequências genéticas escolhidas para aplicação do método HMM são resultantes do consenso obtido entre três espécies (*Bos taurus, Capra hiscus e Ovies aries*) para dois genes distintos, o HSP70.1 e o NRAMP-1 que desenvolvem um papel importante na resposta de resistência à mastite. O Uso dessas sequências tem como principal objetivo a aplicação do método HMM como ferramenta estatística.

3. Resultados e Discussões

A localização das regiões homogêneas com conteúdo de C+G é denotada por segmentação de DNA. Uma analise preliminar desses segmentos em uma determinada sequência foi realizada, observando-se a frequência de C+G e em seguida, notando a existência de um comportamento sistemático através de um gráfico construído a partir de informações binárias, ou seja, (C ou G = 1; A ou T = 0).

Dois pontos devem ser ressaltados:

a) Devido a quantidade de bases de nucleotídeos para as duas primeiras espécies, *Bos taurus e Capra hircus* serem grande, pois as mesmas estão relacionadas ao seqüenciamento completo do gene, observa-se a ocorrência de uma quebra na estrutura do gráfico.
b) Na figura 3 é demonstrada apenas como ilustração a segmentação que ocorre no mesmo gene, mas para a espécie *Ovis Aries.* Visualmente, percebe-se que a quantidade de bases é inferior quando comparadas as anteriores, pois a sequência utilizada não corresponde à completa.

É possível observar os seguimentos dentro das sequências analisadas que podem estar associados a expressão de uma dada característica. As regiões escurecidas representam as áreas segmentadas.

Considerando o gene HSP70.1, temos uma estrutura contendo a segmentação mais presente no início das sequências (Figuras 1 e 2). Com relação a espécie *Ovis aries,* mesmo a sequência não sendo a completa, não é possível observar uma leve formação de padrões, mas nada que possa ser conclusivo com relação a segmentação C+G.

Os mesmos gráficos foram construídos com as sequências genéticas para as três espécies em estudo afim de observar a existência de padrões. Observa-se que para a espécie *Bos taurus,* a formação de regiões é bastante expressiva.

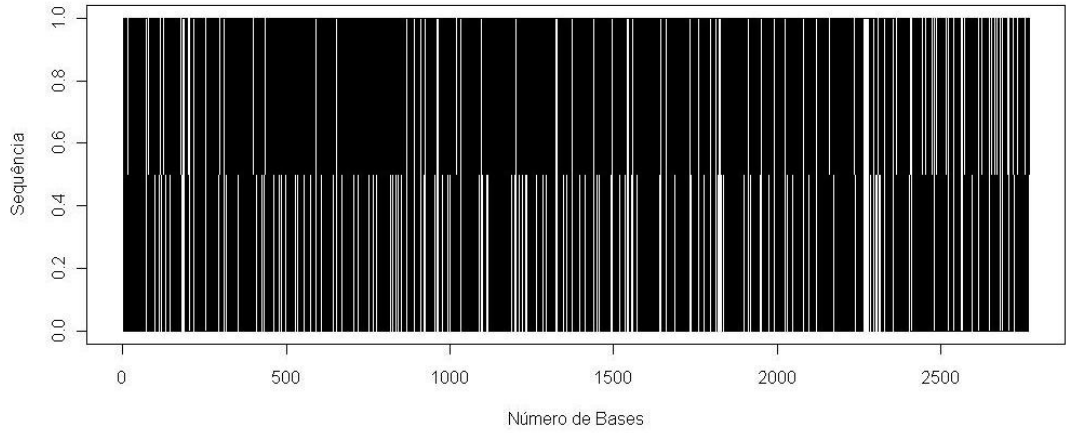

Figura 1: Segmentação do gene HSP70.1 relacionada á espécie *Bos taurus*

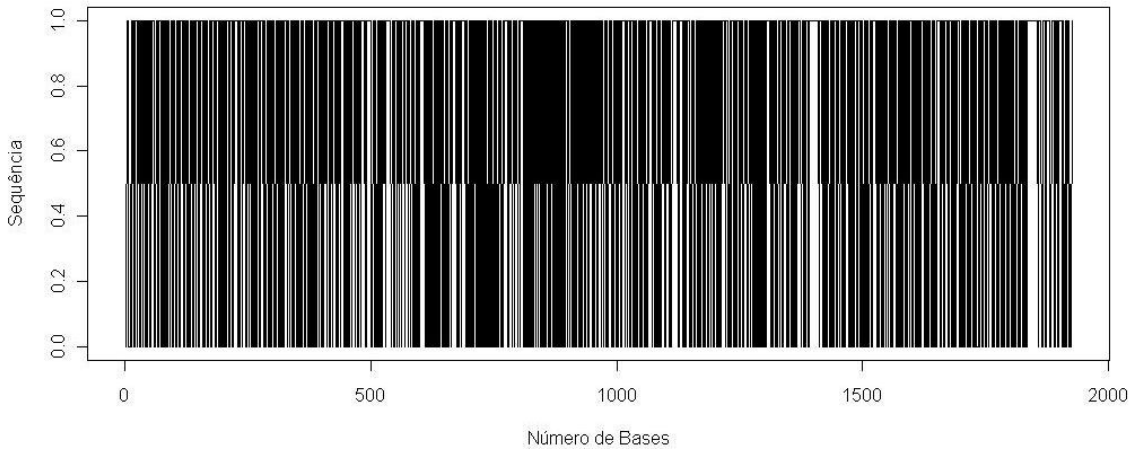

Figura 2: Segmentação do gene HSP70.1 relacionada á espécie Capra hircus

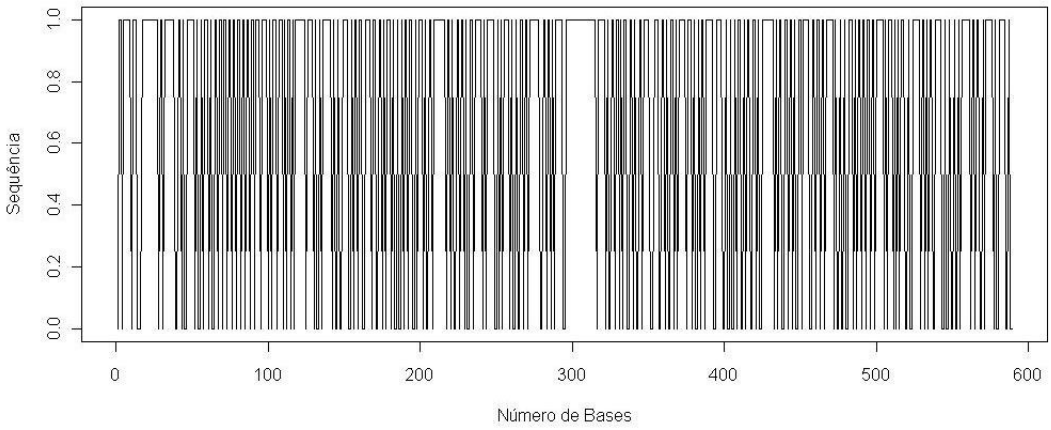

Figura 3: Segmentação do gene HSP70.1 relacionada á espécie *Ovis aries*

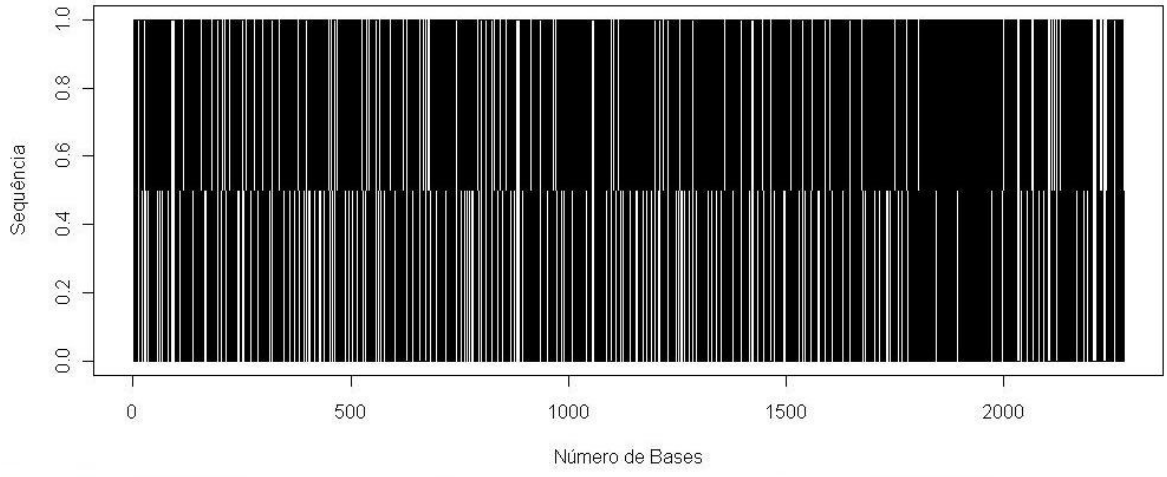

Figura 4: Segmentação do gene NRAMP-1 relacionada á espécie *Bos taurus*

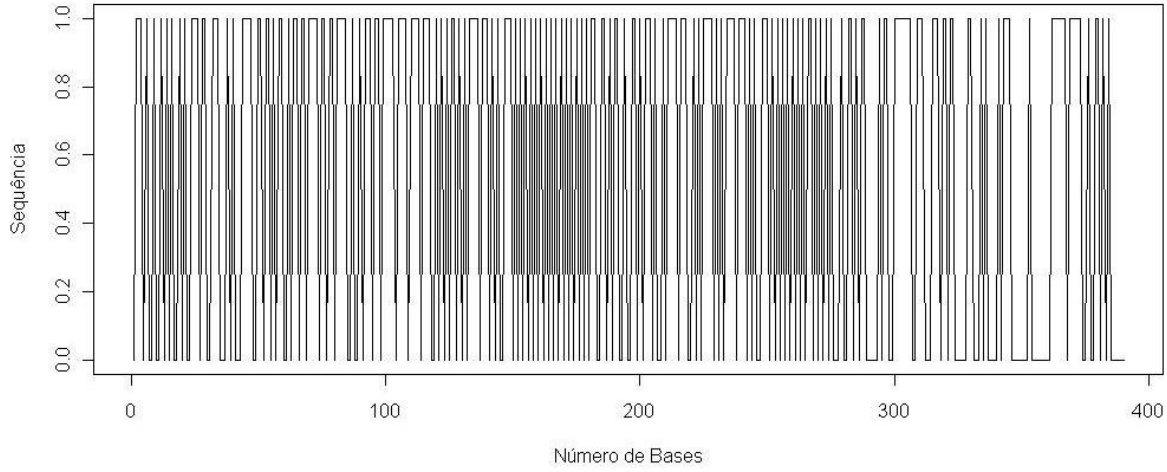

Figura 5: Segmentação do gene NRAMP-1 relacionada á espécie *Capra hircus*

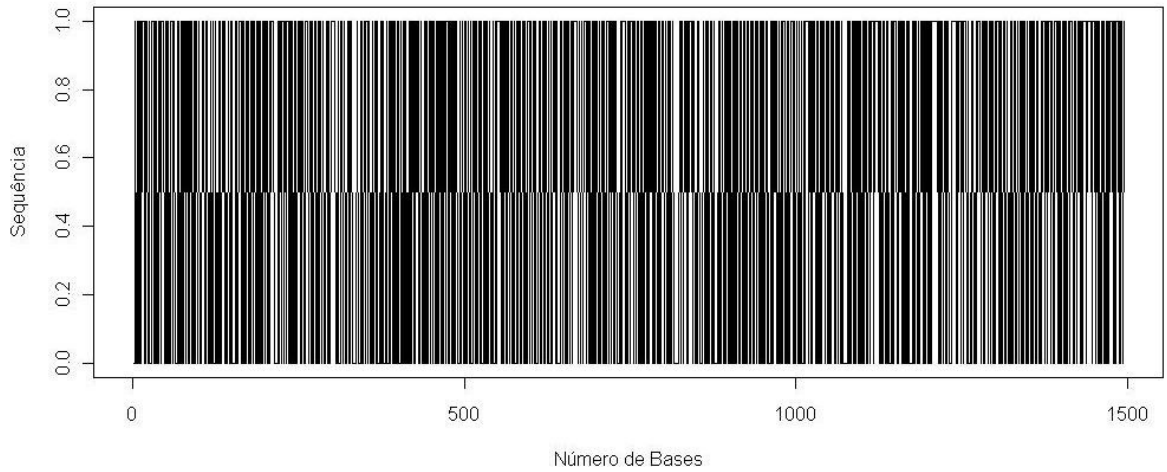

Figura 6: Segmentação do gene NRAMP-1 relacionada á espécie *Ovis aries*

É preciso resolver os três problemas básicos associados ao uso dos HMM's (avaliação, decodificação e treinamento do modelo). O primeiro problema é resolvido a partir da construção das matrizes de transição considerando apenas a mudança de estado inicial *i* para o estado seguinte *j.* Com base nos estados iniciais (A, C, T, e G).

Tabela 1: Matriz de transição entre as bases presentes no consenso com suas respectivas freqüências e probabilidades do gene HSP70.1

Bases	A	C	T	G	Frequência	Probabilidades
A	0,20744	-	-	-	511	0,22521
C	0,27170	0,26367	-	-	622	0,27413
T	0,10973	0,28678	0,23192	-	401	0,17673
G	0,25986	0,26939	0,13605	0,33333	735	0,32393

Tabela 2: Matriz de transição entre as bases presentes no consenso com suas respectivas freqüências e probabilidades do gene NRAMP-1

Bases	A	C	T	G	Frequência	Probabilidades
A	0,23558	-	-	-	1855	0,25286
C	0,307028	0,319124	-	-	1736	0,23664
T	0,145677	0,255276	0,219197	-	1469	0,20025
G	0,278998	0,20123	0,141916	0,377856	2276	0,31025

A partir da definição das matrizes de probabilidade e de transição a aplicação do método HMM pode ser aplicado sem maiores problemas, uma vez que tais informações são requisitos iniciais.

Como já foi definido anteriormente, o objetivo principal na aplicação do método HMM consiste em encontrar a região mais provável que pode aparecer numa determinada sequência genética. Com base na matriz de probabilidade inicial, ou *priores* da sequência genética e a matriz de transição entre as bases A,C,T e G é feito interativamente. O cálculo das probabilidades posteriores de uma determinada base passar do estado *i* para o estado *j* numa dada seqüência de observações e de um determinado modelo de Markov oculto.

O desafio é determinar os parâmetros ocultos a partir dos parâmetros observáveis. Os parâmetros extraídos do modelo podem então ser usados para realizar novas análises, O reconhecimento de padrões é um dos resultados mais relevantes que pode-se fazer uso com aplicação do HMM.

A probabilidade a posteriori de estar em um estado X no instante k pode ser calculada a partir de dois métodos *backward* e *forward*, ver detalhes em (Rabiner, 1989) e (Oliveira, 2005).

$$(X_k = X \mid E_1 = e_1, \ldots, E_n = e_n) = \frac{f(X,k) \cdot b(X,k)}{P(E_1 = e_1, \ldots, E_n = e_n)}, \qquad (9)$$

em que $E_1 = e_1, \ldots, E_n = e_n$ é uma seqüência de emissões observadas e X_k é uma variável aleatória que representa o estado no tempo *k*

Foi considerado somente o algoritmo *forward* (para frente) para o cálculo das posteriores pois a forma de entrada das probabilidades iniciais contempla somente a mudança de um estado inicial para um próximo estado. Com uso da função *forward*, em que a mesma já encontra-se implementada pacote HMM no software R (ver algoritmo em anexo). Os resultados seguem abaixo:

Tabela 3: Matriz com as probabilidades posteriores entre as bases presentes no consenso do gene HSP70.1

Bases	Matriz de Transição			
	A	**C**	**T**	**G**
A	0,22530	-	-	-
C	0,27427	0,27439	-	-
T	0,17679	0,17687	0,17688	-
G	0,32362	0,32376	0,32376	0,32406

Tabela 4: Matriz com as probabilidades posteriores entre as bases presentes no consenso do gene NRAMP-1

Bases	Matriz de Transição			
	A	**C**	**T**	**G**
A	0,24924	-	-	-
C	0,23760	0,23874	-	-
T	0,20152	0,20239	0,20209	-
G	0,31162	0,31321	0,31299	0,31137

O segundo problema do uso dos modelos HMM é a decodificação e este problema é resolvido a partir da implementação do algoritmo de Viterbi que tem como meta calcular a sequência de estados ocultos mais provável em cada modelo HMM. O algoritmo de Viterbi tem o propósito de identificar padrões, que normalmente podem expressar uma característica.

Ainda no pacote HMM é possível fazer uso do algoritmo de Viterbi que já encontra-se implementado (ver algoritmo em anexo). As sequências mais prováveis foram definidas com o seguinte padrão: G,G,G,G. Esse resultado é algo esperado, pois analisando-se as freqüências em ambas as sequências, a guanina é a base que mais aparece. Essa representação também é feita graficamente (Figuras 7 e 8)

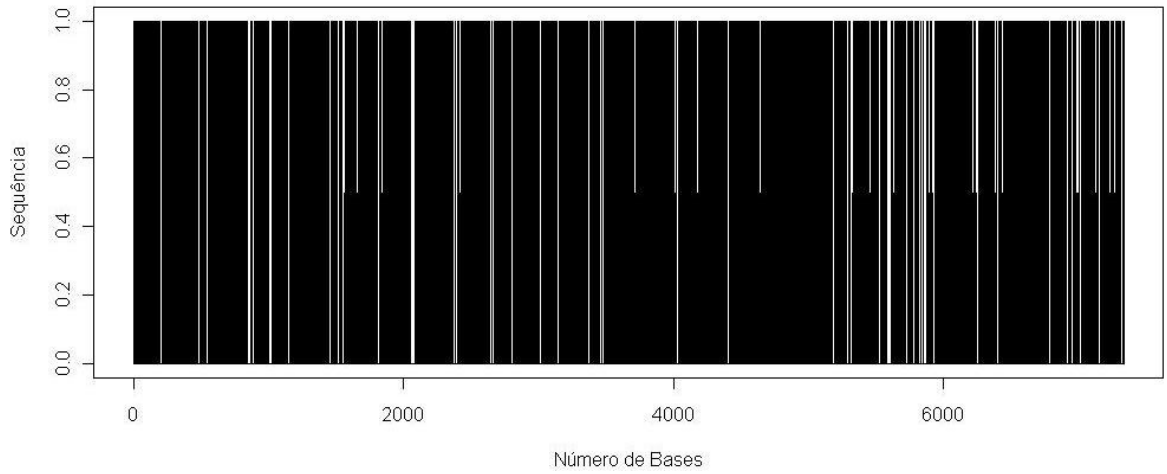

Figura 7: Padrão fornecido pelo algoritmo de Viterbi na sequência consenso do gene NRAMP-1

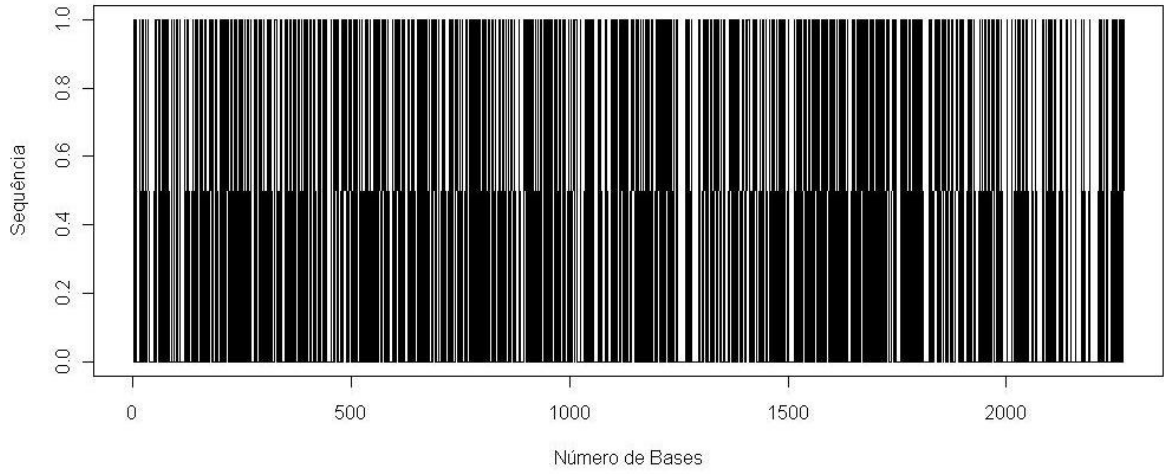

Figura 8: Padrão fornecido pelo algoritmo de Viterbi na sequência consenso do gene HSP70.1

Os gráficos acima demonstram o padrão definido pelo algoritmo de viterbi. As regiões mais densas estão relacionadas à freqüência de Guanina em toda a sequência. É perceptível que graficamente é possível corroborar as probabilidades associadas com as matrizes de transição (Tabelas 3 e 4).

A partir da otimização da probabilidade de observação de uma dada sequência, foi feito uso do algoritmo *Baum-Welch* para solucionar o terceiro problema no uso do HMM, com o objetivo de encontrar os melhores parâmetros. Como já foi dito, não existe uma solução ótima para o problema da determinação dos melhores parâmetros associados ao dado modelo HMM. A partir de uma probabilidade inicial, foi aplicado o método de re-estimação iterativamente para

geração de uma nova sequência de observações, no caso, formada por uma matriz de transição, com a mais alta probabilidade de ocorrência.

O novo modelo, $\lambda'=(A',B',\Pi')$ é calculado com o algoritmo de re-estimação de *Baum-Welch* (ver algoritmo em anexo). E assim, torna-se possível melhorar a probabilidade de uma sequência observada. Ainda com uso do pacote HMM do software R, as reestimativas foram realizadas com a implementação das sequências genéticas. Os resultados são demonstrados em forma de tabela com as chamadas probabilidades de reemissão.

Quando os caminhos que levam a um determinado resultado são desconhecidos, nas reestimativas não existe uma equação fechada para estimar diretamente os valores dos parâmetros, de forma a encontrar o melhor modelo λ que constitui um problema de otimização, ou seja, escolhe-se $\lambda'=(A',B',\Pi')$ tal que a $P(O/\lambda)$ das seqüências do conjunto de treinamento sejam localmente maximizadas. Qualquer um dos algoritmos padrões para otimização de funções contínuas pode ser usado, entretanto, um método iterativo que tem uma interpretação probabilística natural, o algoritmo de *Baum-Welch* é usado para escolher os parâmetros do modelo λ.

Os resultados observados são considerados como sendo ótimos à medida que os resultados se aproximem da distribuição de probabilidade inicial de cada símbolo, as bases de nucleotídeos. Com uso do algoritmo foram realizadas um número de implementações, tantas quantas fossem necessárias para se chegar aproximadamente ao objetivo do método. Depois de algumas tentativas, foi possível observar que o número ótimo para gerar a sequência ótimas seria algo em torno de 500 interações. Observando as probabilidades de remissão (Tabelas 5 e 6), quanto mais próximas as reestimações estiverem da distribuição de probabilidade, melhor será considerado o ajuste do algoritmo. Com isso é possível identificar em termos quantitativos a presença de um gene numa dada sequência ou em termos mais gerais, a probabilidade dessa ocorrência.

Tabela 5: Probabilidades de reemissão do consenso do gene HSP70.1

Bases	Matriz de Reemissão			
	A	**C**	**T**	**G**
A	0,2499900	-	-	-
C	0,2500010	0,2501000	-	-
T	0,2501000	0,2500010	0,2500120	-
G	0,2499910	0,2499010	0,2499800	0,2500200

Tabela 6: Probabilidades de reemissão do consenso do gene NRAMP-1

Bases	Matriz de Reemissão			
	A	**C**	**T**	**G**
A	0,2500200	-	-	-
C	0,2499800	0,2501000	-	-
T	0,2500800	0,2490800	0,2500010	-
G	0,2501800	0,2499800	0,2449800	0,2406800

4. Conclusões

Foi feito uso de uma importante técnica na área da bioinformática para mostrar probabilisticamente a relação de transição entre bases de nucleotídeos. Com uso dessa técnica é possível mostrar uma forma de identificação de padrões em sequências de DNA como também calcular as probabilidades de reemissão associadas a uma dada sequência genética.

A técnica HMM vem como uma análise alternativa em meio a tantas outra que existem, cujo objetivo consiste em identificar padrões. No entanto, o diferencial no HMM é que nesta técnica existe um embasamento teórico bem fundamentado matematicamente que permite a extração de informações estatísticas contidas em uma determinada quantidade de seqüências.

Diante da grande massa de dados que são gerados todos os dias, análises na área de genética molecular que mostrem informações relevantes são necessárias, para que além de identificar, também seja possível mensurar o comportamento das inter-relações entre diversos organismos.

A técnica HMM permite a extração de informações estatísticas contidas em seqüências genéticas. Como foi mostrado neste trabalho, é possível identificar e reestimar regiões homogêneas, o que vem a fornecer padrões que em muitos dos casos podem estar associados a expressão de determinadas características.

5. Trabalhos Futuros

Embora ainda tenha muito caminho a ser percorrido, pelo que existe na literatura hoje, As Cadeias de Markov Ocultos possuem um grande potencial pra modelagem estatística de seqüências de DNA e apresentam-se promissoras e com muito ainda a ser explorado.

- ✓ Realização de estudos comparativos do comportamento assintótico dos algoritmos aqui utilizados;

- ✓ Realizar uma implementação da técnica HMM considerando as sequências genéticas no contexto de séries temporais

Referências Bibliográficas

AAS, K.; EIKVIL, L.; Text page recognition using grey-level features and hidden Markov models. Pattern recognition, v. 29, n. 6, p. 977-985, jun. 1996.

BAUM, L. E.; PETRIE, T.; Statistic inference for probabilistic functions of finite state Markov chain. Annals of Mathematical Statistics, v. 37, p. 1554-1563, 1966.

CHURCHILL, G.; Hidden Markov Chain and the analysis of genome structure. Computer Chemistry, v. 16, n. 2, p. 107-115, 1992.

CHURCHILL, G; Stochastic models for Heterogeneous DNA sequences. Bulletin of Mathematical Biology, v. 51, n.1, p. 79-94, 1989.

CRUZ, C. D.; REGAZZI, A. J. Modelos biométricos aplicados ao melhoramento genético. UFV, 2001. 390p.

DURBIN, R; EDDY, S.; KROGH, A; MITCHISON, G.; Biological Sequence Analysis: Probabilistic Models o Proteins and Nucleic Acids. Cambridge: Cambridge University Press, 1999. 368 p.

EPHRAIM Y., MERHAV N.: Hidden Markov processes. IEEE Trans. Inform. Theory 48 p.1518-1569, 2002

FORNEY, G. D. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

FUTUYMA, DJ. 1992. Biologia evolutiva. 2. ed Sociedade Brasileira de Genética.

GRAUR, D. and Li, W.-H. 1999. Fundamentals of Molecular Evolution. Second Edition, Sinauer Associates.

GUSFIELD, DAN. Algorithms on *Strings*, Trees, and Sequences – Computer Science and Computational Biology. Cambrigde University Prees, 1997

HINKLEY, D. V; Inference about the change point in a sequence of random variables. Biometrika, v. 57, n.1, p.1-17, 1970.

HOEL, P. G.; PORT. S. C. & STONE. C. J. Introduction to Stochastic Processes. Houghton-Mifflin, Boston. 1972.

HUGHES, J.P.; GUTTORP, P.; CHARLES, S.P.; A non-homogeneous hidden Markov model for precipitation occurrence. Applied Statistics, v. 48, n. 1. p. 15-31, 1999.

LAWRENCE R. Rabiner: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77(2) p.257-286, 1989.

CAPPE O., MOULINES E., RYDEN T.: Inference in Hidden Markov Models. Springer. ISBN 0-387-40264-0.

OLIVEIRA, D. C. de. *Cadeias de Markov com Estados Latentes com aplicações em análises de seqüências de DNA*. 2004. 190f. Dissertação (Mestrado em Agronomia - Estatística e Experimentação Agropecuária) – Departamento de Ciências Exatas, Universidade Federal de Lavras, Lavras, 2005.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

RABINER, L.; A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, v. 77, n. 2, p.
257-286, fev. 1989.

Ross, S.M. (2003). Introduction to Probability Models. 8a edição

RYDEN, T.; TERASVIRTA, T.; ASBRINK, S.; Stylized Facts of Daily Return Series and the Hidden Markov Model. Journal of applied econometrics, v. 13, n. 3, p. 217-230, mai. 1998.

Scientific Software Development - Dr. Lin Himmelmann and www.linhi.com (2010). HMM: HMM - Hidden Markov Models. R package version 1.0. http://CRAN.R-project.org/package=HMM

SMITH, A.F.M.; A bayesian approach to inference about a change-point in a sequence of random variables. Biometrika, v. 62, n. 2, p. 407-416, 1975.

Suzuki, DT et al. 1998. Introdução a genética. 6 ed. Guanabara Koogan

VLONTZOS, J.; KUNG, S.; Hidden Markov Models for character Recognition. IEEE transactions on image processing, v. 1, n. 4, p. 539-543, out. 1992.

Anexos

############## Script do R ##############

```
###### Importando as sequências genéticas ######
###### para contagem das frequências      ######

###### Sequência do consenso do gene HSP70.1 ######
seqa=read.fasta("C:\\Rita
Project\\DISERTACAO\\KMO\\FINAL\\seq1.fasta")
table(seqa)

###### Sequência do consenso do gene NRAMP-1 ######
seqb=read.fasta("C:\\Rita
Project\\DISERTACAO\\KMO\\FINAL\\seq2.fasta")
table(seqb)


###### Iniciando a aplicação do HMM ######

# Initialise HMM
hmm = initHMM(c("A","C","T","G"),
c("Adenina","Citosina","Timina","Guanina"),startProbs=c(0.2252
1,0.27413,0.17673,0.32393),transProbs=t(matrix(c(0.20744,0.283
76,0.17025,0.33855,0.27170,0.26367,0.19453,0.27010,0.10973,0.2
8678,0.23192,0.37157,0.25986,0.26939,0.13605,0.33333),4)),)
emissionProbs=NULL)
print(hmm)

# Construindo a sequências de observações para posteriori

# Sequence of observations
observations=c("Adenina","Citosina","Timina","Guanina")

# Calculate posterior probablities of the states
posterior = posterior(hmm,observations)
print(posterior)

simHMM(hmm, 1000)
table(simHMM(hmm, 100))

# Calculando o algoritmo de Viterbi
viterbi = viterbi(hmm,observations)
print(viterbi)

# Implementando o Algoritmo BaumWelch nas sequências genética

# Sequence of observations
observations = c("Adenina","Citosina","Timina","Guanina")
# Calculate forkward probablities
logForkwardProbabilities = forkward(hmm,observations)
```

```
print(exp(logBackwardProbabilities))

# Sequence of observation
a =
sample(c(rep("Adenina",100),rep("Citosina",100),rep("Timina",1
00),rep("Guanina",100)))
b =
sample(c(rep("Adenina",100),rep("Citosina",100),rep("Timina",1
00),rep("Guanina",100)))
observation = c(a,b)
# Baum-Welch
bw = baumWelch(hmm,observation,100)
print(bw$hmm)

# Construção das regiões homogêneas

regiao1=c(0,1,0,1,...,0,0,0,1) # Aqui consta apenas uma parte
do vetor
plot(a1,type="l",xlab="Número de Bases",ylab="Sequência")
```

############## Algoritmos do pacote HMM no R ##############

 # Algoritmo forward

```
function (hmm, observation)
{
    hmm$transProbs[is.na(hmm$transProbs)] = 0
    hmm$emissionProbs[is.na(hmm$emissionProbs)] = 0
    nObservations = length(observation)
    nStates = length(hmm$States)
    f = array(NA, c(nStates, nObservations))
    dimnames(f) = list(states = hmm$States, index =
1:nObservations)
    for (state in hmm$States) {
        f[state, 1] = log(hmm$startProbs[state] *
hmm$emissionProbs[state,
            observation[1]])
    }
    for (k in 2:nObservations) {
        for (state in hmm$States) {
            logsum = -Inf
            for (previousState in hmm$States) {
                temp = f[previousState, k - 1] +
log(hmm$transProbs[previousState,
                  state])
                if (temp > -Inf) {
                  logsum = temp + log(1 + exp(logsum - temp))
                }
            }
            f[state, k] = log(hmm$emissionProbs[state,
observation[k]]) +
                logsum
        }
    }
    return(f)
}
```

 # Algoritmo Viterbi

```
function (hmm, observation)
{
    hmm$transProbs[is.na(hmm$transProbs)] = 0
    hmm$emissionProbs[is.na(hmm$emissionProbs)] = 0
    nObservations = length(observation)
    nStates = length(hmm$States)
    v = array(NA, c(nStates, nObservations))
    dimnames(v) = list(states = hmm$States, index =
1:nObservations)
    for (state in hmm$States) {
```

```
        v[state, 1] = log(hmm$startProbs[state] *
hmm$emissionProbs[state,
            observation[1]])
    }
    for (k in 2:nObservations) {
        for (state in hmm$States) {
            maxi = NULL
            for (previousState in hmm$States) {
                temp = v[previousState, k - 1] +
log(hmm$transProbs[previousState,
                  state])
                maxi = max(maxi, temp)
            }
            v[state, k] = log(hmm$emissionProbs[state,
observation[k]]) +
                maxi
        }
    }
    viterbiPath = rep(NA, nObservations)
    for (state in hmm$States) {
        if (max(v[, nObservations]) == v[state,
nObservations]) {
            viterbiPath[nObservations] = state
            break
        }
    }
    for (k in (nObservations - 1):1) {
        for (state in hmm$States) {
            if (max(v[, k] + log(hmm$transProbs[,
viterbiPath[k +
                1]])) == v[state, k] +
log(hmm$transProbs[state,
                viterbiPath[k + 1]])) {
                viterbiPath[k] = state
                break
            }
        }
    }
    return(viterbiPath)
}
<environment: namespace:HMM>
```

 # Algoritmo Baum-Welch

```
function (hmm, observation, maxIterations = 100, delta = 1e-
09,
    pseudoCount = 0)
{
    tempHmm = hmm
```

```
    tempHmm$transProbs[is.na(hmm$transProbs)] = 0
    tempHmm$emissionProbs[is.na(hmm$emissionProbs)] = 0
    diff = c()
    for (i in 1:maxIterations) {
        bw = baumWelchRecursion(tempHmm, observation)
        T = bw$TransitionMatrix
        E = bw$EmissionMatrix
        T[!is.na(hmm$transProbs)] = T[!is.na(hmm$transProbs)]
+
            pseudoCount
        E[!is.na(hmm$emissionProbs)] =
E[!is.na(hmm$emissionProbs)] +
            pseudoCount
        T = (T/apply(T, 1, sum))
        E = (E/apply(E, 1, sum))
        d = sqrt(sum((tempHmm$transProbs - T)^2)) +
sqrt(sum((tempHmm$emissionProbs -
            E)^2))
        diff = c(diff, d)
        tempHmm$transProbs = T
        tempHmm$emissionProbs = E
        if (d < delta) {
            break
        }
    }
    tempHmm$transProbs[is.na(hmm$transProbs)] = NA
    tempHmm$emissionProbs[is.na(hmm$emissionProbs)] = NA
    return(list(hmm = tempHmm, difference = diff))
}
<environment: namespace:HMM>
```