LUCEMBERG DE ARAÚJO PEDROSA

COMPARAÇÃO ENTRE DISTÂNCIA ENTRÓPICA E DISTÂNCIA GENÉTICA PARA ANÁLISE DE SEQUÊNCIAS DE DNA

RECIFE-PE – JULHO/2013.

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

COMPARAÇÃO ENTRE DISTÂNCIA ENTRÓPICA E DISTÂNCIA GENÉTICA PARA ANÁLISE DE SEQUÊNCIAS DE DNA

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

Área de Concentração: Biometria e Estatística Aplicada

Orientador: Prof. Dr. Wilson Rosa de Oliveira Junior

Co-orientador: Prof. Dr. Kleber Régis Santoro

RECIFE-PE – JULHO/2013.

Ficha catalográfica

P372c Pedrosa, Lucemberg de Araújo
Comparação entre distância entrópica e distância genética para análise de sequências de DNA / Lucemberg de Araújo Pedrosa. – Recife, 2013.
94 f. : il.

Orientador: Wilson Rosa de Oliveira Junior.
. Dissertação (Mestrado em Biometria e Estatística Aplicada) – Universidade Federal Rural de Pernambuco, Departamento de Biometria e Estatística Aplicada, Recife, 2013.
Inclui referências e apêndice(s).

1. Alinhamento 2. Correlação 3. Distância genética 4. Entropia 5. Sequências genéticas I. De Oliveira, Wilson Rocha, orientador II. Título

CDD 310

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

COMPARAÇÃO ENTRE DISTÂNCIA ENTRÓPICA E DISTÂNCIA GENÉTICA PARA ANÁLISE DE SEQUÊNCIAS DE DNA

Lucemberg de Araújo Pedrosa

Dissertação julgada adequada para obtenção do título de mestre em Biometria e Estatística Aplicada, defendida e aprovada por unanimidade em 29/07/2013 pela Comissão Examinadora.

Orientador:

_____

Prof. Dr. Wilson Rosa de Oliveira Junior
Universidade Federal Rural de Pernambuco

Banca Examinadora:

_____

Profª. Dra. Teresa Bernarda Ludermir – membro externo
Universidade Federal de Pernambuco

_____

Prof. Dr. Kleber Régis Santoro – membro interno
Universidade Federal Rural de Pernambuco

_____

Profª. Dra. Tatijana Stosic – membro interno
Universidade Federal Rural de Pernambuco

Dedicatória

Dedico este trabalho a meus pais, minha vó “Inha”, meu irmão Léo e minha noiva Nayza.

Agradecimentos

Agradeço primeiramente a Deus, pois sem Ele nada seria possível, e nos momentos de angústia era a quem eu recorria e tenho a certeza de que Ele sempre esteve comigo.

A minha vó “Inha” pelas sábias palavras e a preocupação que sempre teve comigo, sempre foi o meu porto seguro.

A minha mãe Luzia pelos ensinamentos desde criança, principalmente pela ética e a garra. Ao meu pai Laércio que sempre torceu por mim em cada batalha.

Ao meu irmão Léo, que apesar se não sermos gêmeos, sempre fomos muito parecidos em questão de caráter, nas muitas vezes da solidão da madrugada, era o único que eu encontrava acordado para conversar e relaxar um pouco.

A minha noiva Nayza, que sempre compreendeu meus momentos de ausência devido o momento ao qual eu passava, pelo companheirismo e pelas críticas feitas a esse trabalho.

Ao meu orientador Wilson pelos ensinamentos, paciência e amizade nessa caminhada.

Ao meu co-orientador Kleber pela valiosa contribuição na construção deste trabalho e amizade.

Ao programa de Mestrado em Biometria e Estatística Aplicada.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, pela bolsa concedida durante os anos do curso.

A todos os meus colegas da Pós-Graduação, em especial aos que me acompanham desde a graduação.

A todos que de uma maneira ou de outra contribuíram para a conclusão desse trabalho.

“As nuvens mudam sempre de posição, mas são sempre as nuvens no céu. Assim devemos ser todo dia, mutantes, porém leais com o que pensamos e sonhamos; lembre-se, tudo se desmancha no ar, menos os pensamentos.”

Paulo Beleki

Resumo

Diante da grande quantidade de dados que é gerada na área da genética molecular, sente-se cada vez mais a necessidade de novos métodos para análise dos dados de maneira rápida e precisa. Assim, no presente estudo abordamos o conceito de entropia para análise de sequências de DNA, onde utiliza-se a distância entrópica. Para verificar a sua eficiência foi realizado o teste de Mantel, que faz a correlação entre as duas matrizes de distâncias, a distância entrópica e a distância genética, na qual foi utilizada a distância de Jukes – Cantor (JC69). Foram analisadas nove tipos de sequências genéticas, primeiramente o gene BoLA, onde foi realizado um estudo mais detalhado, mostrando uma estatística descritiva e realizando os dendogramas. Posteriormente analisamos sequências de tamanho maiores, que foram as abelhas sem ferrão *Melipona quinquefasciata*, e em seguida analisamos sequências muito maiores, o cromossomo 5 do *Homo Sapiens*. Foram realizadas também 6 simulações para verificar o comportamento dos resultados, em cada uma dessas nove análises foi realizado o alinhamento antes da distância de JC69. Foi possível obter bons resultados quando utilizou-se entropia condicional, onde foi introduzido o conceito de entropia em blocos, o método mostrou ser mais eficiente em sequências de grande comprimento.

Palavras-chave: Alinhamento, correlação, distância genética, entropia, sequências genéticas.

Abstract

Due to the large amount of data that is generated in the area of molecular genetics the need of new fast and precise methods for data analysis is increasingly needed. In the present study we discuss the concept of entropy to analyse DNA sequences, where the notion of entropic distance is defined. In order to check its efficiency the Mantel test was performed, which makes the correlation between the two distances matrices, the entropic distance and the genetic distance in which the Jukes - Cantor (JC69) distance. Nine types of genetic sequences were analysed, first the gene BoLA, where was conducted a more detailed study, showing a descriptive statistics and performing the dendograms. Subsequently sequences of larger size were analyzed, the stingless bees *Melipona quinquefasciata*, and finaly a much larger sequence, the chromosome 5 of Homo Sapiens. 6 simulations were also conducted to verify the behavior of results, in each of these three analysis the alignment was performed before the JC69 distance. It was possible to obtain good results when used the conditional entropy, which was introduced the concept of entropy in blocks, the method proved to be more effective in very long sequences.

Keywords: Alignment, correlation, genetic distance, entropy, genetic sequences.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

As informações genéticas são armazenadas nos ácidos nucléicos – o ácido desoxirribonucleico (DNA) e o ácido ribonucléico (RNA). O DNA é uma molécula que existe dentro das células de todos os seres vivos. O DNA é encontrado mais especificamente nos cromossomos. O RNA, por sua vez é encontrado principalmente no citoplasma, havendo muito pouco nos cromossomos. Na composição do DNA entram quatro bases nitrogenadas chamadas de *nucleotídeos*, ou simplesmente, de *bases*: adenina (A), guanina (G), timina (T) e citosina (C). O RNA por sua vez contém uracila (U) em vez da timina (T). A molécula de DNA tem uma estrutura semelhante a de uma escada torcida, formando um espiral. Os nucleotídeos formam os degraus, estando a adenina emparelhada com a timina e a guanina com a citosina constituindo uma dupla sequência de bases como mostra a figura 1. As sequências de DNA podem ser representadas através de longas cadeias de letras A, C, G, T. A localização física do nucleotídeo na sequência de DNA é denominada sítio (WATSON, 1992; GRIFFITHS, 2000).

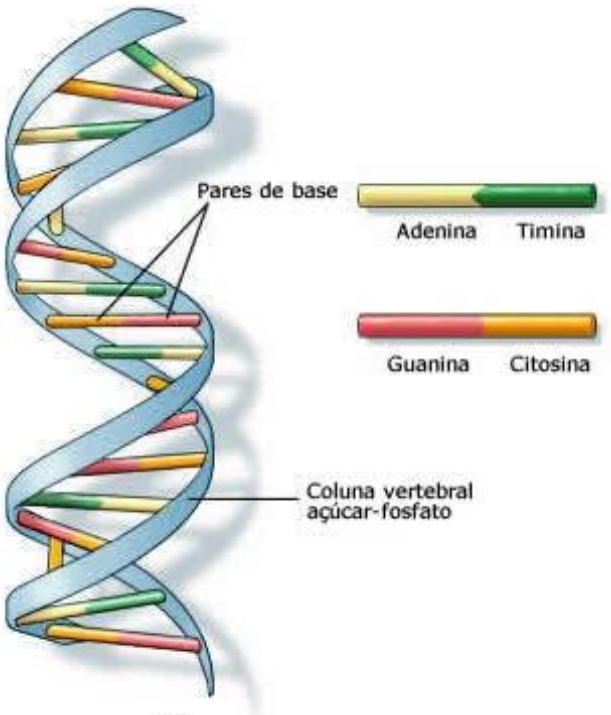

Figura 1: Estrutura do DNA
Fonte: HowStuffWorks – Como tudo funciona UOL
Disponível em: <http://saude.hsw.uol.com.br/dna1.htm>. Acesso em: 03 jul. 2013

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 cromossomo 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 a área de estatística e de bioinformática para análise desses dados (WATERMAN, 1995).

A bioinformática surgiu à partir da biologia molecular no momento em que se iniciou a utilização de ferramentas computacionais para a análise de dados genéticos. Ela envolve a união de diversas áreas do conhecimento: a computação, a matemática, a estatística e a biologia molecular, e tem como finalidade principal analisar a grande quantidade de dados que vem sendo obtidos através de sequência de DNA e proteínas. Como o volume de informações biológicas e sequencias genômicas se multiplicam a um ritmo muito elevado devido as modernas técnicas de sequenciação, torna-se cada vez mais importante o desenvolvimento de algoritmos que permitam compactar esse volume de informação de forma a otimizar o armazenamento e processamento das informações.

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, como por exemplo, o GenBank que é um banco de dados mundial que contém sequencias de DNA disponíveis de mais de 140.000 organismos diferentes (IDALINO, 2010).

Essas sequências são obtidas principalmente através de submissões de dados de sequência de laboratórios e submissões em lotes de projetos de sequenciamento de grande escala. O GenBank é mantido pelo “National Center for Biotechnology Information” (NCBI). O último registro disponível mostra que até o ano de 2012 existiam mais de 140 bilhões de pares de bases e aproximadamente 160 milhões de sequências presentes no site do NCBI disponível em <http://www.ncbi.nlm.nih.gov>.

Esse trabalho tem como objetivo encontrar boas correlações entre as distâncias por entropias e as distância genéticas através de uma detecção de similaridade por uma metodologia de custo computacional mais baixo, especialmente quando os objetivos são cadeias de grande comprimento. As distâncias genéticas utilizam o

alinhamento das sequências, estas dependem da quantidade de sequências que serão alinhadas (analisadas), o alinhamento é feito par a par.

Atualmente existem cada vez mais sequências de DNA de tamanhos maiores para serem analisadas e a quantidade de sequências que se está analisando faz com que computacionalmente tenha-se um processo muito lento. Por outro lado**,** tem-se a entropia, que informa o grau de desordem de um sistema. A vantagem de utilizá-la é que cada sequência terá sua PRÓPRIA entropia, ou seja, para calculá-la precisa-se unicamente de uma sequência. O custo computacional será basicamente o mesmo para todas as sequências que se deseja analisar, independentemente da quantidade de sequências que se está analisando. A desvantagem é que, ao contrário das distâncias genéticas que considera a ordem ou posição de como as coisas acontecem em uma determinada sequência, a entropia considera a frequência (quantidade) como um todo. A grande motivação é tentar encontrar alguma entropia que substitua a metodologia tradicional das distâncias genéticas. Para isso é necessário que seja levado em consideração algum tipo de ordem ou posição de como as coisas estão acontecendo na sequência. Esse é o ponto principal para se obter uma boa correlação entre a distância por entropia e a distância genética.

2. REVISÃO DA LITERATURA

Morgan, entre 1909 e início da década de 40, conseguiu desvendar o mistério da localização dos genes nos cromossomos, estabelecendo que eles eram responsáveis pela hereditariedade. Usando a frequência estatística de determinados acontecimentos foram deduzidas algumas distâncias entre genes, mesmo não conhecendo a sua composição e natureza. Assim, teve início a cartografia genética que tenta mostrar a disposição dos genes ao longo dos cromossomos.

Em 1944, Avery associa o ácido desoxirribonucleico (DNA) à hereditariedade e desvenda a sua natureza. Em 1953, Watson e Crick estabeleceram a estrutura da molécula de DNA, formada por duas cadeias enroladas em hélice, constituída por sequências de nucleotídeos. Como reconhecimento desse feito eles receberam o prêmio Nobel de Fisiologia/Medicina em 1962.

A biologia molecular permite medições em termos de pares de bases (bp) e muitas pesquisas até hoje se concentram para decodificar o código genético com o

alfabeto de quatro letras A, C, G, T. O código genético do homem tem cerca de três bilhões de letras. Jones (1995) refere-se que *"A genética é, em si mesma, uma linguagem, um conjunto de instruções herdadas, transmitidas de geração em geração. Tem um vocabulário (os próprios genes), uma gramática (a forma como a informação herdada está disposta) e uma literatura (as milhares de instruções para compor o ser humano)"*.

Com a descoberta do DNA começou o desenvolvimento de técnicas para a sua análise, principalmente a partir dos anos 60. Hoje em dia a genética está presente em várias áreas do conhecimento, como por exemplo: medicina, agricultura, farmácia, história, identificação de paternidade e resolução de alguns crimes.

São inúmeros os artigos publicados em revistas relacionadas à genética ou probabilidade e estatística. Por exemplo: *Science, Genomics, American Journal of Human Genetics*, *Biometrics, Statistical Science, The Annals of Applied Probability* e *Journal of the American Statistical Association*.

Em meados da década de 80 houve uma grande revolução na área da genética médica quando foram mapeados mais de 1.000 genes que causam doenças no homem. O sequenciamento 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) (WATERMAN, 1995).

Na evolução das espécies, ao longo do tempo, os modelos Markovianos têm sido utilizados para descrever alterações ocorridas no DNA que podem dar origem a espécies diferentes (Kelly, 1994). Quanto mais próximas estiverem duas espécies, mais semelhante é o seu DNA, pelo que na evolução usam-se algumas medidas descritas, por exemplo, em Weir (1990) para avaliar essa proximidade ou distanciamento a partir de um ancestral comum.

Um grupo de pesquisadores das universidades de Málaga (Espanha) e Harvard (Estados Unidos) descobriu que o DNA contém estruturas maiores, muito além das letras individuais dos nucleotídeos (CARPENA et al., 2011). Eles procuraram por regiões ricas em GC e pobres em GC nas sequências dos cromossomos humanos, utilizando um algoritmo baseado no conceito de entropia. Carpena e seus colegas dividiram a sequência de nucleotídeos em segmentos que maximizam a diferença de

entropia entre segmentos individuais e a sequência como um todo. Eles descobriram que cada cromossomo humano é dividido em grandes segmentos com dezenas de milhões de nucleotídeos de comprimento, mais do que qualquer estrutura organizacional previamente conhecida no genoma. Esses segmentos, que eles batizaram de "superestruturas", têm cerca de duas centenas de genes em média (CARPENA et al., 2011).

3. MATERIAL E MÉTODOS

3.1 DADOS

No presente trabalho utilizou-se 9 tipos de dados. Primeiramente foram as sequências do gene BoLA que foi o objeto inicial e principal da pesquisa. Após as análises do gene BoLA, procurou-se um maior embasamento das análises em sequências de tamanhos maiores, utilizando-se, então, sequências de abelhas sem ferrão *Melipona quinquefasciata* das regiões 18s e sequências do cromossomo 5 do *Homo Sapiens*. Posteriormente foram realizadas 6 simulações de modo a verificar os resultados considerando sequências aleatórias.

3.1.1 GENE BOLA

As sequências utilizadas são provenientes de rebanhos controlados no estado de Pernambuco, localizados na bacia leiteira do Estado, sendo os rebanhos da Estação Experimental do IPA em São Bento do Una e Estação Experimental de Arcoverde, respectivamente, totalizando 145 animais. Os dados fazem parte da dissertação de Luciana Florêncio Vilaça (VILAÇA, 2012), orientada de Kleber Régis Santoro no Programa de Pós-Graduação em Ciência Animal e Pastagens (PPGCAP) da Unidade Acadêmica de Garanhuns (UAG/UFRPE), que tem como parceiro fundamental o Instituto Agronômico de Pernambuco (IPA). Os dados são em relação ao gene BoLA – DRB3.2.

O gene BoLA-DRB3, pertencente a família dos genes BoLA (Bovine Lymphocyte Antigen), que se localizam no Complexo Principal de Histocompatibilidade (MHC – Major Histocompatibility Complex) do genoma bovino,

está envolvido no processo molecular de resistência e susceptibilidade à mastite destes animais. Eles são responsáveis por codificar as proteínas presentes na superfície das células e envolvidas na relação entre antígenos e anticorpos. Ele é altamente polimórfico.

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, que é a inflamação da glândula mamária pode ser causada por muitos fatores, sendo as bactérias os principais agentes infecciosos. 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 sensíveis de inflamação da glândula mamária. (BRANT & FIGUEIREDO, 1994)

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 consequências mais sérias estão relacionadas com as perdas econômicas acarretadas 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).

3.1.2 *MELIPONA QUINQUEFASCIATA*

As sequências das regiões 18s das abelhas sem ferrão *Melipona quinquefasciata* fazem parte da dissertação de Patrícia Silva do Nascimento Barros

(BARROS, 2011) orientada de Wilson Rosa de Oliveira Junior no Programa de Pós-Graduação em Biometria e Estatística Aplicada (PPGBEA) da UFRPE, onde foram utilizadas 6 sequências obtidas de várias colônias silvestres, em localidades distintas da Chapada do Araripe – CE, Chapada da Ibiapaba – CE, cidade do Canto do Buriti – PI e Luziânia – GO.

A distribuição geográfica de *Melipona quinquefasciata* mostra sua ocorrência apenas em estados do sul do Brasil, do Sul do Espírito Santo ao Rio Grande do Sul, incluindo áreas de Minas Gerais, Goiás, Mato Grosso, Mato Grosso do Sul, Bolívia (parte do Sul); Paraguai e Norte-Nordeste de Argentina (MARIANO-FILHO, 1911; SCHWARZ, 1932; KERR, 1948; MOURE, 1948, 1975; VIANA,1987).

Esta espécie de Melipona possui a particular característica de nidificar sob o solo, podendo suas câmaras chegar a uma profundidade de até 4 metros (KERR et.al, 2001; NOGUEIRA-NETO, 1997), utilizando-se de formigueiros ou cupinzeiros abandonados. O mel dessa abelha é bastante saboroso e embora de difícil coleta tem comércio garantido, sendo então cobiçado pelos meleiros do local. A extração do mel leva ainda à morte da colônia, já que nesta época a rainha está em seu período fértil estando com sobrepeso, não permitindo seu voo de fuga para uma futura formação de colônia como acontece com outras espécies (KERR et.al, 2001). Possui um comprimento de 9 a 10,5mm, com cinco faixas de coloração amarelada no abdômen. Infelizmente, a M. quinquefasciata já consta em listas de espécies em extinção, como mencionado no Livro Vermelho da Fauna Ameaçada no Estado do Paraná <http://www.maternatura.org.br>.

3.1.3 CROMOSSOMO 5

O cromossomo 5 é um dos 23 pares de cromossomos do genótipo humano e é um dos maiores cromossomos, representando quase 6% do total de DNA nas células. As pessoas, tem duas cópias deste cromossomo. Foram utilizadas 10 sequências do cromossomo 5 do Homo Sapiens obtidas do NCBI (NCBI 2013).

3.1.4 SIMULAÇÕES

Foram realizados 6 tipos de simulações, sendo duas simulações para cada 3 diferentes tamanho de sequências, em cada simulação foram geradas 10 sequências aleatórias. O tamanho das sequências realizadas foi de 400, 2.000 e 100.000 caracteres respectivamente, contendo as letras A, C, G, T. No 1º tipo de simulação considerou-se a probabilidade de ocorrência de cada base igual, ou seja, 0,25 para cada uma. No 2º tipo de simulação considerou-se a probabilidade 0,4; 0,3; 0,2; 0,1 para as bases A, C, G, T respectivamente.

3.2 ALINHAMENTO DE SEQUÊNCIAS (CLUSTALW)

O alinhamento de sequências consiste no processo de comparar duas ou mais sequências de nucleotídeos de forma a se observar seu nível de similaridade. Após realizar o alinhamento é possível localizar trechos conservados, que são regiões em que todas as sequências apresentam a mesma base em uma determinada posição, como também comparar uma sequência desconhecida com bancos de dados de sequências conhecidas. As posições dos nucleotídeos são ajustadas, se necessário, usando espaços (gaps), porém não pode no alinhamento ter uma posição em que todas as sequências alinhadas tenham gaps. O 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.

A sequência consenso refere-se à região de consenso em sequências alinhadas de forma a maximizar suas homologias. Para que uma sequência seja aceita como consenso, cada base individual deve ser predominante na respectiva posição. Estando alinhadas n sequências, ao final desse alinhamento será identificada uma nova sequência com as similaridades, polimorfismo e os gaps. A sequência consenso é de fundamental importância para identificar indivíduos que sofrem mutações numa dada característica num determinado gene.

Em alinhamentos locais, segmentos das sequencias 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 sequencias que são similares ao longo de um trecho de suas sequencias e mais dissimilares em outros trechos, ou então, sequencias que diferem muito no tamanho ou que compartilham uma região ou domínio conservado.

O alinhamento ClustalW é um método de alinhamento múltiplo global, ou seja, a similaridade é considerada ao longo de todas as sequências, e segue uma abordagem progressiva; começa com as sequências mais similares organizando a partir destes, uma árvore filogenética. O alinhamento global tem como objetivo tentar alinhar as sequências completamente usando o máximo de caracteres possíveis em toda a extensão.

Para escolher o melhor alinhamento é utilizada a técnica do algoritmo de programação dinâmica, onde é criado um sistema de pontuação (*score*), os critérios são *matches* que corresponde às bases iguais, *mismatches* são as bases diferentes e os *gaps* que representam as inserções entre as sequências. Esse valor de pontuação é computado com o intuito de penalizar as diferenças das sequências e privilegiar as similaridades. O melhor alinhamento é aquele que maximiza o score. O alinhamento de 2 sequências é representado como uma matriz de 2 linhas por C colunas, onde C é o tamanho das sequências após o alinhamento, similarmente, um alinhamento de 3 sequências é representado como uma matriz de 3 linhas por C colunas.

O alinhamento ClustalW é feito em 3 passos. No 1º passo é feito o alinhamento par a par para todos os possíveis conjuntos de sequências e a distância do alinhamento de cada par é obtida através do algoritmo de programação dinâmica. Considerando um conjunto de n sequências, o algoritmo de programação dinâmica será realizado $n(n-1)/2$ vezes, essa é a fase mais crítica em relação ao tempo de processamento. No 2º passo é construída uma árvore guia que conduzirá o alinhamento múltiplo. O 3º passo é o alinhamento progressivo, onde se começa pelas duas sequências mais similares encontradas no passo 1 e em seguida é utilizada a árvore guia do passo 2 que vai adicionando as sequências e alinhando-a ao alinhamento já existente. Para um conjunto de n sequências, são necessários $n-1$ alinhamentos nessa última etapa, ou seja, o algoritmo de programação dinâmica é realizado $n-1$ vezes.

O alinhamento múltiplo passa a ser um problema quando a quantidade de sequências aumenta significativamente, ainda mais considerando que o volume de dados a serem analisados aumenta diariamente. A fase mais crítica do alinhamento múltiplo progressivo é o alinhamento par a par, pois demanda grande custo computacional.

Seja $s_1, s_2, \dots, s_k$ sequências de DNA, um alinhamento de $s_1, s_2, \dots, s_k$ é uma matriz $A = A_{ij}$ de dimensão $k\ X\ n$. Dizemos que dois caracteres $s_i[j]$ e $s_{i'}[j']$ estão alinhados se ambos estão na mesma coluna da A. Onde s_i é a sequência e j é a posição do nucleotídeo na sequência.

3.3 DISTÂNCIA DE JUKES – CANTOR (JC69)

A distância genética é uma medida da diferença de material genético entre diferentes espécies ou indivíduos da mesma espécie. 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.

O modelo proposto em 1969 por Thomas H. Jukes e Charles R. Cantor (JUKES E CANTOR, 1969) assume que todas as bases tem igual probabilidade de ocorrência. Através deste modelo obtém-se a distância genética entre duas sequências.

$$d_{ij} = -\frac{3}{4}\ln(1 - \frac{4}{3}p_{ij})$$

em que p_{ij} é a proporção de diferenças de nucleotídeos observada entre duas sequências i e j, e

$$p_{ij} = \frac{n_p}{n}$$

Em que n_p é a quantidade de nucleotídeos diferentes entres as duas sequências e n é a quantidade de nucleotídeos.

3.4 MÉTODO DE AGRUPAMENTO PELA MÉDIA DAS DISTÂNCIAS

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

Este método define a distância entre dois grupos como sendo a média das distâncias entre todos os pares de elementos, sendo um em cada grupo. Este procedimento pode ser utilizado tanto para medidas de similaridade como de distância, contanto que o conceito de uma medida média seja aceitável (Everitt, 1974). Os grupos são reunidos em um novo grupo quando a média das distâncias entre seus elementos é mínima.

No método das médias das distâncias se define a distância entre dois grupos, i e j, como sendo a média das distâncias entre todos os pares de objetos constituídos por elementos dos dois grupos. A estratégia é que o valor médio tem a vantagem de evitar valores extremos e de tomar em consideração toda a informação dos grupos. Um grupo passa a ser definido como um conjunto de indivíduos no qual cada um tem mais semelhanças, em média, com todos os membros do mesmo grupo do que com todos os elementos de qualquer outro grupo (Reis, 1997).

Algoritmo:

Passo 1. Determina-se a matriz de distâncias inicial;

Passo 2. Localizam-se os dois elementos que apresentam a menor distância, reunindo em um único grupo;

Passo 3. Calcula-se a distância entre os diversos pares de grupos como sendo a média das distâncias entre todos os pares de seus elementos, sendo um elemento de cada um dos grupos;

Passo 4. Os dois grupos que apresentam menor distância são reunidos em um único grupo.

Média das distâncias:

$$d^*k_{ij} = \left[\frac{n_i}{n_i + n_j}\right].d_{ki} + \left[\frac{n_j}{n_i + n_j}\right].d_{kj}$$

em que:

d^*k_{ij}, d_{ki}e d_{kj} são as distâncias entre os elementos de k e agrupamento ij, k e i, k e j, e i e j, respectivamente;

n_i e n_j são a quantidade de elementos nos agrupamentos i e j, respectivamente.

As medidas de dissimilaridade entre as observações podem ser várias, a mais comum delas é a distância euclidiana. A distância euclidiana entre dois elementos $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p}]$ e $\boldsymbol{Y} = [\boldsymbol{y_1}, \boldsymbol{y_2}, \dots, \boldsymbol{y_p}]$ é definido por:

$$d_{xy} = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

3.6 COEFICIENTE DE CORRELAÇÃO COFENÉTICA (CCC)

O coeficiente de correlação cofenética mede o grau de ajuste entre a matriz de dissimilaridade (Matriz Fenética F) e a matriz resultante da simplificação proporcionada pelo método de agrupamento (matriz cofenética C), (FARRIS, 1969). Quanto maior o valor obtido para o coeficiente de correlação, menor será a distorção provocada pelo agrupamento de todos os indivíduos. A correlação foi calculada conforme Bussab et al., (1990):

$$rcof = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\overline{c})(d_{ij}-\overline{d})}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\overline{c})^2}\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(d_{ij}-\overline{d})^2}}$$, em que;

c_{ij}= valor de dissimilaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética;

d_{ij}= valor de dissimilaridade entre os indivíduos i e j, obtidos a partir da matriz de dissimilaridade;

$$\overline{c} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_{ij}$$

$$\overline{d} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}$$

Nota-se que essa correlação equivale à correlação de Pearson entre a matriz de dissimilaridade original e aquela obtida após a construção do dendrograma. Assim, quanto mais próximo de 1, menor será a distorção provocada pelo agrupamento dos indivíduos com os métodos.

Para Bussab et al., (1990), o desafio é responder se o valor observado é alto ou baixo. Responder a isto é tão difícil como responder, na maioria das situações, o que é um alto coeficiente de correlação entre duas variáveis. Depende da área de estudo e de padrões que vão se desenvolvendo com a prática. Pode-se adiantar que em análise de agrupamento, algo em torno de 0,8 já pode ser considerado um bom ajuste.

3.6 ENTROPIA

O conceito de entropia na teoria da informação tem origem no trabalho de Shannon (1948), onde mostra que processos aleatórios tais como a fala ou a música tem uma complexidade abaixo da qual o sinal não pode ser comprimido. A esta complexidade ele chamou entropia, nome derivado da física, em particular da física estatística, onde era usado como medida do estado de desordem de um sistema. Entropia tem a ver com a probabilidade das possibilidades. Se o sistema está muito ordenado, a entropia é zero, à medida que o sistema se desordena, a entropia vai atingir o seu valor máximo.

3.6.1 ENTROPIA DE SHANNON

Suponhamos que x é uma variável aleatória discreta de X e distribuição de probabilidade $p(x) = \Pr(X = x), x \in X$. A entropia $H(X)$ de uma v.a discreta X é definida por (SHANNON, 1948) :

$$H_S(X) = -\sum_{x \in X} p(x) log_2 p(x)$$

3.6.2 ENTROPIA DE RÉNYI

Rényi (1976) propôs uma definição mais geral de entropia na qual a entropia de Shannon aparece como um caso particular. Para uma v.a discreta a entropia de Rényi é dada por:

$$H_R(X) = \frac{1}{1-a} log_2 \sum_{x \in X} p(x)^a$$

onde a é um parâmetro livre denominado ordem.

3.6.3 ENTROPIA DE TSALLIS

A entropia de Tsallis é uma generalização da entropia de Boltzmann-Gibbs. Ela foi formulada por Constatino Tsallis em 1988 e é definida como:

$$H_T(X) = \frac{1}{q-1}(1 - \sum_{x \in X} p^q(x))$$

onde q é um parâmetro real.

3.6.4 ENTROPIA CONDICIONAL

A entropia conjunta $\boldsymbol{H(X,Y)}$ das variáveis aleatórias $\boldsymbol{X}$ e $\boldsymbol{Y}$ com distribuição de probabilidade conjunta $\boldsymbol{p(x,y)}$ é definida como:

$$H(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log\ p(x,y)$$

A entropia de uma variável aleatória **X**, condicionada pela presença (ou conhecimento) de outra variável **Y**, mede a incerteza de **X** quando **Y** é conhecida. Se condicionarmos um valor específico $Y = y$, as probabilidades condicionais $\{p(x|y), x \in X\}$ podem ser usadas na definição original, pois verificam 0 ≤ $p(x|y)$ ≤1 e

$$\sum_{x \in X} p(x|y) = 1,$$

qualquer que seja $y \in Y$. Surge assim a entropia (incerteza) de X, condicionada a que $Y = y$, dado por

$$H(X|Y = y) = -\sum_{x \in X} p(x|y) log p(x|y)$$

3.7 TESTE DE MANTEL

O teste de Mantel (MANTEL, 1967) tem como objetivo comparar duas matrizes de **semelhança** (distância ou similaridade) derivada de dois conjuntos de dados multidimensionais. A estatística do ***r*** de Mantel varia de -1 a +1. Quanto maior a correlação entre as matrizes de distância, maior o valor de *r.*

Seja uma variável observada em n localizações, obtêm-se duas matrizes A e B, de dimensão $n\ x\ n$, simétricas, cujos elementos representam distâncias, em alguma métrica, entre as observações.

$$A = \begin{bmatrix} 0 & a_{21} & \dots & a_{n1} \\ a_{21} & 0 & \dots & a_{n2} \\ \vdots & \vdots & 0 & \vdots \\ a_{n1} & a_{n2} & \dots & 0 \end{bmatrix} \text{ e } B = \begin{bmatrix} 0 & b_{21} & \dots & b_{n1} \\ b_{21} & 0 & \dots & b_{n2} \\ \vdots & \vdots & 0 & \vdots \\ b_{n1} & b_{n2} & \dots & 0 \end{bmatrix}$$

A estatística do teste é dada pelo coeficiente de correlação de Pearson entre os elementos correspondentes de A e B, isto é,

$$r = \frac{m \sum_{i<j} a_{ij} b_{ij} - \sum_{i<j} a_{ij} \sum_{i<j} b_{ij}}{\sqrt{\left[m \sum_{i<j} a_{ij}^2 - \left(\sum_{i<j} a_{ij}\right)^2\right]\left[m \sum_{i<j} b_{ij}^2 - \left(\sum_{i<j} b_{ij}\right)^2\right]}}$$

Que produz o valor r_o quando calculada para os valores observados. A seguir permutam-se várias vezes linhas e colunas de uma das matrizes, e obtêm-se os valores da estatística dos dados aleatorizados r_{Ai}. A proporção *p de valores* $r_{Ai} \geq r_o$ é comparada com um valor de nível de significância pré-fixado α e rejeita-se a hipótese nula se $p < \alpha$. A hipótese nula é de que não existe relação linear entre correspondentes das matrizes de distância.

Sendo as matrizes A e B simétricas, a correlação entre todos os elementos fora da diagonal principal é a mesma que a correlação de $m = \frac{n(n-1)}{2}$ elementos na parte triangular superior ou inferior da matriz. Note-se ainda que o único termo que é alterado no valor de **r** pela mudança da ordem dos elementos em uma das matrizes A ou B é a soma dos produtos $\sum_{i<j} a_{ij} b_{ij}$.

4. RESULTADOS E DISCUSSÕES

4.1 ANÁLISE GENÉTICA

Foram analisadas 145 sequências através do software DAMBE 5.3.2 (Data Analysis in Molecular Biology and Evolution) e MEGA 5.05 (Molecular Evolutionary Genetics Analysis) utilizando a ferramenta ClustalW e também o software R 2.15.1

Com o alinhamento foi possível observar as regiões onde os nucleotídeos se repetiam e desta forma foi gerada uma sequência padrão, denominada de sequência consenso. Após a construção e análise dos alinhamentos foi feito um estudo descritivo da composição de cada sequência. Em seguida foram calculadas as distâncias genéticas e realizados os agrupamentos.

Pelo alinhamento podemos observar uma maior concentração de regiões similares nas posições 117 até 173 sendo a maioria o nucleotídeo G.

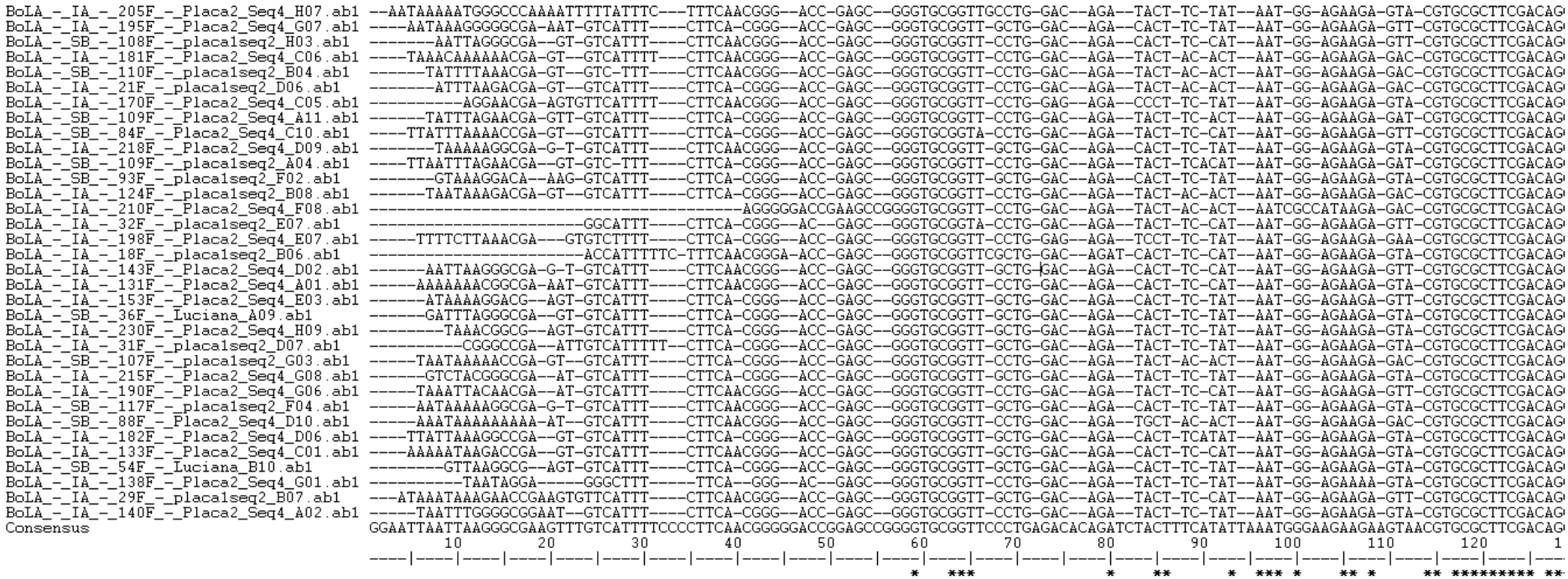

Figura 2: Alinhamento de parte das sequências da posição 1 até 128

Observa-se nas tabelas 1 a 5 que a média de cada base não se aproxima da quantidade de cada base da sequência consenso, ao contrário das probabilidades de cada base em que os valores são bem semelhantes.

Tabela 1: Quantidade e frequência de cada base (A, C, G, T) em cada sequência

Sequências	A	C	G	T	Total	PA	PC	PG	PT
1-BoLA_-_IA_-_200F	70	59	98	46	273	0,26	0,22	0,36	0,17
2-BoLA_-_IA_-_155F	66	60	100	51	277	0,24	0,22	0,36	0,18
3-BoLA_-_IA_-_145F	61	61	101	49	272	0,22	0,22	0,37	0,18
4-BoLA_-_SB_-_61F	59	64	99	48	270	0,22	0,24	0,37	0,18
5-BoLA_-_SB_-_51F	64	59	104	44	271	0,24	0,22	0,38	0,16
6-BoLA_-_IA_-_30F	60	62	96	52	270	0,22	0,23	0,36	0,19
7-BoLA_-_SB_-_102F	61	62	105	46	274	0,22	0,23	0,38	0,17
8-BoLA_-_SB_-_98F	62	63	97	52	274	0,23	0,23	0,35	0,19
9-BoLA_-_IA_-_06F	59	62	100	45	266	0,22	0,23	0,38	0,17
10-BoLA_-_IA_-_161F	64	61	103	46	274	0,23	0,22	0,38	0,17
11-BoLA_-_SB_-_75F	61	63	96	51	271	0,23	0,23	0,35	0,19
12-BoLA_-_IA_-_127F	58	64	103	46	271	0,21	0,24	0,38	0,17
13-BoLA_-_IA_-_219F	60	62	96	52	270	0,22	0,23	0,36	0,19
14-BoLA_-_SB_-_118F	65	63	96	50	274	0,24	0,23	0,35	0,18
15-BoLA_-_IA_-_164F	62	67	91	50	270	0,23	0,25	0,34	0,19
16-BoLA_-_SB_-_63F	62	58	101	49	270	0,23	0,21	0,37	0,18
17-BoLA_-_IA_-_122F	60	64	97	47	268	0,22	0,24	0,36	0,18
18-BoLA_-_IA_-_05F	57	66	96	45	264	0,22	0,25	0,36	0,17
19-BoLA_-_IA_-_01F	57	51	93	35	236	0,24	0,22	0,39	0,15
20-BoLA_-_SB_-_66F	59	60	100	47	266	0,22	0,23	0,38	0,18
21-BoLA_-_IA_-_148F	62	60	101	47	270	0,23	0,22	0,37	0,17
22-BoLA_-_SB_-_100F	60	63	102	47	272	0,22	0,23	0,38	0,17
23-BoLA_-_IA_-_17F	61	59	102	48	270	0,23	0,22	0,38	0,18
24-BoLA_-_IA_-_142F	64	59	102	47	272	0,24	0,22	0,38	0,17
25-BoLA_-_IA_-_183F	64	65	94	49	272	0,24	0,24	0,35	0,18
26-BoLA_-_IA_-_15F	60	65	96	47	268	0,22	0,24	0,36	0,18
27-BoLA_-_IA_-_28F	58	66	98	45	267	0,22	0,25	0,37	0,17
28-BoLA_-_IA_-_173F	62	61	97	46	266	0,23	0,23	0,36	0,17
29-BoLA_-_SB_-_85F	63	62	98	43	266	0,24	0,23	0,37	0,16
30-BoLA_-_IA_-_144F	60	63	99	49	271	0,22	0,23	0,37	0,18
31-BoLA_-_SB_-_112F	63	59	101	46	269	0,23	0,22	0,38	0,17
32-BoLA_-_SB_-_52F	62	61	103	47	273	0,23	0,22	0,38	0,17
33-BoLA_-_SB_-_99F	62	70	100	45	277	0,22	0,25	0,36	0,16
34-BoLA_-_IA_-_203F	62	60	100	48	270	0,23	0,22	0,37	0,18
35-BoLA_-_IA_-_24F	56	61	101	47	265	0,21	0,23	0,38	0,18
36-BoLA_-_IA_-_121F	59	62	99	50	270	0,22	0,23	0,37	0,19
37-BoLA_-_IA_-_141F	64	62	97	50	273	0,23	0,23	0,36	0,18
38-BoLA_-_IA_-_151F	64	61	105	50	280	0,23	0,22	0,38	0,18
39-BoLA_-_IA_-_206F	55	60	101	51	267	0,21	0,22	0,38	0,19
40-BoLA_-_IA_-_202F	59	59	106	50	274	0,22	0,22	0,39	0,18
41-BoLA_-_IA_-_147F	65	57	101	48	271	0,24	0,21	0,37	0,18
42-BoLA_-_IA_-_191F	83	91	124	63	361	0,23	0,25	0,34	0,17
43-BoLA_-_IA_-_06F	64	60	103	48	275	0,23	0,22	0,37	0,17
44-BoLA_-_IA_-_10F	68	64	96	46	274	0,25	0,23	0,35	0,17

Tabela 2: Quantidade e frequência de cada base (A, C, G, T) em casa sequência (continuação)

Sequências	A	C	G	T	Total	PA	PC	PG	PT
45-BoLA_-_IA_-_128F	66	64	103	45	278	0,24	0,23	0,37	0,16
46-BoLA_-_SB_-_76F	68	58	99	46	271	0,25	0,21	0,37	0,17
47BoLA_-_IA_-_35F	57	60	95	39	251	0,23	0,24	0,38	0,16
48-BoLA_-_IA_-_188F	62	61	96	52	271	0,23	0,23	0,35	0,19
49-BoLA_-_SB_-_92F	67	66	99	50	282	0,24	0,23	0,35	0,18
50-BoLA_-_IA_-_224F	63	59	98	51	271	0,23	0,22	0,36	0,19
51-BoLA_-_SB_-_49F	65	60	103	46	274	0,24	0,22	0,38	0,17
52-BoLA_-_SB_-_101F	69	59	101	44	273	0,25	0,22	0,37	0,16
53-BoLA_-_IA_-_171F	62	61	97	49	269	0,23	0,23	0,36	0,18
54-BoLA_-_IA_-_222F	70	67	96	46	279	0,25	0,24	0,34	0,16
55-BoLA_-_SB_-_67F	74	88	129	61	352	0,21	0,25	0,37	0,17
56-BoLA_-_IA_-_20F	60	61	94	44	259	0,23	0,24	0,36	0,17
57-BoLA_-_SB_-_88F	63	60	105	43	271	0,23	0,22	0,39	0,16
58-BoLA_-_IA_-_152F	68	62	98	43	271	0,25	0,23	0,36	0,16
59-BoLA_-_IA_-_129F	60	62	100	50	272	0,22	0,23	0,37	0,18
60-BoLA_-_IA_-_167F	67	69	95	44	275	0,24	0,25	0,35	0,16
61-BoLA_-_IA_-_169F	65	62	102	48	277	0,23	0,22	0,37	0,17
62-BoLA_-_SB_-_60F	65	58	103	44	270	0,24	0,21	0,38	0,16
63-BoLA_-_SB_-_93F	59	62	101	47	269	0,22	0,23	0,38	0,17
64-BoLA_-_IA_-_136F	60	64	95	45	264	0,23	0,24	0,36	0,17
65-BoLA_-_IA_-_174F	61	64	97	48	270	0,23	0,24	0,36	0,18
66-BoLA_-_SB_-_42F	66	60	104	44	274	0,24	0,22	0,38	0,16
67-BoLA_-_IA_-_162F	67	67	103	51	288	0,23	0,23	0,36	0,18
68-BoLA_-_IA_-_177F	61	59	101	47	268	0,23	0,22	0,38	0,18
69-BoLA_-_IA_-_157F	60	64	98	49	271	0,22	0,24	0,36	0,18
70-BoLA_-_SB_-_98F	54	57	88	46	245	0,22	0,23	0,36	0,19
71-BoLA_-_SB_-_114F	8	77	138	69	392	0,28	0,20	0,35	0,18
72-BoLA_-_IA_-_178F	65	61	96	47	269	0,24	0,23	0,36	0,17
73-BoLA_-_SB_-_91F	67	63	96	45	271	0,25	0,23	0,35	0,17
74-BoLA_-_IA_-_125F	55	59	105	50	269	0,20	0,22	0,39	0,19
75-BoLA_-_IA_-_208F	58	63	100	49	270	0,21	0,23	0,37	0,18
76-BoLA_-_IA_-_179F	60	63	103	49	275	0,22	0,23	0,37	0,18
77-BoLA_-_IA_-_10F	66	62	95	50	273	0,24	0,23	0,35	0,18
78-BoLA_-_SB_-_102F	60	62	102	47	271	0,22	0,23	0,38	0,17
79-BoLA_-_IA_-_139F	59	63	102	46	270	0,22	0,23	0,38	0,17
80-BoLA_-_SB_-_80F	65	61	99	44	269	0,24	0,23	0,37	0,16
81-BoLA_-_IA_-_172F	81	86	131	62	360	0,23	0,24	0,36	0,17
82-BoLA_-_IA_-_34F	59	64	100	47	270	0,22	0,24	0,37	0,17
83-BoLA_-_SB_-_79F	63	61	98	49	271	0,23	0,23	0,36	0,18
84-BoLA_-_IA_-_132F	84	91	122	63	360	0,23	0,25	0,34	0,18
85-BoLA_-_IA_-_126F	55	62	101	45	263	0,21	0,24	0,38	0,17
86-BoLA_-_IA_-_14F	60	63	95	45	263	0,23	0,24	0,36	0,17
87-BoLA_-_IA_-_166F	66	62	96	48	272	0,24	0,23	0,35	0,18

Tabela 3: Quantidade e frequência de cada base (A, C, G, T) em casa sequência (continuação)

Sequências	A	C	G	T	Total	PA	PC	PG	PT
88-BoLA_-_IA_-_149F	60	62	98	51	271	0,22	0,23	0,36	0,19
89-BoLA_-_IA_-_11F	63	61	102	46	272	0,23	0,22	0,38	0,17
90-BoLA_-_SB_-_89F	60	63	99	48	270	0,22	0,23	0,37	0,18
91-BoLA_-_IA_-_214F	56	61	100	51	268	0,21	0,23	0,37	0,19
92-BoLA_-_IA_-_158F	66	60	97	49	272	0,24	0,22	0,36	0,18
93-BoLA_-_SB_-_59F	64	62	97	49	272	0,24	0,23	0,36	0,18
94-BoLA_-_IA_-_03F	58	61	95	47	261	0,22	0,23	0,36	0,18
95-BoLA_-_IA_-_194F	62	57	102	51	272	0,23	0,21	0,38	0,19
96-BoLA_-_SB_-_72F	93	89	135	65	382	0,24	0,23	0,35	0,17
97-BoLA_-_SB_-_95F	62	61	96	48	267	0,23	0,23	0,36	0,18
98-BoLA_-_IA_-_156F	58	64	99	48	269	0,22	0,24	0,37	0,18
99-BoLA_-_IA_-_160F	58	62	97	53	270	0,21	0,23	0,36	0,20
100-BoLA_-_IA_-_165F	73	79	120	77	349	0,21	0,23	0,34	0,22
101-BoLA_-_SB_-_43F	57	61	102	49	269	0,21	0,23	0,38	0,18
102-BoLA_-_IA_-_22F	57	60	102	50	269	0,21	0,22	0,38	0,19
103-BoLA_-_IA_-_25F	62	59	102	48	271	0,23	0,22	0,38	0,18
104-BoLA_-_SB_-_78F	63	66	100	46	275	0,23	0,24	0,36	0,17
105-BoLA_-_IA_-_13F	61	63	100	47	271	0,23	0,23	0,37	0,17
106-BoLA_-_SB_-_94F	57	60	101	42	260	0,22	0,23	0,39	0,16
107-BoLA_-_SB_-_99F	83	91	118	69	361	0,23	0,25	0,33	0,19
108-BoLA_-_IA_-_184F	64	64	101	45	274	0,23	0,23	0,37	0,16
109-BoLA_-_IA_-_150F	69	60	97	48	274	0,25	0,22	0,35	0,18
110-BoLA_-_SB_-_120F	5	80	137	71	393	0,27	0,20	0,35	0,18
111-BoLA_-_IA_-_209F	57	59	100	46	262	0,22	0,23	0,38	0,18
112-BoLA_-_IA_-_205F	68	60	101	51	280	0,24	0,21	0,36	0,18
113-BoLA_-_IA_-_195F	63	59	104	47	273	0,23	0,22	0,38	0,17
114-BoLA_-_SB_-_108F	59	63	100	49	271	0,22	0,23	0,37	0,18
115-BoLA_-_IA_-_181F	95	97	144	57	393	0,24	0,25	0,37	0,15
116-BoLA_-_SB_-_110F	65	60	99	47	271	0,24	0,22	0,37	0,17
117-BoLA_-_IA_-_21F	63	62	99	45	269	0,23	0,23	0,37	0,17
118-BoLA_-_IA_-_170F	75	78	132	62	347	0,22	0,22	0,38	0,18
119-BoLA_-_SB_-_109F	63	58	99	51	271	0,23	0,21	0,37	0,19
120-BoLA_-_SB_-_84F	63	64	95	50	272	0,23	0,24	0,35	0,18
121-BoLA_-_IA_-_218F	62	61	101	46	270	0,23	0,23	0,37	0,17
122-BoLA_-_SB_-_109F	64	57	100	51	272	0,24	0,21	0,37	0,19
123-BoLA_-_SB_-_93F	63	62	100	45	270	0,23	0,23	0,37	0,17
124-BoLA_-_IA_-_124F	66	60	100	44	270	0,24	0,22	0,37	0,16
125-BoLA_-_IA_-_210F	59	62	95	35	251	0,24	0,25	0,38	0,14
126-BoLA_-_IA_-_32F	57	60	95	43	255	0,22	0,24	0,37	0,17
127-BoLA_-_IA_-_198F	62	58	102	49	271	0,23	0,21	0,38	0,18
128-BoLA_-_IA_-_18F	56	64	98	45	263	0,21	0,24	0,37	0,17
129-BoLA_-_IA_-_143F	61	62	98	49	270	0,23	0,23	0,36	0,18
130-BoLA_-_IA_-_131F	66	64	99	44	273	0,24	0,23	0,36	0,16

Tabela 4: Quantidade e frequência de cada base (A, C, G, T) em casa sequência (continuação)

Sequências	A	C	G	T	Total	PA	PC	PG	PT
131-BoLA_-_IA_-_153F	62	61	99	47	269	0,23	0,23	0,37	0,17
132-BoLA_-_SB_-_36F	58	61	103	48	270	0,21	0,23	0,38	0,18
133-BoLA_-_IA_-_230F	60	61	102	45	268	0,22	0,23	0,38	0,17
134-BoLA_-_IA_-_31F	60	60	103	47	270	0,22	0,22	0,38	0,17
135-BoLA_-_SB_-_107F	90	83	117	57	347	0,26	0,24	0,34	0,16
136-BoLA_-_IA_-_215F	58	61	103	47	269	0,22	0,23	0,38	0,17
137-BoLA_-_IA_-_190F	63	65	92	51	271	0,23	0,24	0,34	0,19
138-BoLA_-_SB_-_117F	64	61	100	46	271	0,24	0,23	0,37	0,17
139-BoLA_-_SB_-_88F	71	60	97	44	272	0,26	0,22	0,36	0,16
140-BoLA_-_IA_-_182F	60	62	104	47	273	0,22	0,23	0,38	0,17
141-BoLA_-_IA_-_133F	64	67	94	48	273	0,23	0,25	0,34	0,18
142-BoLA_-_SB_-_54F	58	62	101	47	268	0,22	0,23	0,38	0,18
143-BoLA_-_IA_-_138F	63	56	99	43	261	0,24	0,21	0,38	0,16
144-BoLA_-_IA_-_29F	64	64	101	49	278	0,23	0,23	0,36	0,18
145-BoLA_-_IA_-_140F	63	59	103	47	272	0,23	0,22	0,38	0,17
Média	63,9	62,9	101,2	47,5	275,5	0,23	0,23	0,37	0,17

Tabela 5: Quantidade e frequência de cada base (A, C, G, T) na sequência consenso

Sequências	A	C	G	T	Total	PA	PC	PG	PT
Consenso	112	101	162	66	441	0,25	0,23	0,37	0,15

Após a análise descritiva das sequências, calculou-se as distâncias genéticas utilizando a distância de Jukers – Cantor (JC69). Pelo histograma da Figura 3, e tabela 6, podemos observar que os valores das distâncias genéticas se concentram em torno de 0.1 e registram uma média de 0.09568 e uma mediana de 0.09224, o que indica que os dados não são dispersos. Pelo Box plot da figura 4, visualizamos a presença de vários outliers, que são pontos atípicos que apresentam um grande afastamento dos demais dados, nesse caso podemos interpretar como duas sequências bem diferentes.

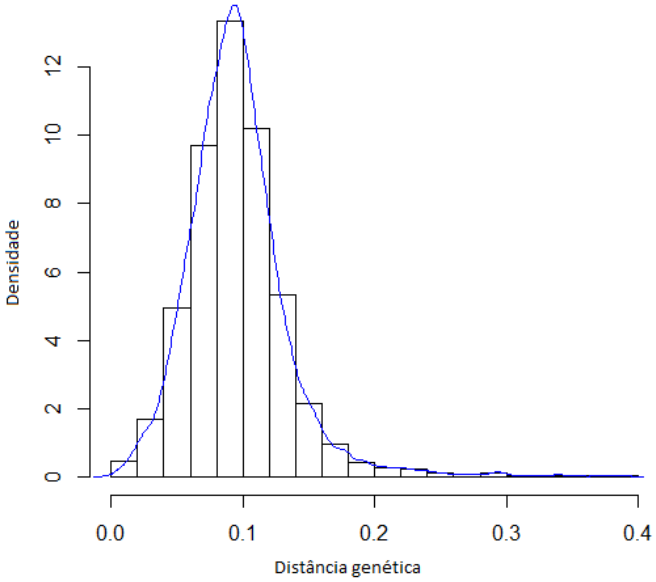

Figura 3: Histograma das distâncias genéticas e função densidade

Tabela 6: Análise descritiva da distância genética.

Distância genética	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
Estatísticas	0.00000	0.07260	0.09224	0.09568	0.11300	0.39160

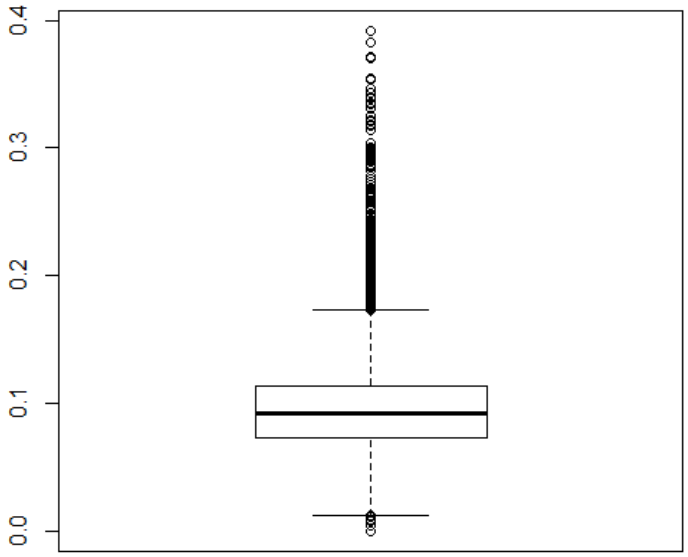

Figura 4: Box plot das distâncias genéticas

Foi realizado também um agrupamento pela média das distâncias com o intuito de identificar grupos com maiores similaridades. Pela figura 5 é possível destacar 5 grupos, sendo 2 grupos isolados; o primeiro contendo a sequência 141 e o segundo, a sequência 144 que definimos de grupo 1 e grupo 2, respectivamente. Um terceiro grupo foi formado pelas sequências 15, 36 e 37. O quarto grupo é o que contém a maior quantidade de sequências, sendo um total de 123. E finalmente o quinto, localizado no canto direito da figura 5 contendo 17 sequências. A correlação cofenética deste método de agrupamento foi de 71% o que indica que os dados não estão tão bem agrupados.

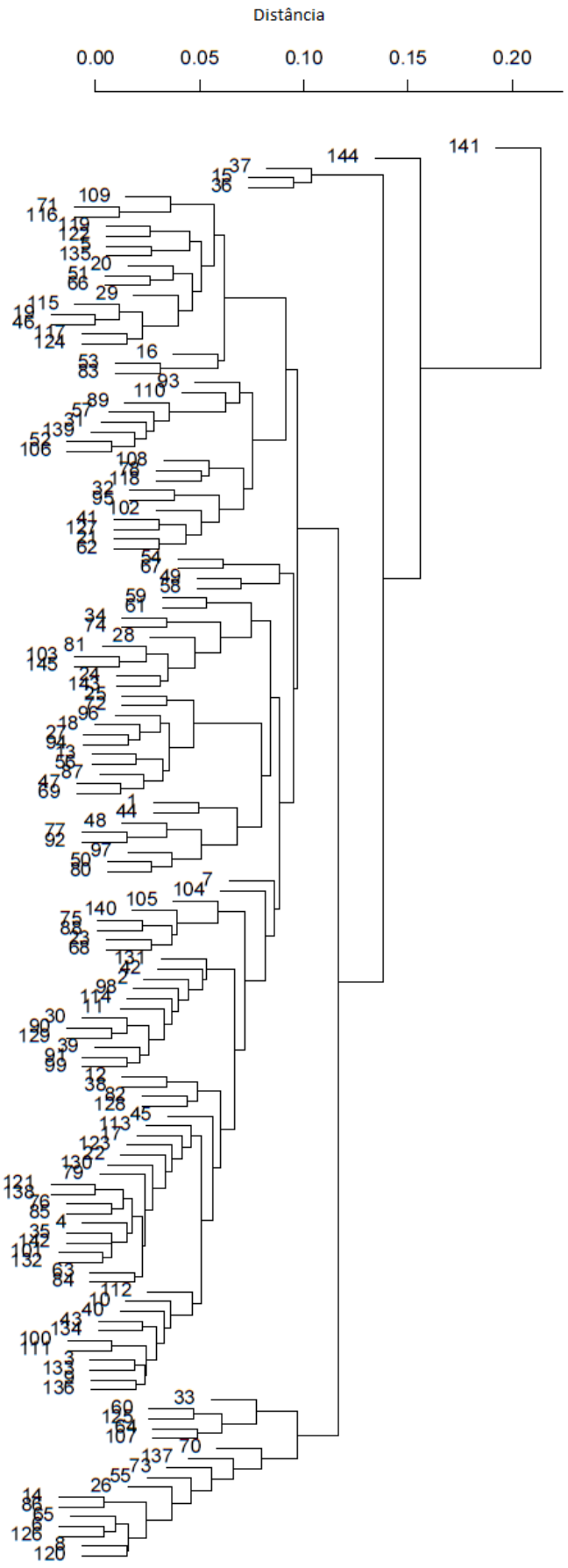

Figura 5: Dendograma do método de agrupamento da média

4.2 ENTROPIAS

A análise das entropias foram feitas através do software R 2.15.1 utilizando as bibliotecas seqinr (Charif e Lobry, 2007) e vegan (Jari Oksanen, et al., 2008).

Primeiramente foram calculados vários tipos de entropias sem considerar nenhum tipo de ordem, apenas todas as possíveis combinações, variando o tamanho da caixa de 1 a 10, que nas sequências são as combinações dos quatro nucleotídeos (A, C, G, T). Para o tamanho de caixa 1, os possíveis eventos são os próprios nucleotídeos, a partir do tamanho de caixa 2 os possíveis eventos são todas as possíveis combinações.

Tabela 7: Eventos possíveis para cada tamanho de caixa

Tamanho da caixa	Possíveis eventos	Total de combinações
1	A, C, G, T	4
2	AA, AC, AG,..., TT	16
3	AAA, AAC,..., TTT	64
4	AAAA, AAAC,..., TTTT	256
5	AAAAA,..., TTTTT	1024
6	AAAAAA,..., TTTTTT	4096
7	AAAAAAA,..., TTTTTTT	16384
8	AAAAAAAA,..., TTTTTTTT	65536
9	AAAAAAAAA,..., TTTTTTTTT	262144
10	AAAAAAAAAA,..., TTTTTTTTTT	1048576
11	...	4^n

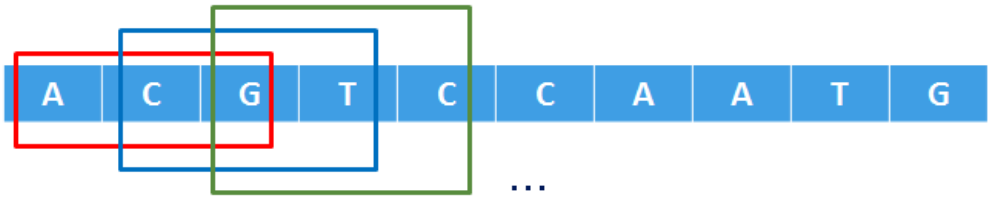

Figura 6: Ilustração de uma sequência com tamanho de caixa 3

Foram calculadas as entropias de Shannon, Tsallis e Rényi para as 145 sequências de DNA. Foi variado o tamanho da caixa de 1 a 10, com o objetivo de encontrar a entropia que apresentar a melhor correlação de Mantel com as distâncias genéticas das 145 sequências de DNA.

Nas Tabelas 8, 9, 10, 11, 12 observa-se que quando aumentamos o tamanho da caixa para maior que 3, e considerando um nível de significância de 5%, não se rejeita hipótese de não correlação entre a distâncias por entropia e a distância genética das sequências de DNA. Os valores das entropias foram normalizados pelo valor máximo para uma melhor interpretação, visto que quando se aumenta o tamanho da caixa, consequentemente se aumenta o valor das entropias.

Tabela 8: Estatísticas da entropia de Shannon

Estatísticas	Tamanho da caixa									
	1	2	3	4	5	6	7	8	9	10
Correlação	0,184	0,159	0,159	0,071	0,034	0,031	0,034	0,042	0,044	0,045
Coeficiente de correlação cofenética	0,768	0,704	0,758	0,778	0,944	0,965	0,975	0,977	0,977	0,977
P_valor	0,001	0,002	0,002	0,074	0,220	0,265	0,235	0,189	0,197	0,189
Máx	0,998	0,999	1,000	0,999	0,995	0,997	0,999	0,998	0,999	1,000
Mín	0,969	0,966	0,946	0,926	0,916	0,914	0,914	0,913	0,912	0,912
Média	0,985	0,984	0,976	0,959	0,945	0,944	0,944	0,940	0,940	0,939
1º quartil	0,982	0,980	0,968	0,951	0,938	0,938	0,938	0,935	0,935	0,935
Mediana	0,985	0,984	0,975	0,958	0,941	0,941	0,941	0,936	0,936	0,936
3º quartil	0,989	0,988	0,981	0,965	0,946	0,944	0,944	0,938	0,937	0,937

Tabela 9: Estatísticas da entropia de Tsallis para q=0,5

Estatísticas	Tamanho da caixa									
	1	2	3	4	5	6	7	8	9	10
Correlação	0,183	0,154	0,125	0,073	0,031	0,035	0,038	0,045	0,046	0,047
Coeficiente de correlação cofenética	0,761	0,759	0,725	0,867	0,954	0,968	0,976	0,977	0,977	0,977
P_valor	0,002	0,003	0,007	0,081	0,239	0,229	0,224	0,173	0,171	0,195
Máx	0,999	0,999	1,000	0,993	0,990	0,993	1,000	0,994	0,997	0,999
Mín	0,979	0,967	0,907	0,816	0,774	0,766	0,764	0,759	0,757	0,756
Média	0,990	0,985	0,961	0,897	0,848	0,842	0,841	0,829	0,827	0,827
1º quartil	0,988	0,980	0,947	0,876	0,830	0,824	0,825	0,814	0,814	0,813
Mediana	0,990	0,985	0,960	0,891	0,838	0,832	0,831	0,817	0,817	0,817
3º quartil	0,993	0,989	0,972	0,910	0,850	0,839	0,837	0,823	0,821	0,820

Tabela 10: Estatísticas da entropia de Tsallis para q=2,0

Estatísticas	Tamanho da caixa									
	1	2	3	4	5	6	7	8	9	10
Correlação	0,183	0,151	0,145	0,050	0,049	0,026	0,027	0,038	0,047	0,041
Coeficiente de correlação cofenética	0,731	0,725	0,729	0,774	0,879	0,942	0,966	0,976	0,977	0,977
P_valor	0,001	0,001	0,002	0,140	0,159	0,293	0,264	0,225	0.218	0,194
Máx	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Mín	0,970	0,987	0,993	0,996	0,998	0,998	0,998	0,998	0,998	0,998
Média	0,985	0,994	0,997	0,998	0,999	0,999	0,999	0,999	0,999	0,999
1º quartil	0,982	0,993	0,996	0,997	0,998	0,999	0,999	0,999	0,999	0,999
Mediana	0,985	0,994	0,997	0,998	0,998	0,999	0,999	0,999	0,999	0,999
3º quartil	0,989	0,996	0,997	0,998	0,999	0,999	0,999	0,999	0,999	0,999

Tabela 11: Estatísticas da entropia de Rényi para q=0,5

Estatísticas	Tamanho da caixa									
	1	2	3	4	5	6	7	8	9	10
Correlação	0,183	0,153	0,122	0,072	0,029	0,033	0,036	0,036	0,044	0,045
Coeficiente de correlação cofenética	0,761	0,760	0,728	0,848	0,952	0,967	0,975	0,978	0,978	0,978
P_valor	0,001	0,001	0,004	0,088	0,267	0,223	0,256	0,256	0,193	0,190
Máx	0,999	1,000	1,000	0,997	0,997	0,998	1,000	1,000	0,999	1,000
Mín	0,984	0,982	0,958	0,927	0,915	0,914	0,914	0,914	0,912	0,912
Média	0,993	0,992	0,983	0,960	0,945	0,944	0,944	0,944	0,940	0,939
1º quartil	0,991	0,989	0,977	0,952	0,938	0,938	0,939	0,939	0,935	0,935
Mediana	0,993	0,993	0,982	0,958	0,941	0,941	0,941	0,941	0,936	0,936
3º quartil	0,995	0,994	0,988	0,966	0,946	0,943	0,943	0,943	0,937	0,937

Tabela 12: Estatísticas da entropia de Rényi para q=2,0

Estatísticas	Tamanho da caixa									
	1	2	3	4	5	6	7	8	9	10
Correlação	0,186	0,155	0,152	0,055	0,049	0,028	0,031	0,041	0,044	0,044
Coeficiente de correlação cofenética	0,730	0,700	0,725	0,778	0,884	0,949	0,970	0,977	0,978	0,978
P_valor	0,001	0,003	0,001	0,136	0,171	0,237	0,266	0,203	0,197	0,192
Máx	0,994	0,998	0,009	0,995	0,993	0,996	0,998	0,997	0,999	1,000
Mín	0,938	0,942	0,931	0,921	0,917	0,915	0,913	0,913	0,912	0,912
Média	0,970	0,971	0,963	0,951	0,945	0,945	0,945	0,940	0,939	0,939
1º quartil	0,963	0,964	0,954	0,942	0,937	0,937	0,938	0,934	0,935	0,935
Mediana	0,969	0,970	0,962	0,949	0,942	0,941	0,941	0,936	0,936	0,936
3º quartil	0,977	0,978	0,970	0,956	0,947	0,946	0,945	0,938	0,937	0,937

O desafio para encontrar uma boa correlação entre as distâncias por entropia e genética é o fato da distância genética considerar a posição dos nucleotídeos, enquanto que as entropias estão considerando a quantidade dos nucleotídeos nas sequências.

Ao aumentar o tamanho da caixa o valor da entropia é reduzido (com exceção da entropia de Tsallis para q=2,0), isso indica uma menor desorganização dos dados, com isso se tem um melhor agrupamento considerando o coeficiente de correlação cofenética (CCC), porém a correlação é reduzida. Pela figura 7 se observa que a entropia que apresenta as maiores correlações são a de Shannon e Rényi com q=2,0, as duas tem um comportamento bem semelhante também em relação à entropia média e ao o coeficiente de correlação cofenética.

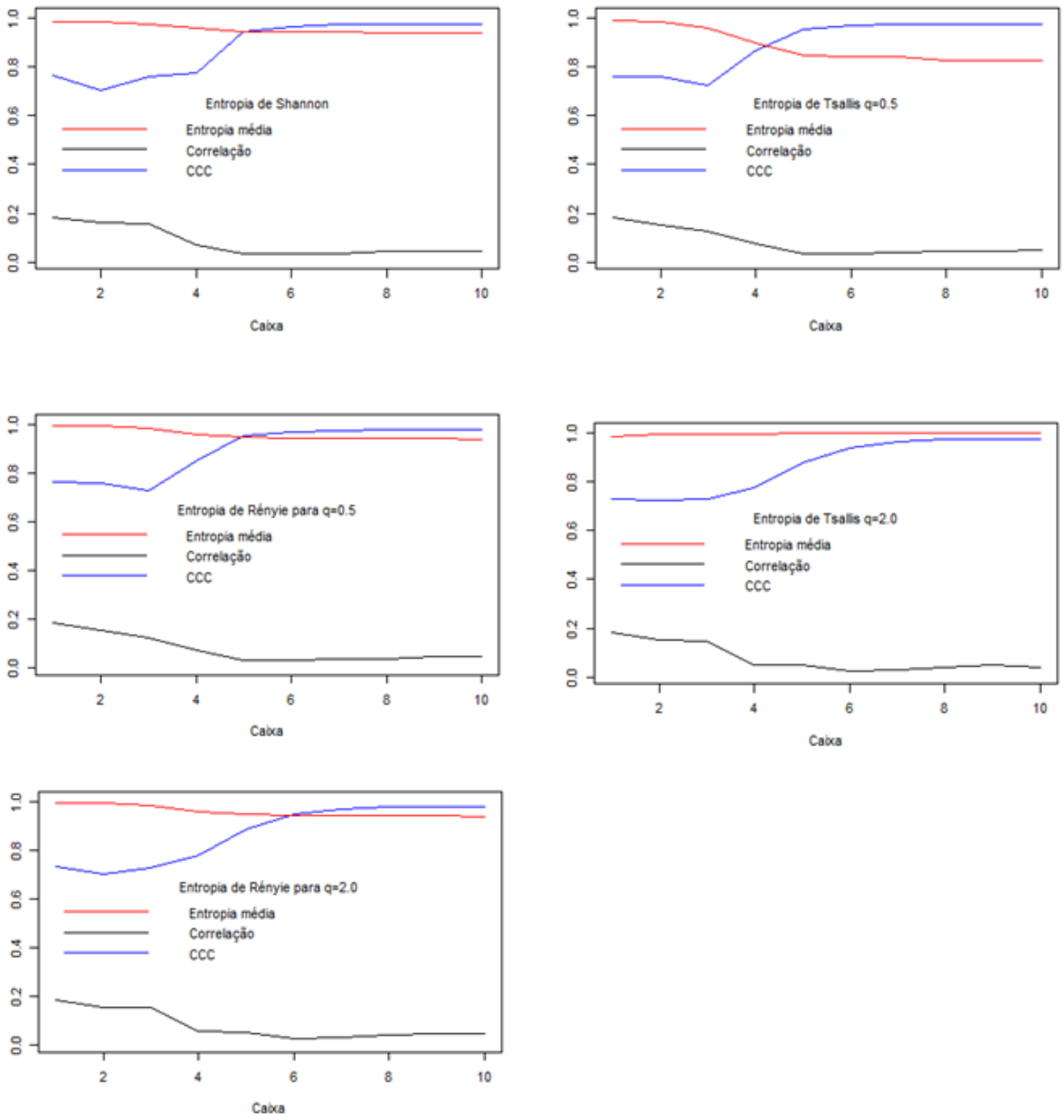

Figura 7: Comportamento das entropias

A partir dos resultados acima, foram realizados os agrupamentos para comparar com o agrupamento das distâncias genéticas. Para cada entropia será mostrado 2 agrupamentos, um com o tamanho de caixa 1, pois é onde se tem a maior correlação, e o outro será um valor extremo, nesse caso utilizaremos o tamanho de caixa 7, pois apesar de não apresentar correlação tem um alto coeficiente de correlação cofenética (CCC).

Para comparar o agrupamento das distâncias genéticas com as distâncias por entropias foi utilizada a mesma quantidade de grupos (5 grupos), identificados na análise da distância genética. Na figura 8 para Shannon com caixa 1, foi identificado um grupo 1 formado pelas sequências 19 e 125, grupo 2 pelas sequências 15,100, 107,137, e em seguida 3 grupos contendo muitas sequências, 68, 60 e 11 sequências em cada grupo respectivamente. Na figura 9 para Shannon com caixa 7, foi identificado um grupo 1 formado pelas sequências 42, 71, 96, 110 e 115, grupo 2 pelas sequências 55, 81, 84, 100, 107, 118 e 135, o grupos 3 formado pela sequência 19 e os grupos 4 e 5 formado por várias sequências, 11 e 121 respectivamente. Comparando esses 2 agrupamentos foram identificadas as sequências 100 e 107 em mesmo grupo.

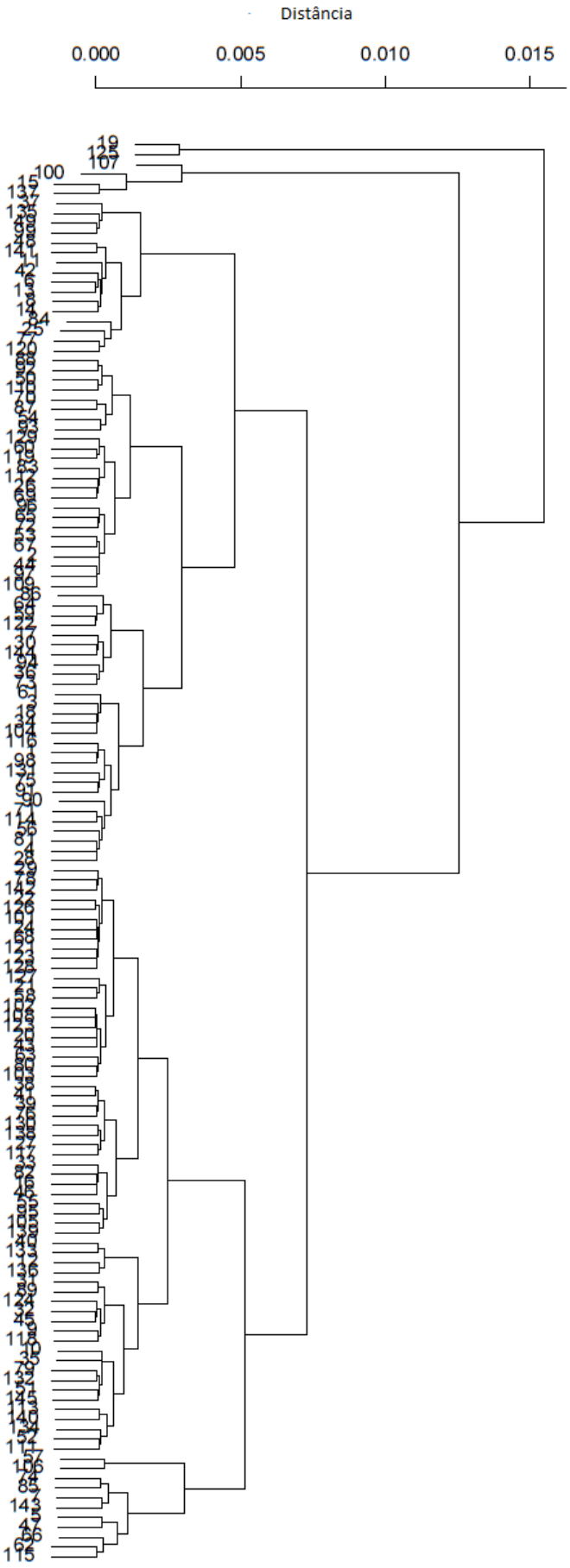

Figura 8: Dendograma pelo método da média da entropia de Shannon com tamanho de caixa 1

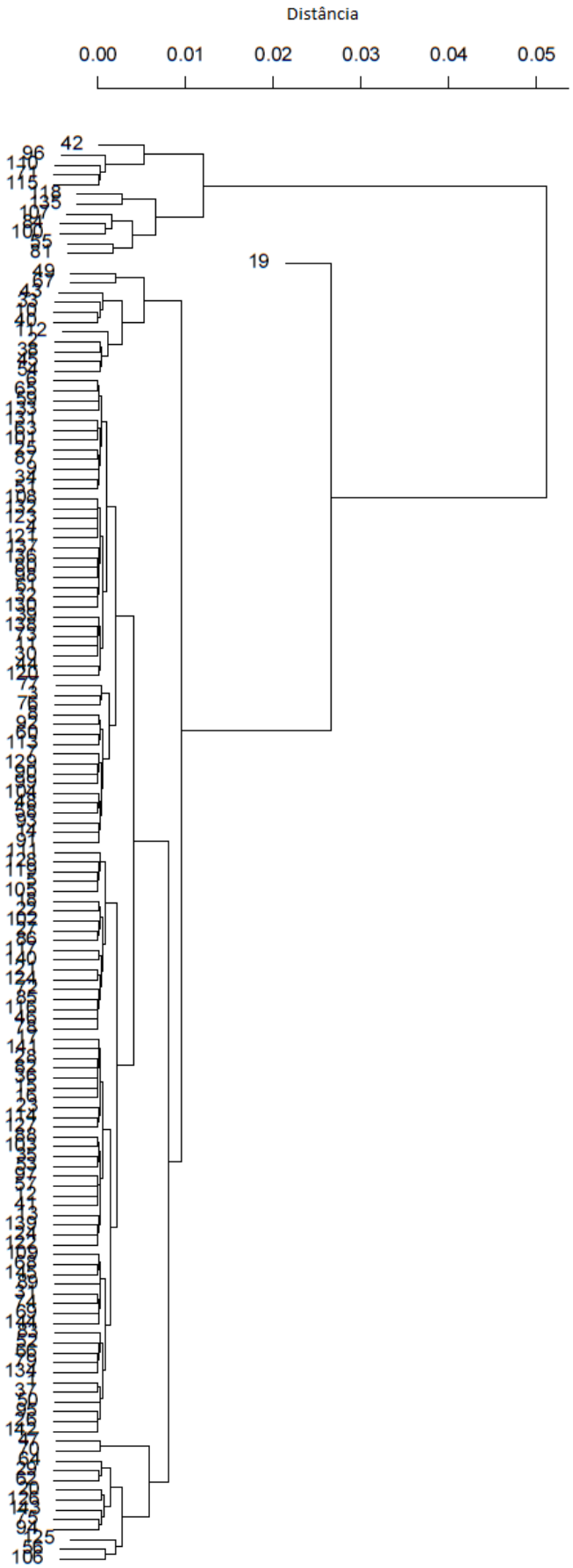

Figura 9: Dendograma pelo método da média da entropia de Shannon com tamanho de caixa 7

Na figura 10, de acordo com a entropia de Tsallis com caixa 1 e q=0.5, foi identificado um primeiro grupo formado com as sequências 19 e 125. Os demais grupos são formados por muitas sequências, 19, 58, 55, e 11, em cada grupo respectivamente. Na figura 11, para Tsallis com caixa 7 e q=0.5, foi identificado um primeiro grupo formado pelas sequências 71, 96, 110 e 115. Os grupos 2, 4 e 5 formados por vários sequências, 8, 11 e 121, respectivamente, e um terceiro grupo formado unicamente pela sequência 19. Comparando esses 2 agrupamentos foi identificado que a sequência 19 aparece nos dois agrupamentos de forma isolada dos demais grupos.

Na figura 12 para Tsallis com caixa 1 e q=2.0, foi identificado um primeiro grupo formado pela sequência 19, um segundo grupo , pelas sequências 57, 106 e 125. Os grupos 3 e 4 são formados por muitas sequências, 97 e 40 respectivamente, e o grupo 5 é formado pelas sequências 15, 100, 107 e 137. Na figura 13 para Tsallis com caixa 7 e q=2.0, foi identificado os grupos 1, 3, 4 e 5 com muitas sequências, 12, 11, 7 e 114 respectivamente, e o grupo 2 contendo a sequência 19. Comparando esses 2 agrupamentos foi identificado que a sequências 19 aparece nos dois agrupamentos de forma isolada dos demais grupos.

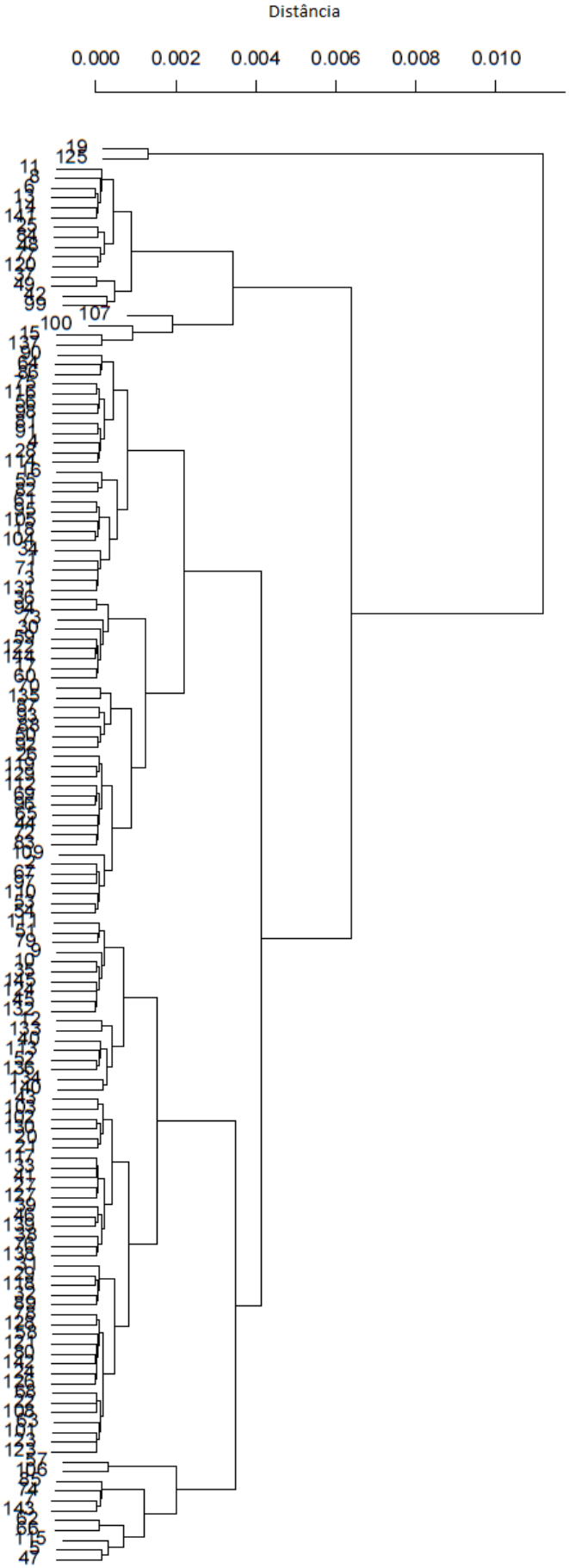

Figura 10: Dendograma pelo método da média da entropia de Tsallis com tamanho de caixa 1 e q=0.5

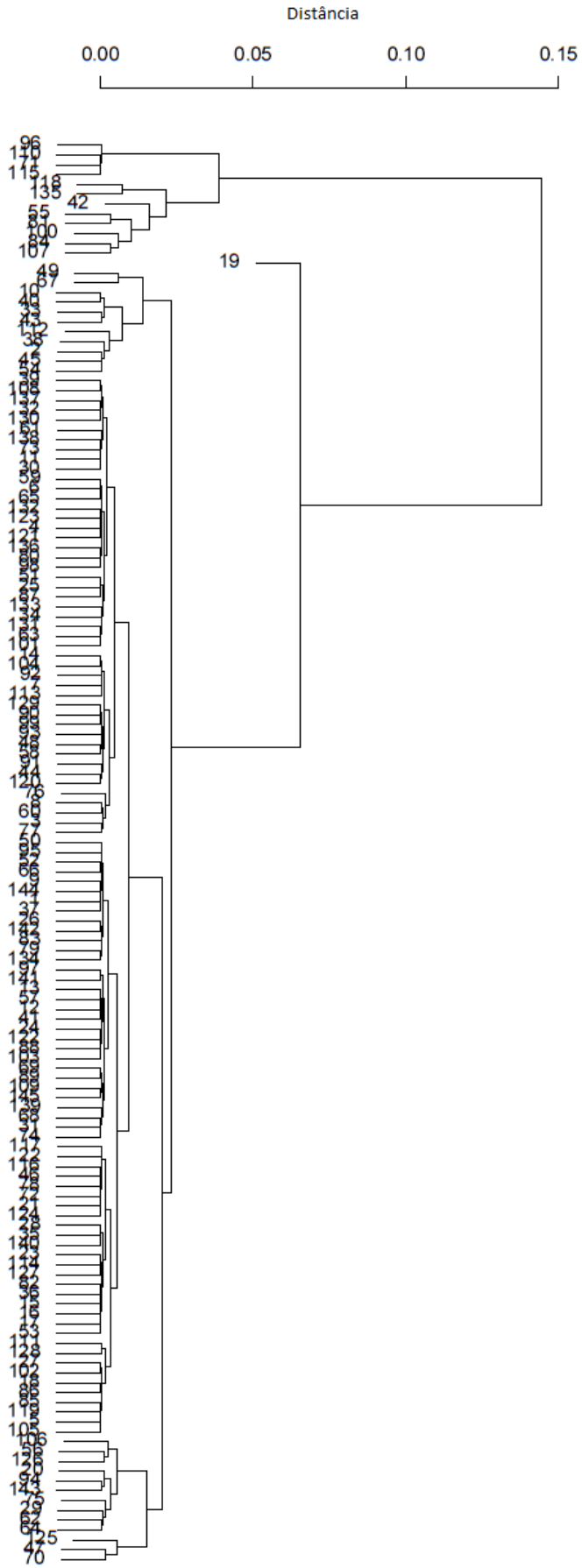

Figura 11: Dendograma pelo método da média da entropia de Tsallis com tamanho de caixa 7 e q=0.5

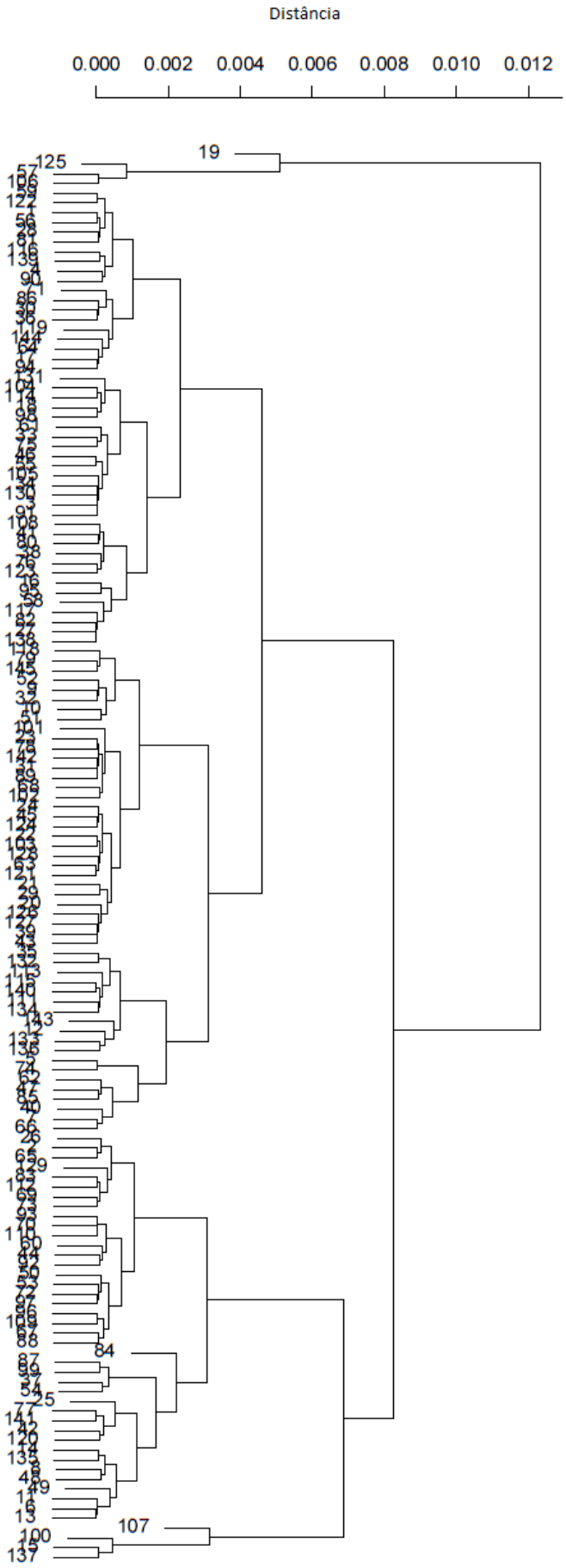

Figura 12: Dendograma pelo método da média da entropia de Tsallis com tamanho de caixa 1 e q=2.0

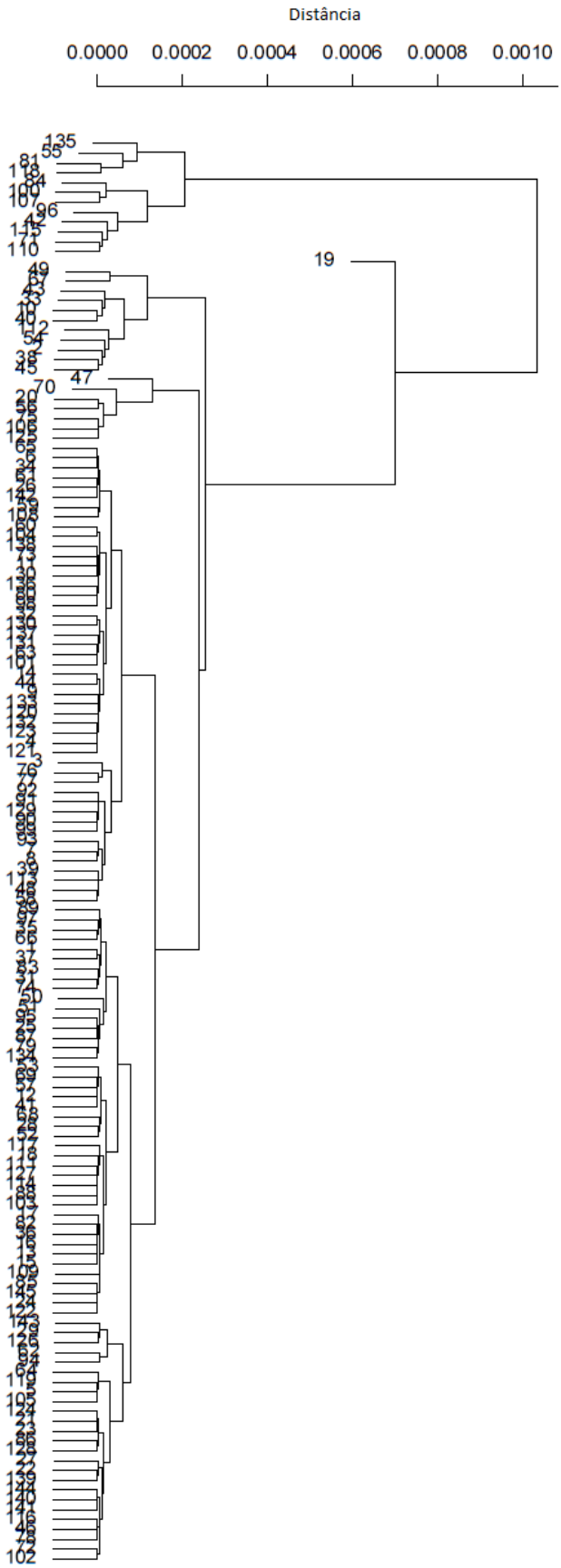

Figura 13: Dendograma pelo método da média da entropia de Tsallis com tamanho de caixa 7 e q=2.0

Na figura 14 para Rényi com caixa 1 e q=0.5, foi identificado um grupo 1 formado pela sequência 19 e 125, e os demais grupos formados por muitas sequências, 19, 58, 55 e 1, respectivamente. Na figura 15 para Rényi com caixa 7 e q=0.5, foi identificado os grupos 2, 4 e 5 com muitas sequências, 8, 11 e 121 respectivamente, grupos 1 contendo as sequências 71, 96, 110 e 115, e o grupo 3 contendo a sequência 19. Comparando esses 2 agrupamentos foi identificado que a sequências 19 aparece nos dois agrupamentos de forma isolada dos demais grupos.

Na figura 16 para Rényi com caixa 1 e q=2.0, foi identificado um grupo 1 formado pela sequência 19, o grupo 2 formado pelas sequências 57, 106 e 125, o grupo 5 formado pelas sequências 15, 100, 107 e 137, e os grupos 3 e 4 formados por muitas sequências, 97 e 40 respectivamente. Na figura 17 para Rényi com caixa 7 e q=2.0, foi identificado os grupos 2, 4 e 5 com muitas sequências, 8, 11 e 121 respectivamente, o grupo 1 contendo as sequências 55, 81, 118 e 135, e o grupo 3 contendo a sequência 19. Comparando esses 2 agrupamentos foi identificado que a sequência 19 aparece nos dois agrupamentos de forma isolada dos demais grupos.

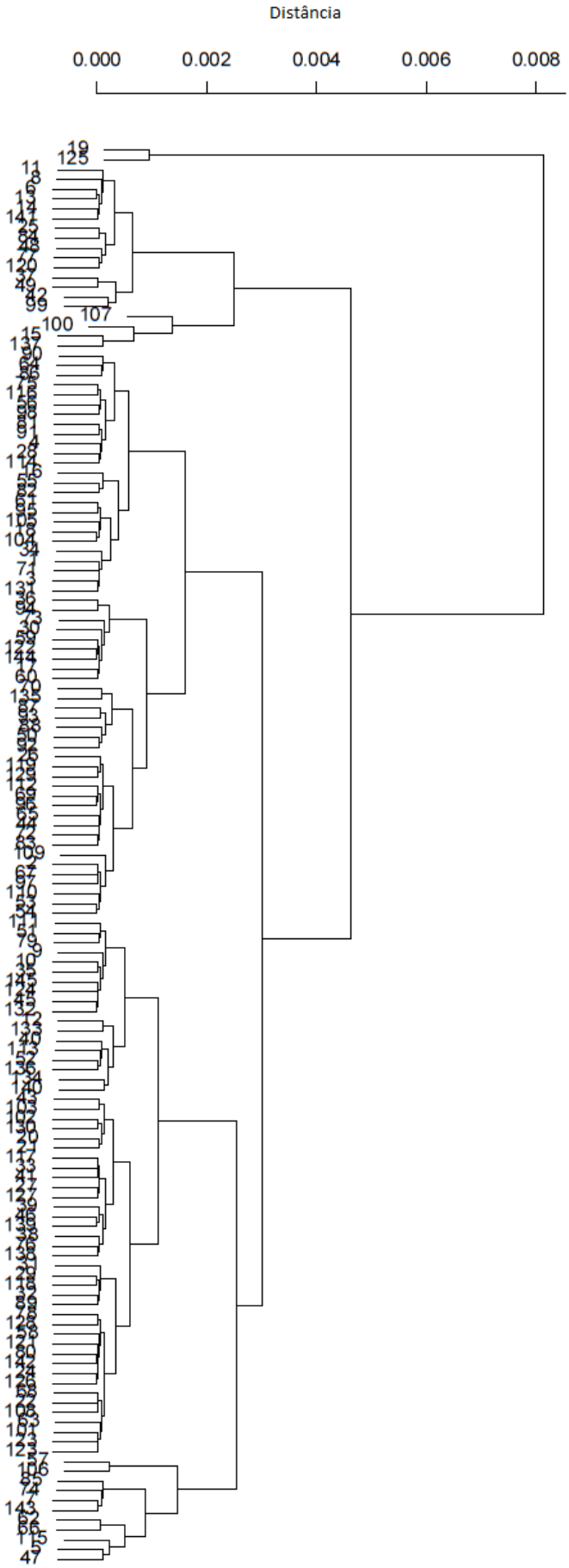

Figura 14: Dendograma pelo método da média da entropia de Rényi com tamanho de caixa 1 e q=0.5

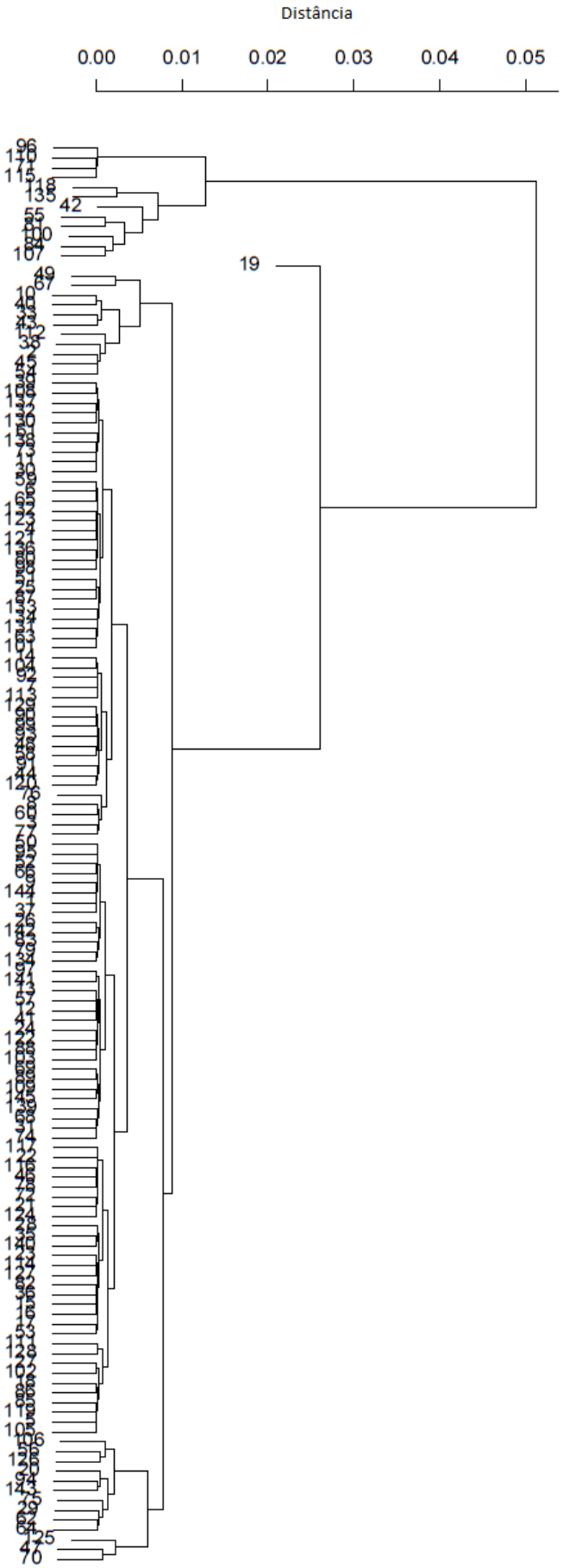

Figura 15: Dendograma pelo método da média da entropia de Rényi com tamanho de caixa 7 e q=0.5

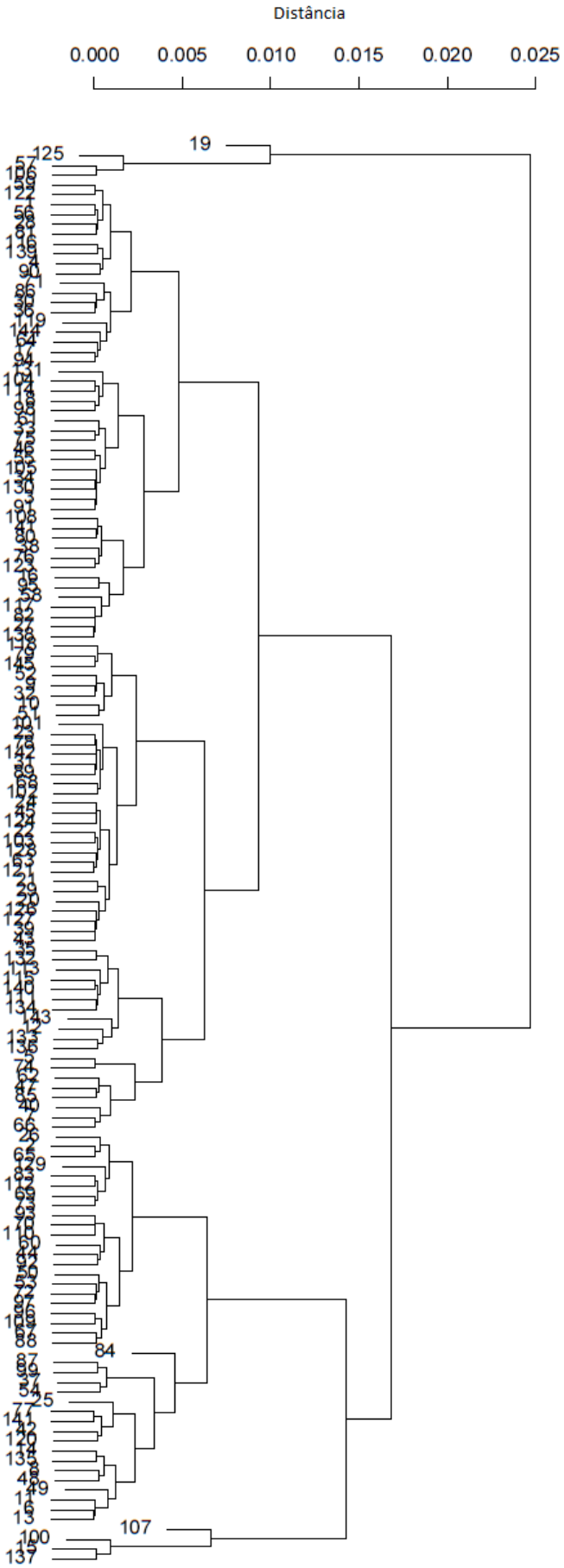

Figura 16: Dendograma pelo método da média da entropia de Rényi com tamanho de caixa 1 e q=2.0

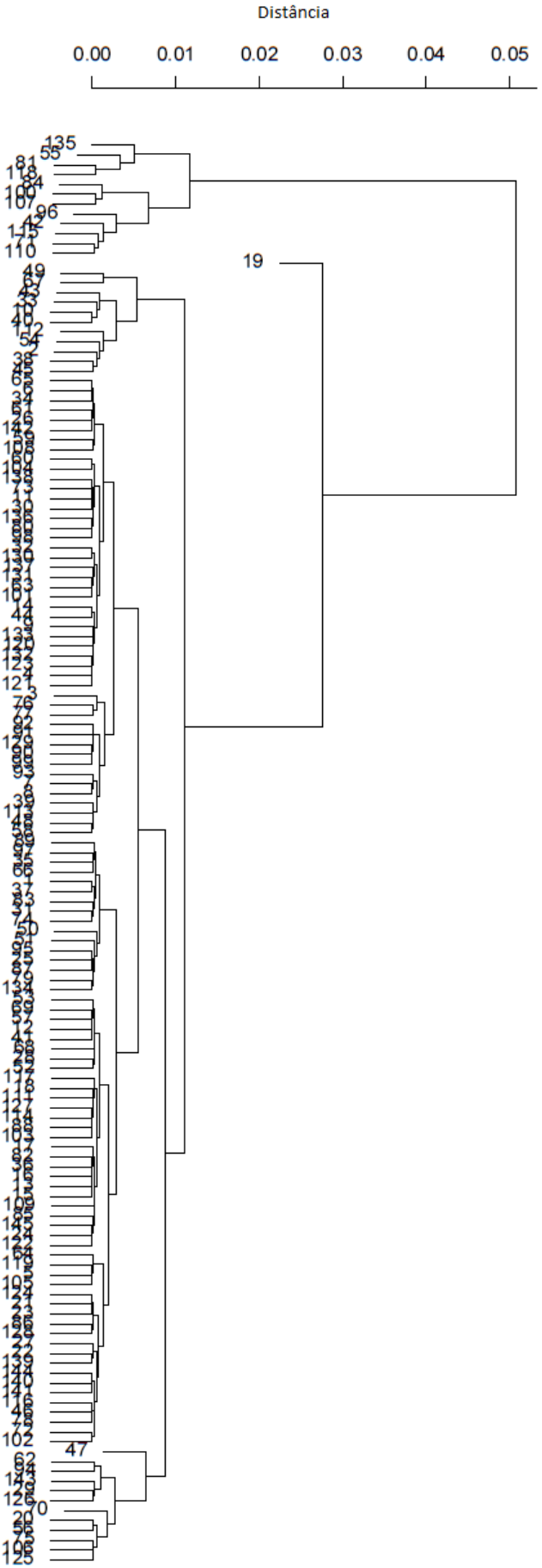

Figura 17: Dendograma pelo método da média da entropia de Rényi com tamanho de caixa 7 e q=2.0

Como esperado, devido a baixa correlação entre as distâncias genéticas e distâncias por entropia, não foram identificados grupos em comum entre os agrupamentos genéticos e por entropia. No agrupamento por entropia, foi encontrada a sequência 19 sempre aparecendo, na maioria das vezes, de forma isolada e em alguns casos agrupada com a sequência 125. Na etapa posterior foram realizadas várias simulações com o mesmo objetivo de tentar encontrar alguma entropia em que a distância por entropia das sequências tenha uma boa correlação com as distâncias genéticas.

4.3 ENTROPIAS EM BLOCOS

Nessa etapa serão calculadas as entropias na qual consideramos um determinado bloco antes ou depois dos possíveis eventos, nesse caso trata-se de entropia condicional. Assim, reduzimos um pouco o viés da entropia pelo fato de não considerar a posição dos eventos.

Tabela 13: Eventos possíveis para cada tamanho de caixa e tamanho de bloco

Tamanho do bloco	Tamanho da caixa	Possíveis eventos (bloco antes)	Possíveis eventos (bloco depois)
1 (A)	1	AA, AC, AG, AT	AA, CA, GA, TA
1 (A)	2	AAA, AAC, AAG,..., ATT	AAA, ACA, AGA,..., TTA
2 (AA)	1	AAA, AAC, AAG, AAT	AAA, CAA, GAA, TAA
2 (AA)	2	AAAA, AAAC,...,AATT	AAAA, ACAA,...,TTAA
.	.	.	.
.	.	.	.
.	.	.	.

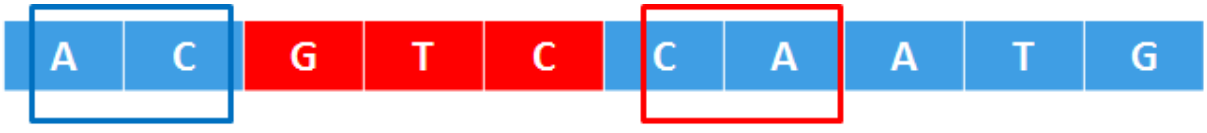

Figura 18: Ilustração de uma sequência com tamanho de bloco 3 e tamanho de caixa 2 antes e depois do bloco

Iniciou-se com blocos de tamanho 1 antes dos eventos até blocos de tamanho 4. Como existem quatro possibilidades de nucleotídeos, haverá no total 4^t blocos, onde t é o tamanho do bloco. Para cada bloco o tamanho da caixa será variado de um a seis. Como as maiores correlações obtidas foram a de Shannon, será utilizada a mesma fórmula matemática da entropia de Shannon, sendo que, agora, considerando uma probabilidade condicional, que é a probabilidade de um determinado evento ocorrer dado que ocorreu um bloco *L*.

$$P(x) = P(x|L) = \frac{P(x \cap L)}{P(L)}, \text{ tal que } P(L) > 0$$

$P(L)$ é a probabilidade de ocorrer o bloco antes ou depois dos possíveis eventos x (a depender se existe interesse no bloco antes ou depois dos possíveis eventos x);
$P(x \cap L)$ é a probabilidade de ocorrer exatamente a sequência desejada, que é o evento x antes ou depois do bloco *L*.

$$H_S(X) = -\sum_{x \in X} p(x) log_2 p(x)$$

Tabela 14: Possíveis blocos de tamanho um

Código do bloco	Bloco
1	A
2	C
3	G
4	T

Pela figura 19, as maiores correlações para o bloco de tamanho 1 antes dos eventos são para os tamanhos de caixa 2 e 3, que apresentaram correlações de 0,1984 e 0,2137, respectivamente, e nos dois casos ocorreu para o bloco 2, ou seja, o bloco C, como se observa na Tabela 14. Nesse caso, para o cálculo das entropias foi considerado apenas os eventos CAA, CAC, CAG, CAT, CCA, CCC, CCG, CCT, CGA, CGC, CGG, CGT, CTA, CTC, CTG e CTT para o tamanho de caixa 2.

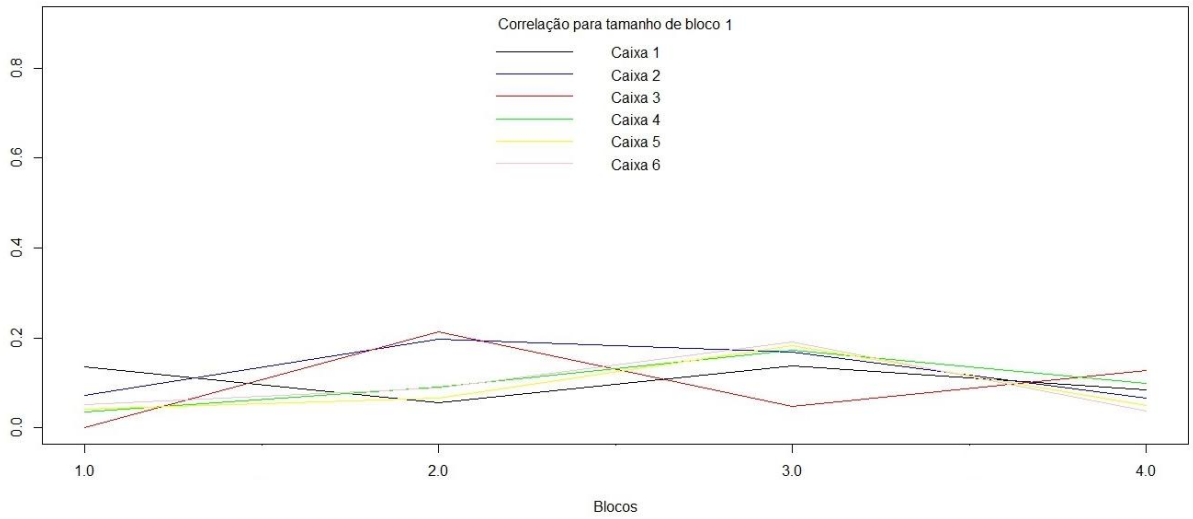

Figura 19: Correlações das distâncias por entropias com tamanho de bloco 1 e tamanho de caixa de 1 a 6 com as distâncias genéticas

Tabela 15: Possíveis blocos de tamanhos dois

Código do bloco	Bloco
1	AA
2	AC
3	AG
4	AT
5	CA
6	CC
7	CG
8	CT
9	GA
10	GC
11	GG
12	GT
13	TA
14	TC
15	TG
16	TT

De acordo com a figura 20 observa-se que as maiores correlações para o tamanho de bloco 2 são para os tamanhos de caixa 4, 5 e 6, que apresentaram correlações de 0,3067, 0,358 e 0,344, respectivamente, e em todos os casos ocorreu para o bloco 3, ou seja, o bloco AG. Se forem divididas as sequências considerando o tamanho de caixa de 4 a 6 após o ocorrer o bloco AG, já foi alcançada uma

correlação bem acima de quando não considerávamos os blocos, um aumento de quase 100%.

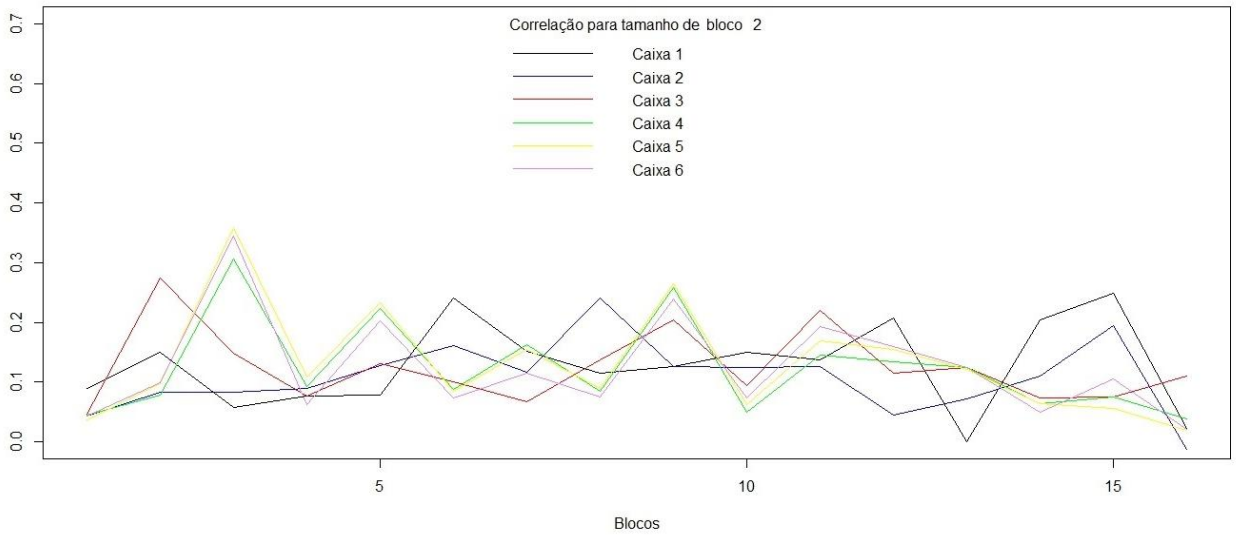

Figura 20: Correlações das distâncias por entropias com tamanho de bloco 2 e tamanho de caixa de 1 a 6 com as distâncias genéticas

A figura 21 mostra 64 possíveis blocos, pois o bloco é de tamanho 3, entre os quais registra uma correlação de 0,50, que ocorre para o tamanho de caixa 1 após o ocorrer o bloco CAC. Se dividir a sequência em blocos de CAC considerando apenas o nucleotídeo após o bloco, se terá melhores resultados do que os obtidos até então, para esse caso específico foram considerados apenas os eventos CACA, CACC, CACG e CACT. Na figura 22 com o bloco de tamanho 4, foi encontrada uma correlação de 0,46 para o tamanho de caixa 1 após o bloco ACTG. Ainda na figura 22 pode se observar alguns cortes no gráfico, isso se dar pelo fato de que alguns blocos de tamanho 4 não ocorrem, sendo assim não teria como calcular as entropias pois a soma da probabilidade dos eventos é zero.

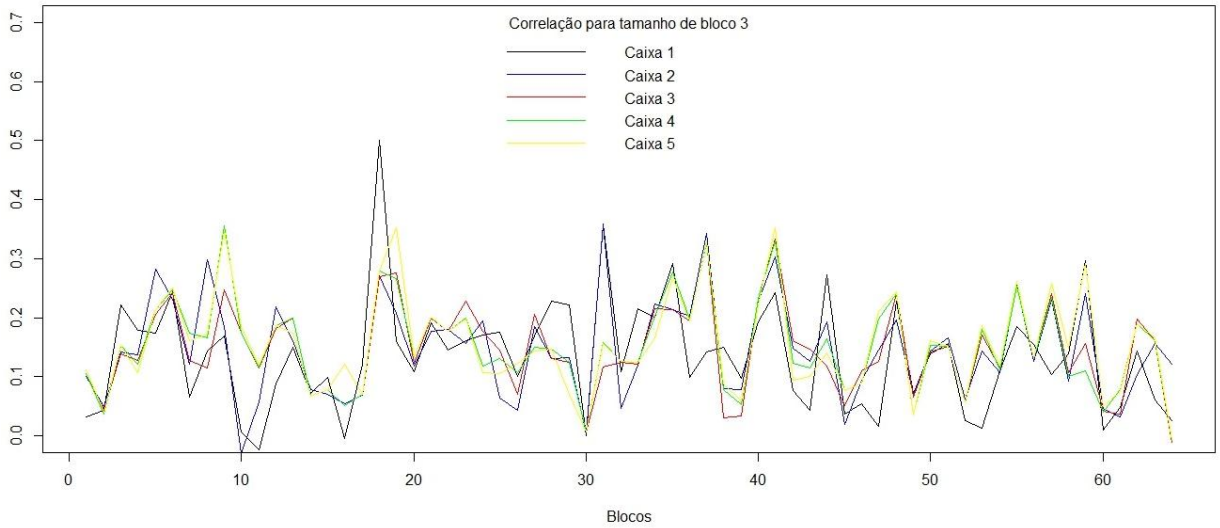

Figura 21: Correlações das distâncias por entropias com tamanho de bloco 3 e tamanho de caixa de 1 a 5 com as distâncias genéticas

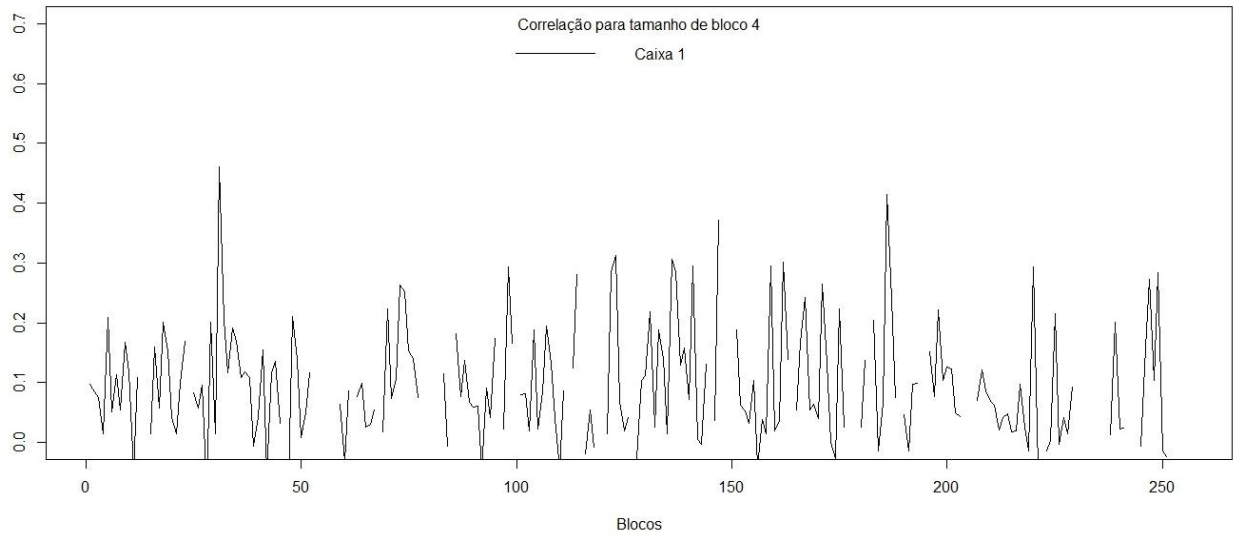

Figura 22: Correlações das distâncias por entropias com tamanho de bloco 4 e tamanho de caixa 1 com as distância genéticas

Foi realizado o procedimento anterior, porém agora considerando-se o bloco após o evento e a variação do tamanho da caixa que será de 1 a 5 com o tamanho do bloco de 1 a 3. Pela figura 23 observa-se que as maiores correlações ocorrem para o tamanho de caixa 1, o qual se refere ao bloco A.

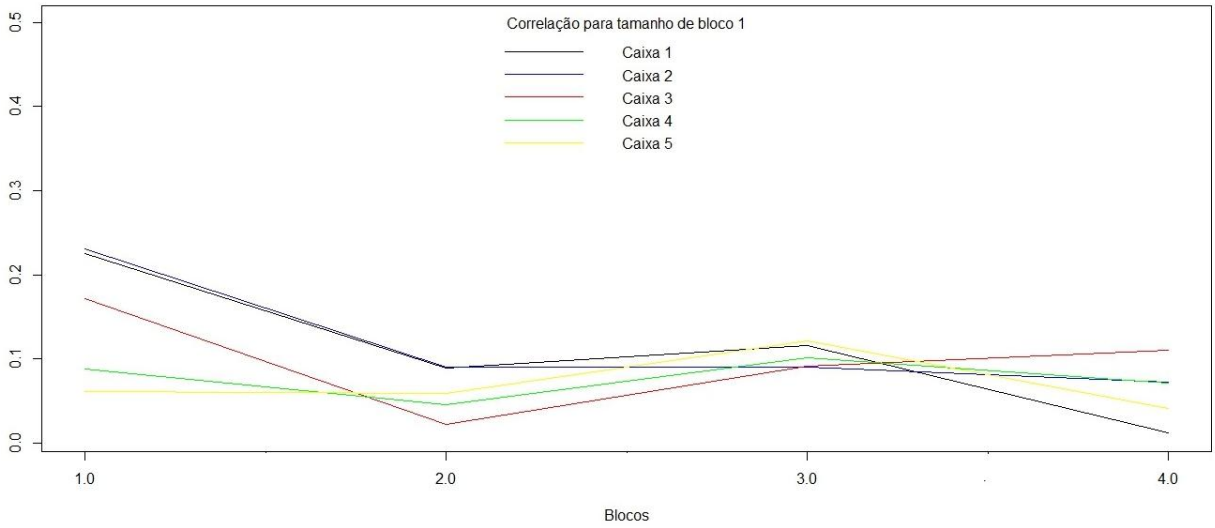

Figura 23: Correlações das distâncias por entropias com tamanho de bloco 1 após os eventos e tamanho de caixa de 1 a 5 com as distâncias genéticas

Na figura 24 as maiores correlações são para os tamanhos de caixa 1 e 5, que apresentaram correlações de 0,2945 e 0,27710, respectivamente, nos dois casos ocorreu para o bloco 3, ou seja, o bloco AG.

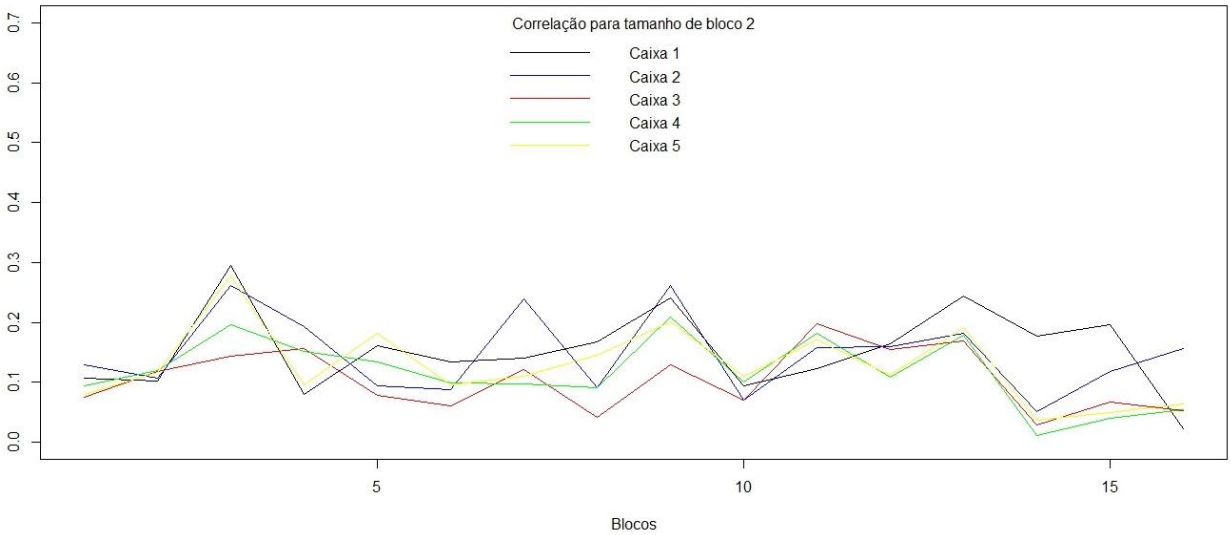

Figura 24: Correlações das distâncias por entropias com tamanho de bloco 2 após os eventos e tamanho de caixa de 1 a 5 com as distâncias genéticas

A figura 25 mostra 64 possíveis blocos, entre os quais registra uma correlação de 0,411 que ocorre para o tamanho de caixa 1 antes de ocorrer o bloco CAG, sendo considerado para esse caso os eventos ACAG, CCAG, GCAG, TCAG.

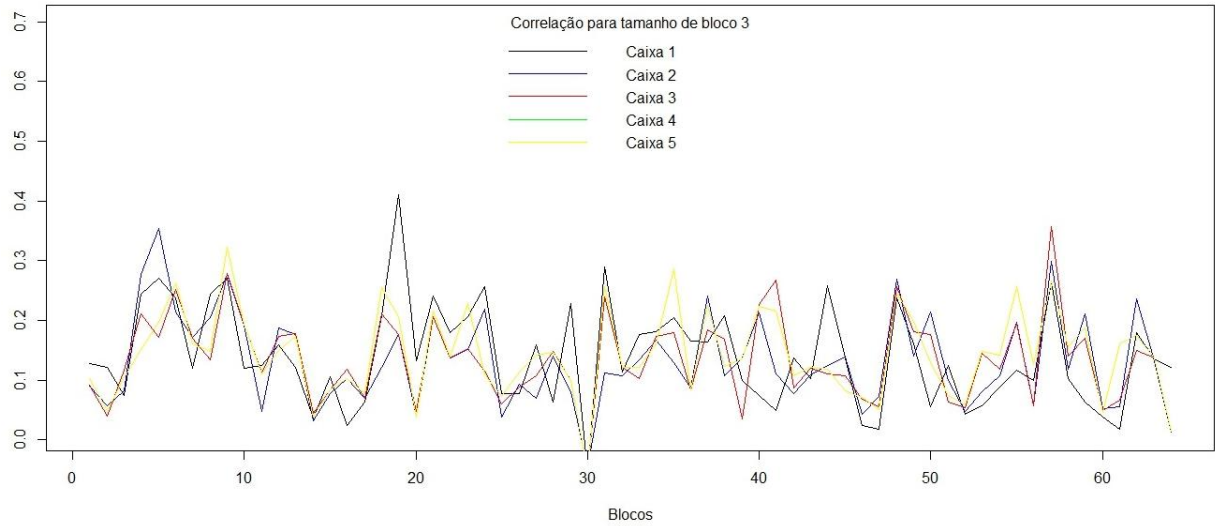

Figura 25: Correlações das distâncias por entropias com tamanho de bloco 3 após os eventos e tamanho de caixa de 1 a 5 com as distâncias genéticas

A partir dos resultados acima das entropias em blocos, foram realizados os agrupamentos para comparar com o agrupamento das distâncias genéticas. O agrupamento foi realizado para as sequências que apresentaram as maiores correlações, que foram os blocos CAC e ACTG antes dos eventos, e ambos com tamanho de caixa 1. Os valores da correlação foram de 0,5009 e 0,4608, respectivamente. Outro agrupamento foi realizado para o bloco CAG depois dos eventos, também com tamanho de caixa 1 e uma correlação de 0,411.

Na figura 26, que é a formada pelo bloco CAC antes dos eventos, foi identificado um grupo 1 formado pelas sequências 15 e 141, os grupos 2, 3 e 5 são formados por muitas sequências, 13, 109 e 20 sequências respectivamente cada grupo, e o grupo 4 formado pela sequência 55, o coeficiente de correlação cofenética é de 0,97. Na figura 27 para o bloco ACTG antes dos eventos foi encontrado um grupo 1 formado pelas sequências 15, 17, 36, 37, 141 e 144, em seguidas se tem os demais grupos formado por muitas sequências, 81, 12, 21 e 25 sequências respectivamente, o coeficiente de correlação cofenética também é de 0,97. Na figura 28 para o bloco CAG depois dos eventos apresentou também um coeficiente de correlação cofenética de 0,97, foi encontrado um grupo 1 formado pelas sequência 141 e 144, o grupos 2 é formando por 140 sequências, e depois os grupos 3, 4 e 5 cada um formado por apenas uma sequência que são as sequências 42, 96 e 143.

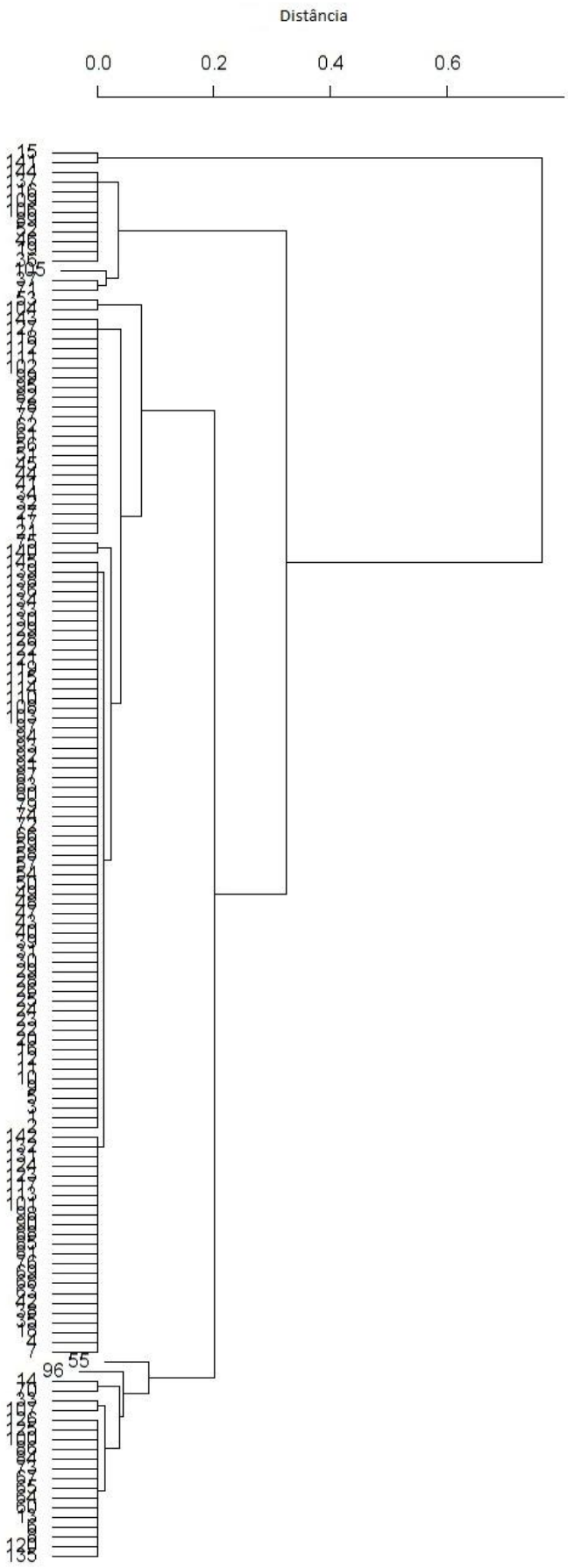

Figura 26: Dendograma pelo método da média da entropia de Shannon para o bloco CAC antes dos eventos com tamanho de caixa 1

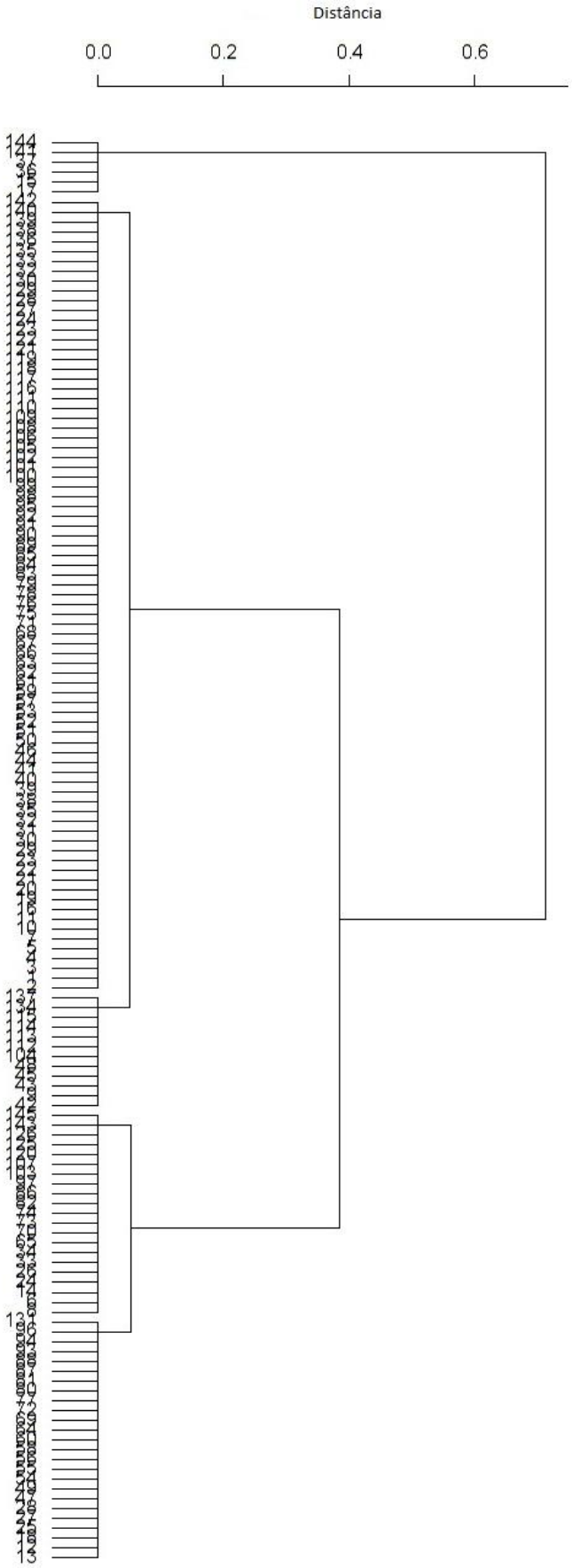

Figura 27: Dendograma pelo método da média da entropia de Shannon para o bloco ACTG antes dos eventos com tamanho de caixa 1

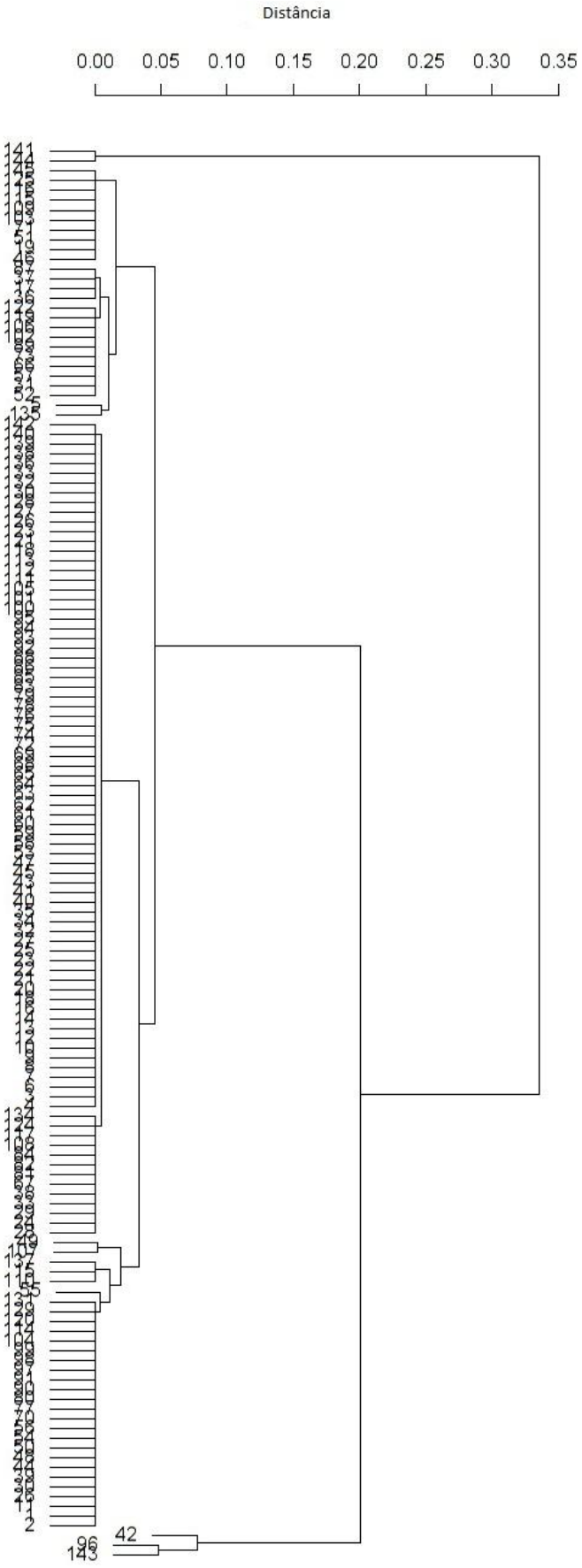

Figura 28: Dendograma pelo método da média da entropia de Shannon para o bloco CAG depois dos eventos com tamanho de caixa 1

4.4 *MELIPONA QUINQUEFASCIATA*

Após obter uma boa correlação entre a distância genética e a distância por entropia para o gene BoLA, foi aplicado o mesmo método para sequências maiores com o intuito de observar o comportamento dos resultados para diferentes tamanhos de sequências. Foram utilizadas sequências de DNA de tamanho 1.869 de abelhas sem ferrão *Melipona quinquefasciata* das regiões 18s, obtidas de várias colônias silvestres, em localidades distintas da Chapada do Araripe – CE, Chapada da Ibiapaba – CE, cidade do Canto do Buriti – PI e Luziânia – GO.

Os resultados podem ser visualizados pelas figuras 29, 30 e 31, onde foi utilizado o tamanho do bloco de 1 a 3, e o tamanho da caixa de 1 a 5. Na figura 29 para o tamanho de bloco 1, a maior correlação foi de 0,55 e ocorreu para o bloco A com tamanho de caixa 1. Na figura 30 para o tamanho de bloco 2, a maior correlação foi de 0,57 para o bloco TG e tamanho de caixa 3. E para o tamanho de bloco 3 (figura 31), a maior correlação foi de 0,61 para o bloco TCC com tamanho de caixa 1.

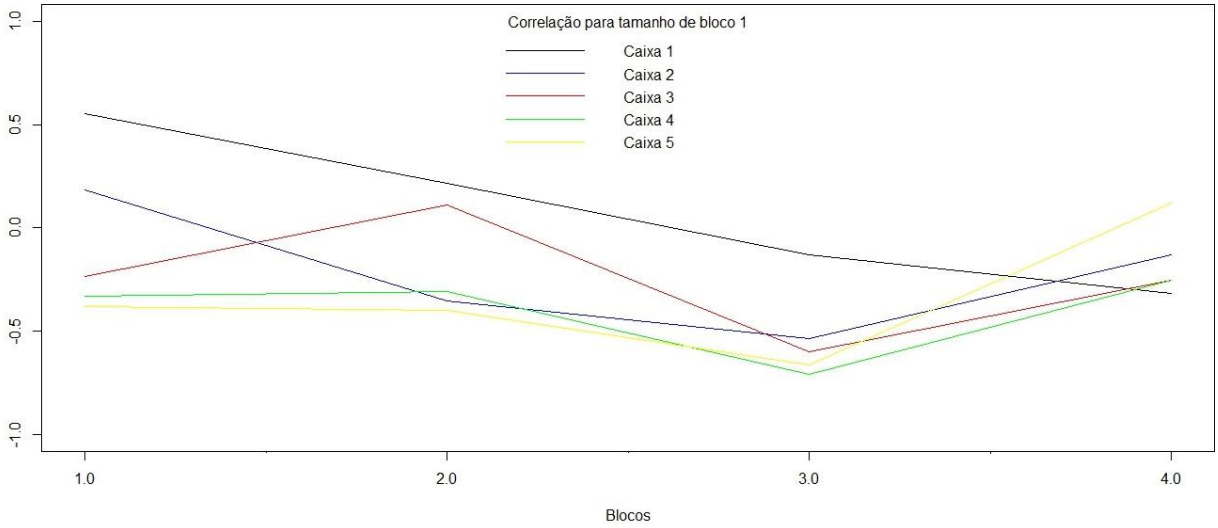

Figura 29: Correlações das distâncias por entropias com tamanho de bloco 1 antes dos eventos e tamanho de caixa de 1 a 5 com as distâncias genéticas das abelhas *Melipona quinquefasciata*

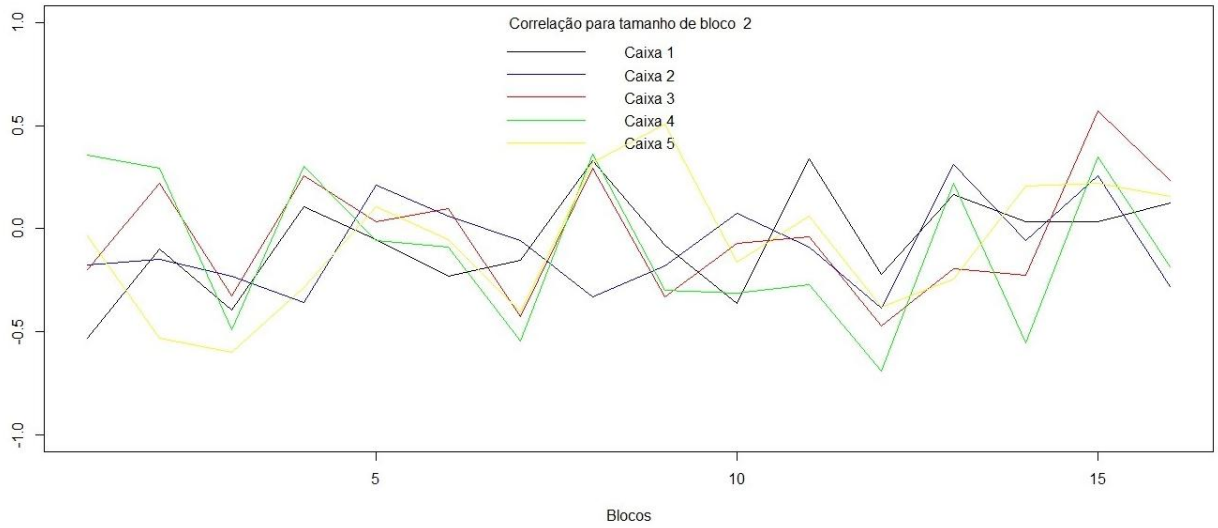

Figura 30: Correlações das distâncias por entropias com tamanho de bloco 2 antes dos eventos e tamanho de caixa de 1 a 5 com as distâncias genéticas das abelhas *Melipona quinquefasciata*

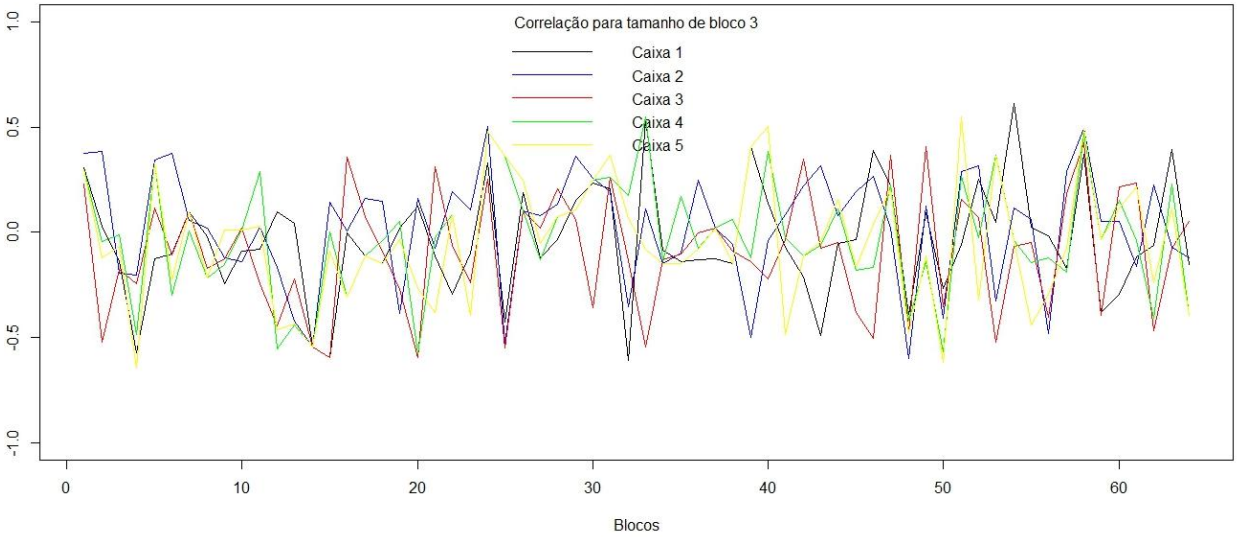

Figura 31: Correlações das distâncias por entropias com tamanho de bloco 3 antes dos eventos e tamanho de caixa de 1 a 5 com as distâncias genéticas das abelhas *Melipona quinquefasciata*

4.5 CROMOSSOMO 5

Outro resultado obtido foi para dez sequências de tamanho 188.332 do cromossomo 5 do *Homo Sapiens* obtidas do NCBI. Por se tratar de sequências muito grandes o alinhamento torna-se muito lento. Para maximizar esse processo foi realizado o alinhamento 2 a 2, totalizando cinco alinhamentos. Mesmo assim, o tempo gasto com o alinhamento variou de 02h45min a 12h.

As correlações obtidas foram muito maiores do que as anteriores, chegando a encontrar uma correlação de até 0,96. Na figura 32 observa-se uma correlação máxima de 0,939 para o bloco C e tamanho de caixa 3. Na figura 33, a maior correlação foi de 0,96 para o bloco CC e também o tamanho de caixa 3. Por último, na figura 34 a maior correlação foi de 0,94 para o bloco ACC e tamanho de caixa 1. Para o cálculo das entropias o maior tempo gasto foi para o tamanho de bloco 3 e tamanho de caixa 3 que correspondeu a 4 minutos.

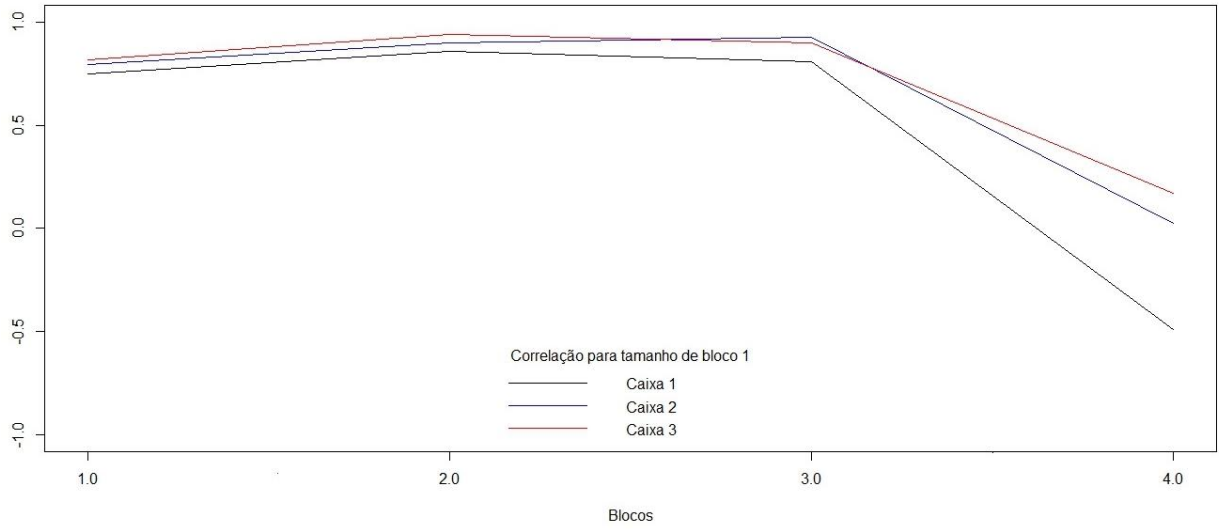

Figura 32: Correlações das distâncias por entropias com tamanho de bloco 1 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas do cromossomo 5 do *Homo Sapiens*

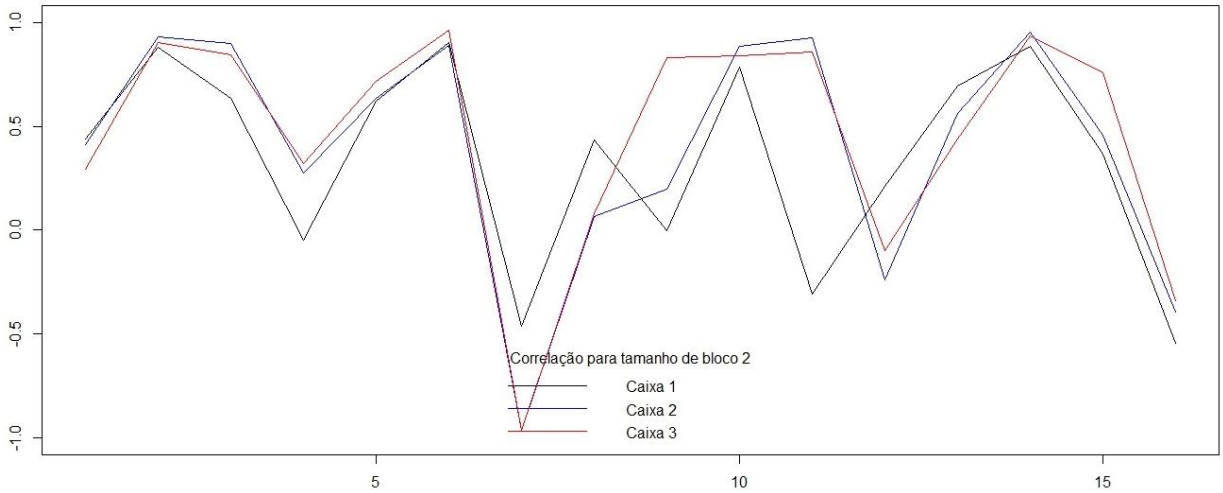

Figura 33: Correlações das distâncias por entropias com tamanho de bloco 2 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas do cromossomo 5 do *Homo Sapiens*

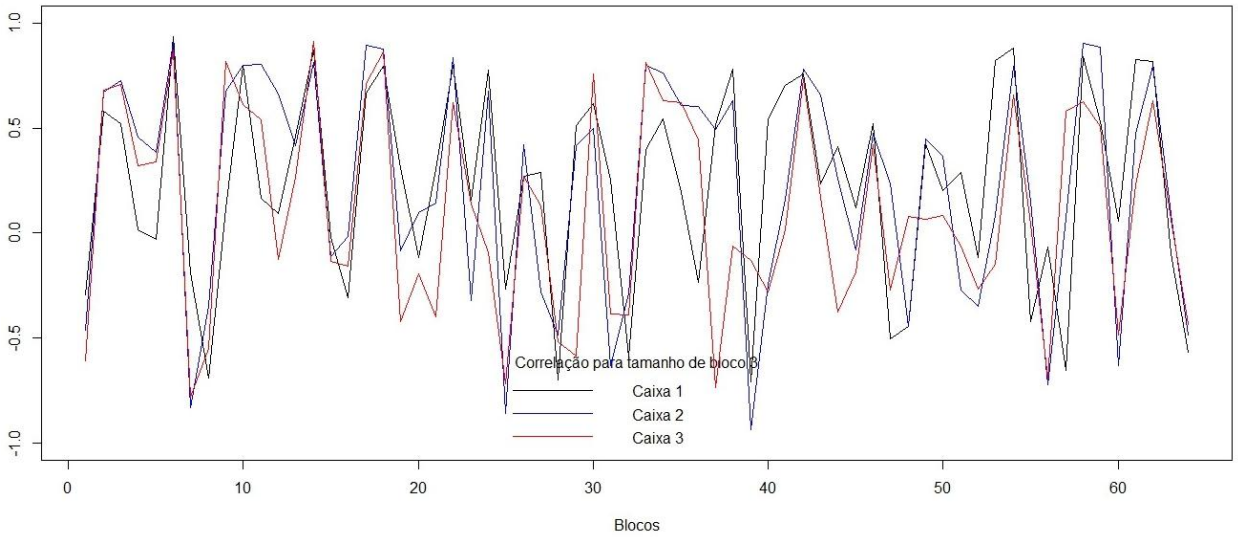

Figura 34: Correlações das distâncias por entropias com tamanho de bloco 3 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas do cromossomo 5 do *Homo Sapiens*

4.6 SIMULAÇÕES

Após obtermos bons resultados para as sequências de DNA utilizando entropia em blocos, testou-se também sua eficiência em sequências aleatórias onde foi determinada a probabilidade de ocorrência de cada base. Para cada tamanho de sequência verificou-se a correlação considerando igual probabilidade de ocorrência das bases, e também considerando probabilidades diferentes de ocorrência. Em cada uma das simulações foram geradas 10 sequências de maneira aleatória.

A primeira simulação realizada foi para sequências de tamanho 400, a maior correlação obtida foi de 0,47 (figura 39) para o bloco de tamanho 2 (GT) e caixa de tamanho 3 considerando as probabilidades diferentes para ocorrências das bases.

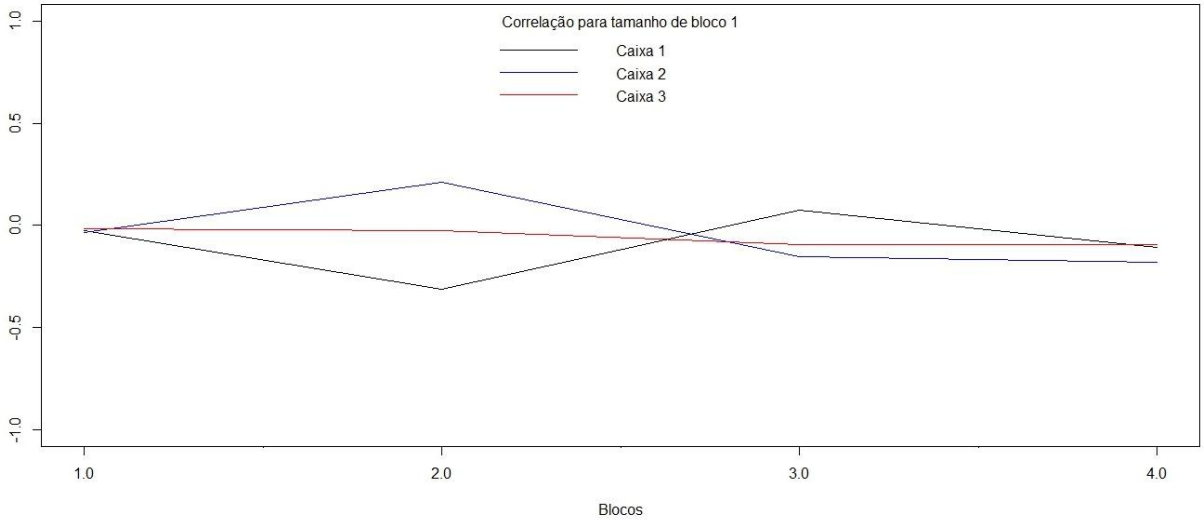

Figura 35: Correlações das distâncias por entropias com tamanho de bloco 1 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 400 e iguais probabilidade de ocorrência das bases

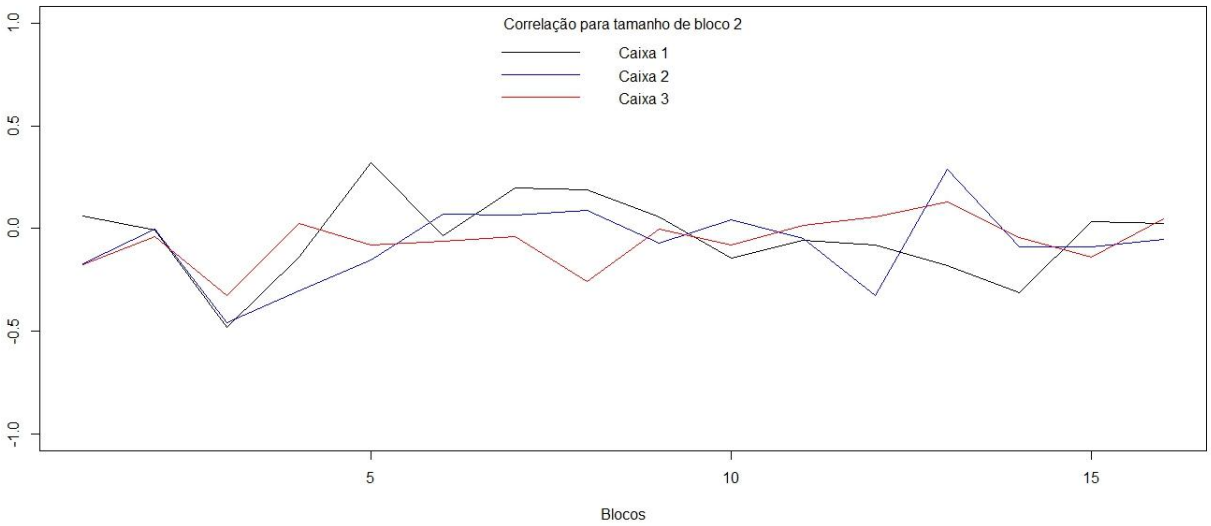

Figura 36: Correlações das distâncias por entropias com tamanho de bloco 2 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 400 e iguais probabilidade de ocorrência das bases

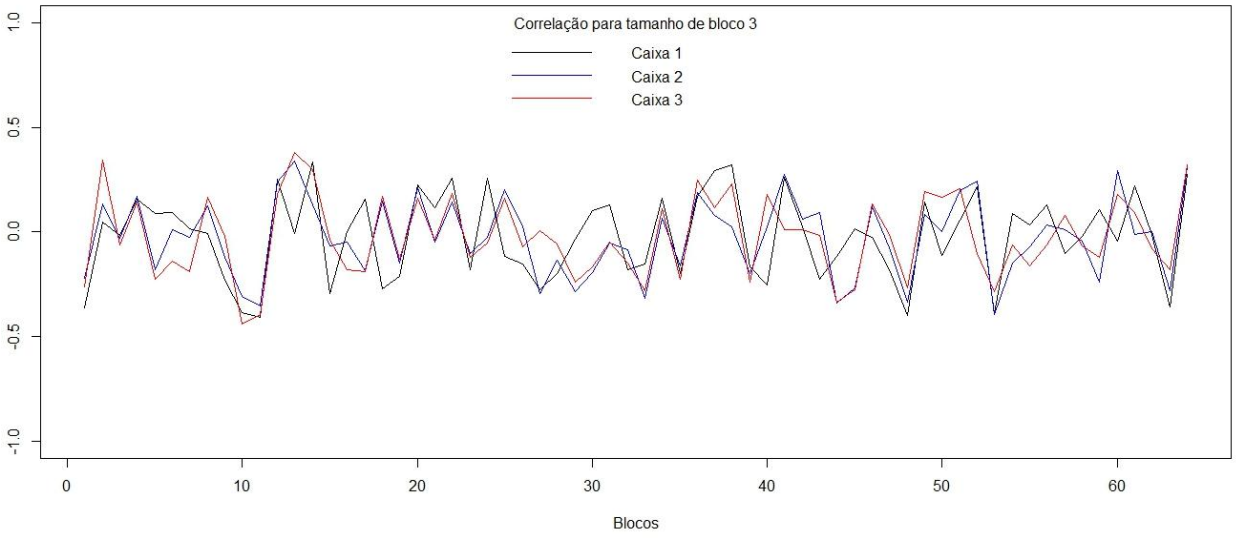

Figura 37: Correlações das distâncias por entropias com tamanho de bloco 3 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 400 e iguais probabilidade de ocorrência das bases

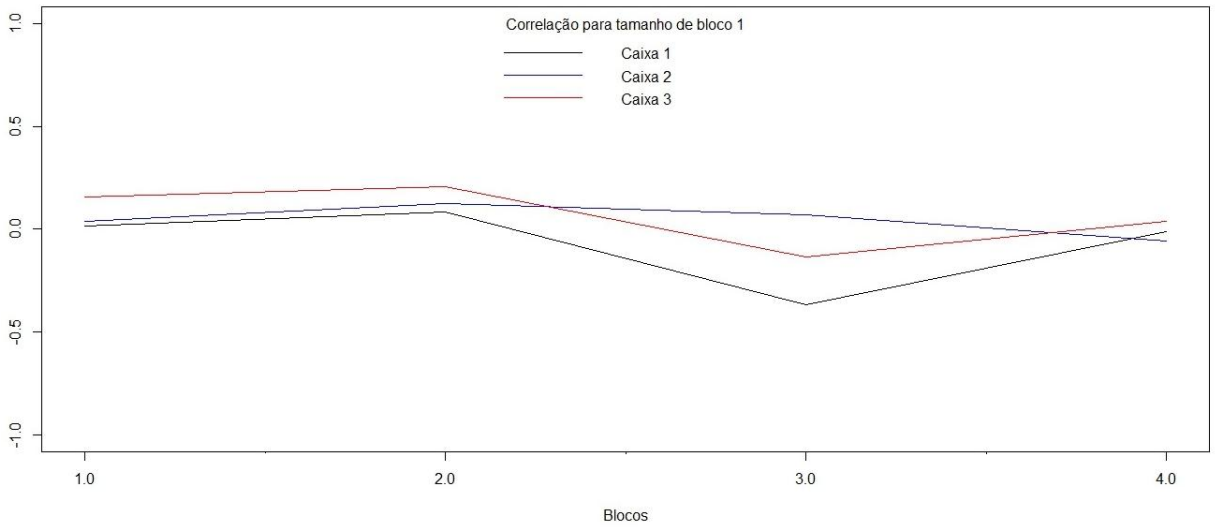

Figura 38: Correlações das distâncias por entropias com tamanho de bloco 1 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 400 e diferentes probabilidades de ocorrência das bases

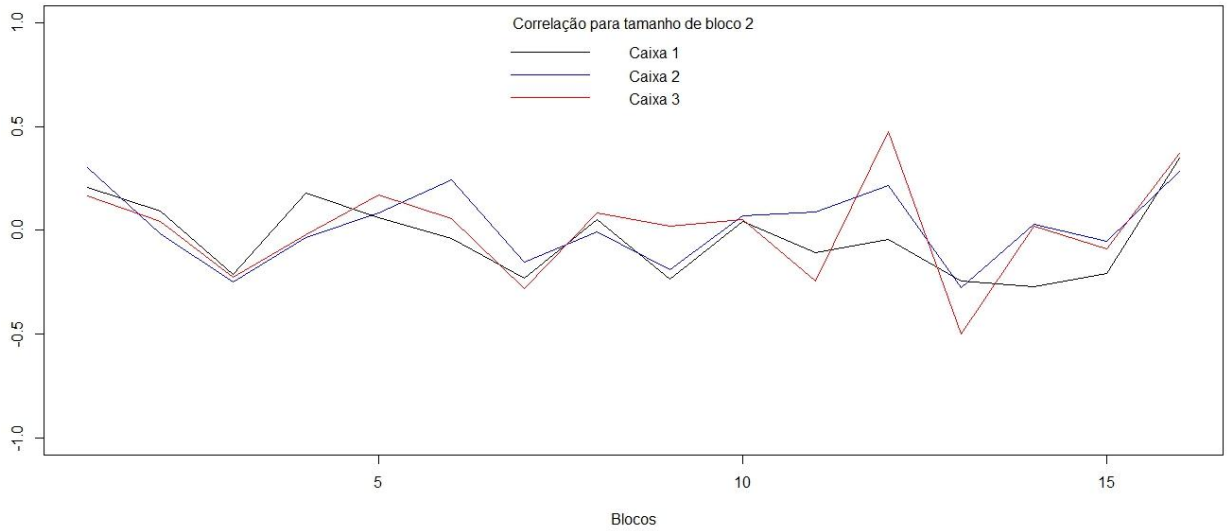

Figura 39: Correlações das distâncias por entropias com tamanho de bloco 2 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 400 e diferentes probabilidades de ocorrência das bases

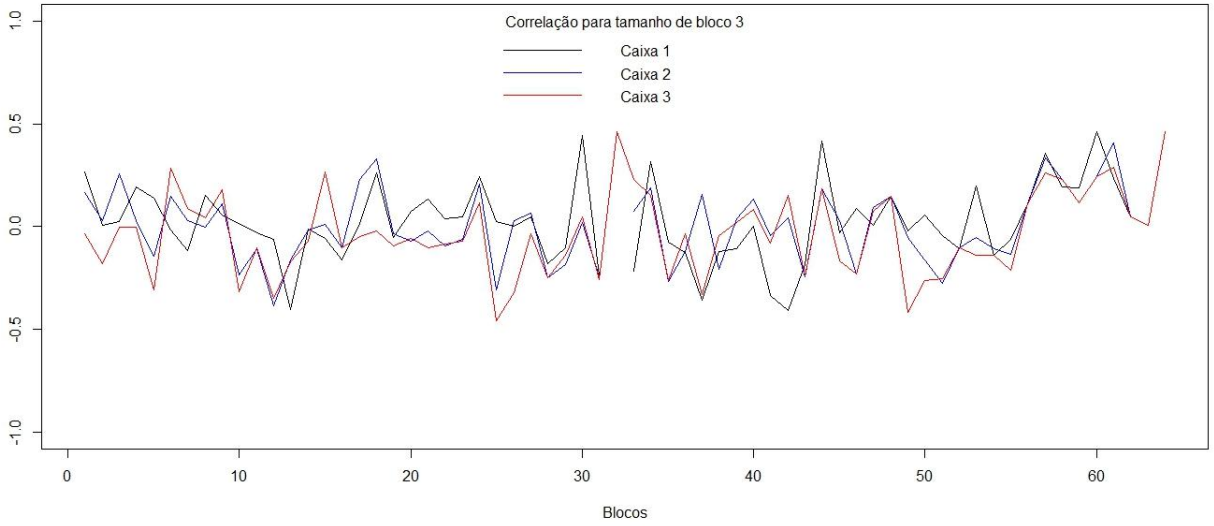

Figura 40: Correlações das distâncias por entropias com tamanho de bloco 3 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 400 e diferentes probabilidades de ocorrência das bases

A segunda simulação realizada foi para sequências de tamanho 2.000, a maior correlação obtida foi de 0,43 (figura 43) para o bloco de tamanho 3 (GCG) e caixa de tamanho 2 considerando as probabilidades iguais para ocorrências das bases.

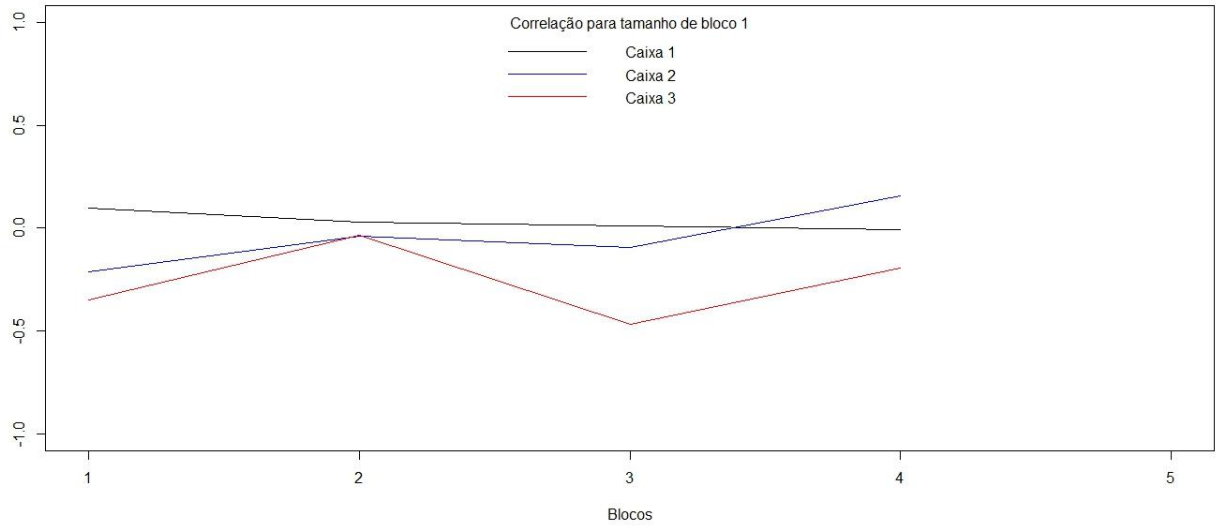

Figura 41: Correlações das distâncias por entropias com tamanho de bloco 1 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 2.000 e igual probabilidade de ocorrência das bases

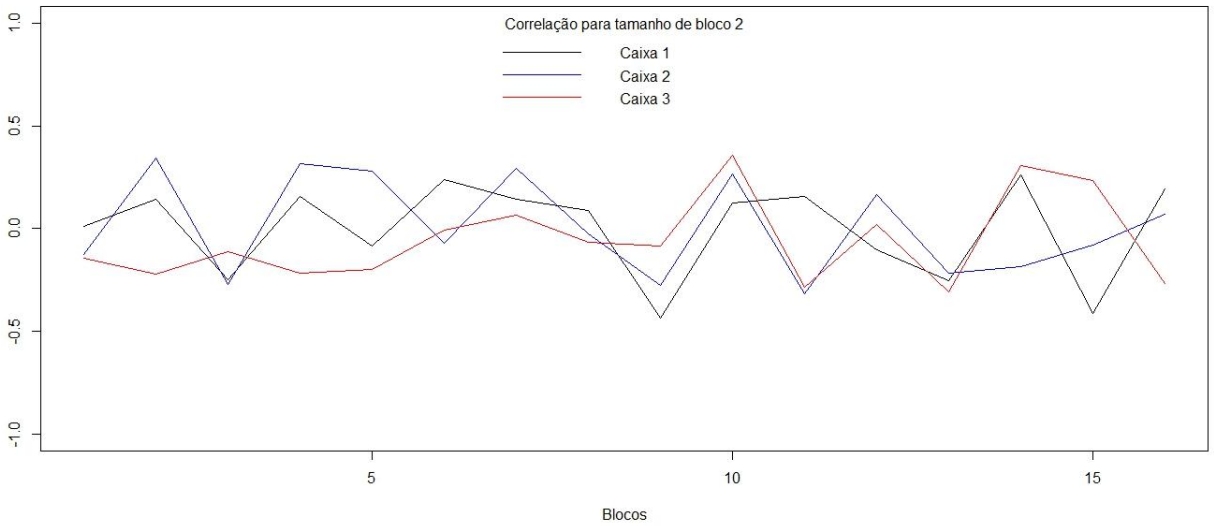

Figura 42: Correlações das distâncias por entropias com tamanho de bloco 2 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 2.000 e igual probabilidade de ocorrência das bases

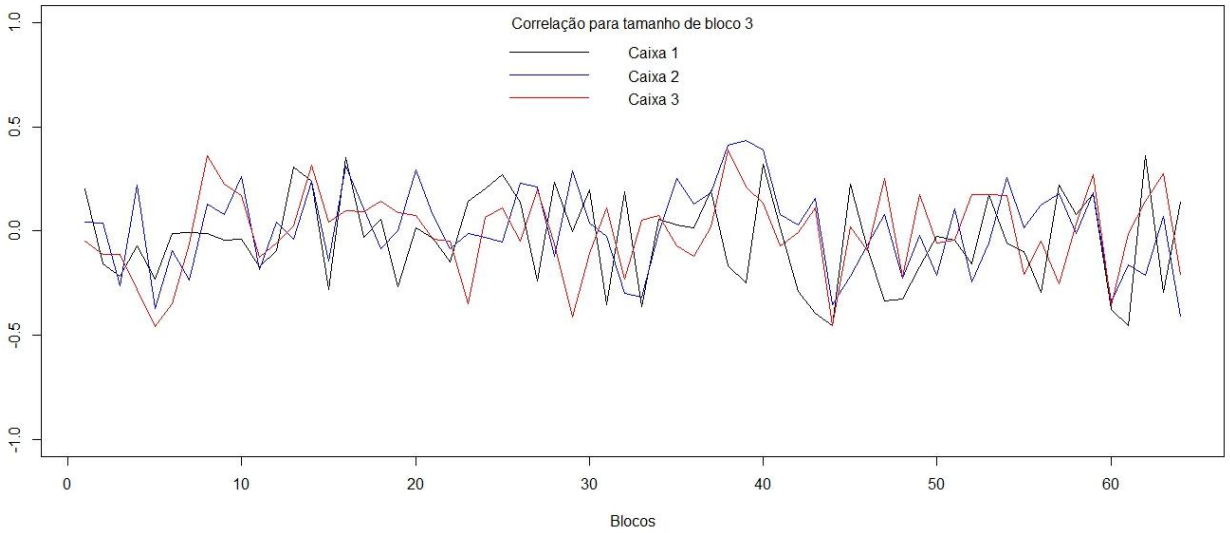

Figura 43: Correlações das distâncias por entropias com tamanho de bloco 3 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 2.000 e igual probabilidade de ocorrência das bases

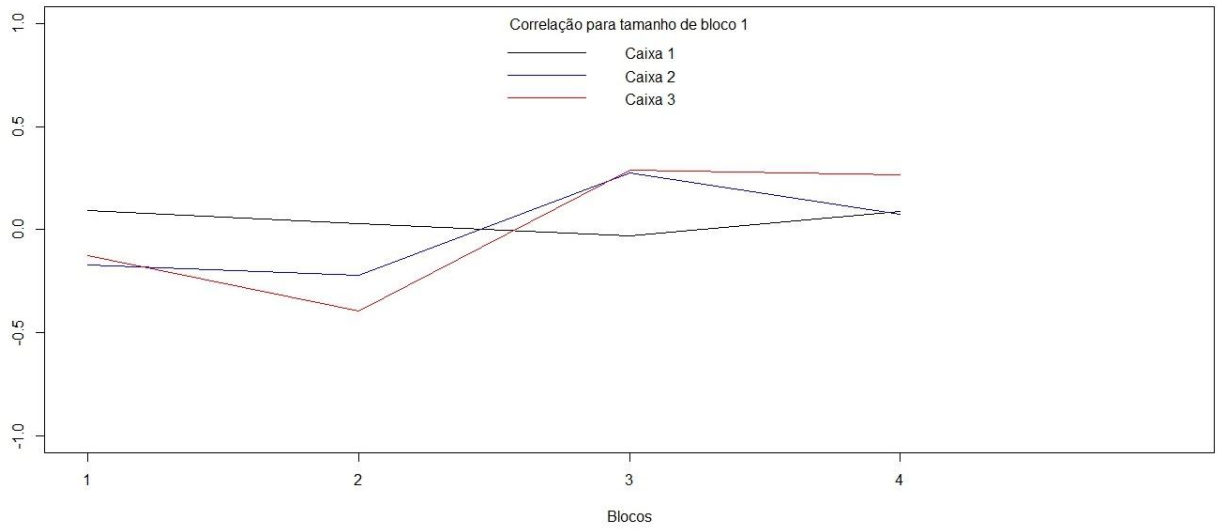

Figura 44: Correlações das distâncias por entropias com tamanho de bloco 1 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 2.000 e diferentes probabilidades de ocorrência das bases

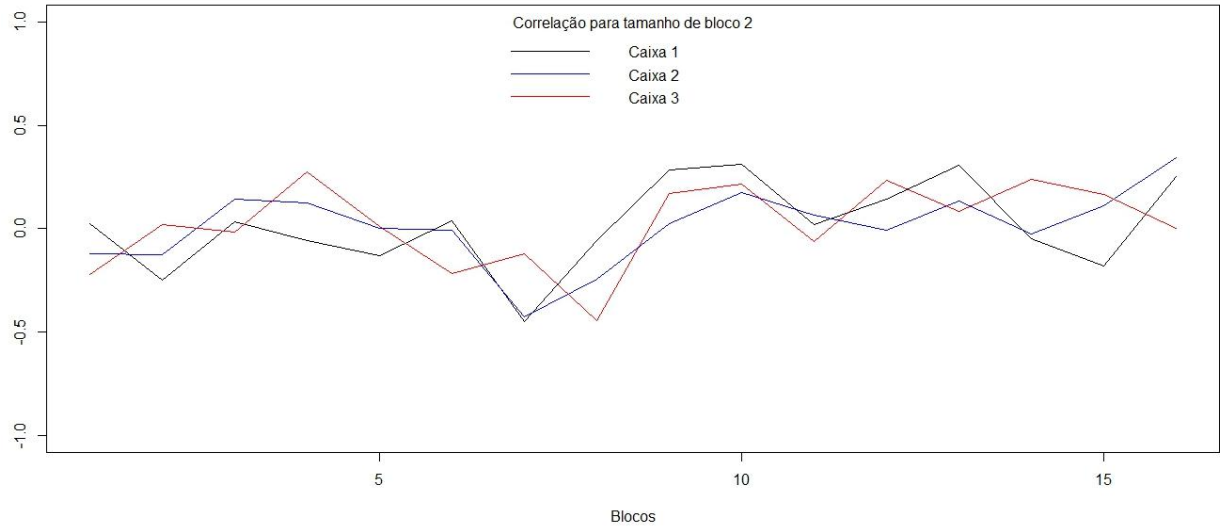

Figura 45: Correlações das distâncias por entropias com tamanho de bloco 2 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 2.000 e diferentes probabilidades de ocorrência das bases

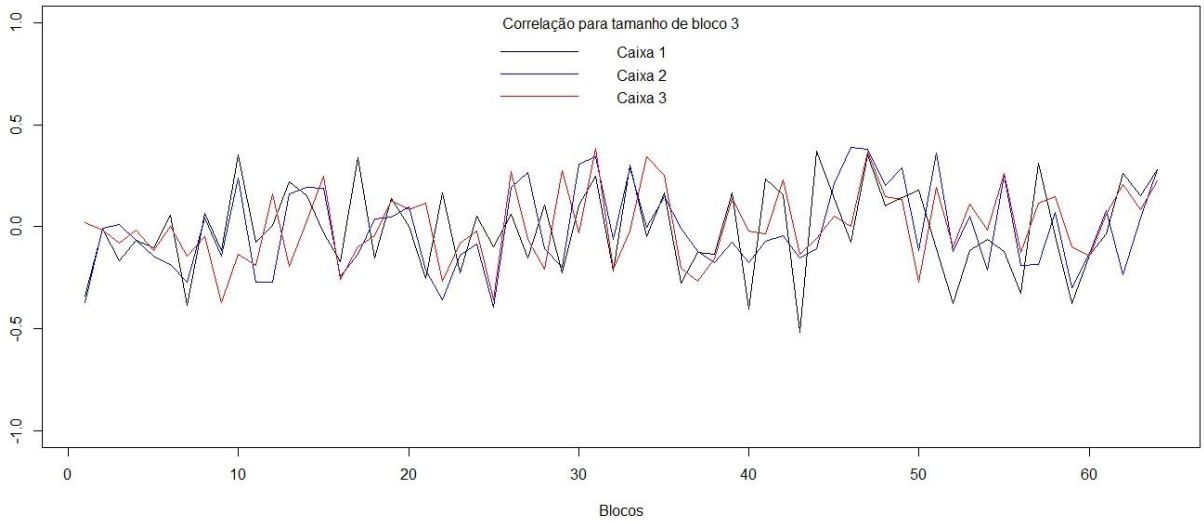

Figura 46: Correlações das distâncias por entropias com tamanho de bloco 3 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 2.000 e diferentes probabilidades de ocorrência das bases

A terceira e última simulação realizada foi para sequências de tamanho 100.000, a maior correlação obtida foi de 0,58 (figura 49) para o bloco de tamanho 3 (CGT) e caixa de tamanho 1, considerando as probabilidades iguais para ocorrências das bases. O tempo de alinhamento das sequências durou 23 horas na 1ª simulação e 72 horas na 2ª simulação, para o cálculo das entropias o tempo máximo durou menos de 1 minuto.

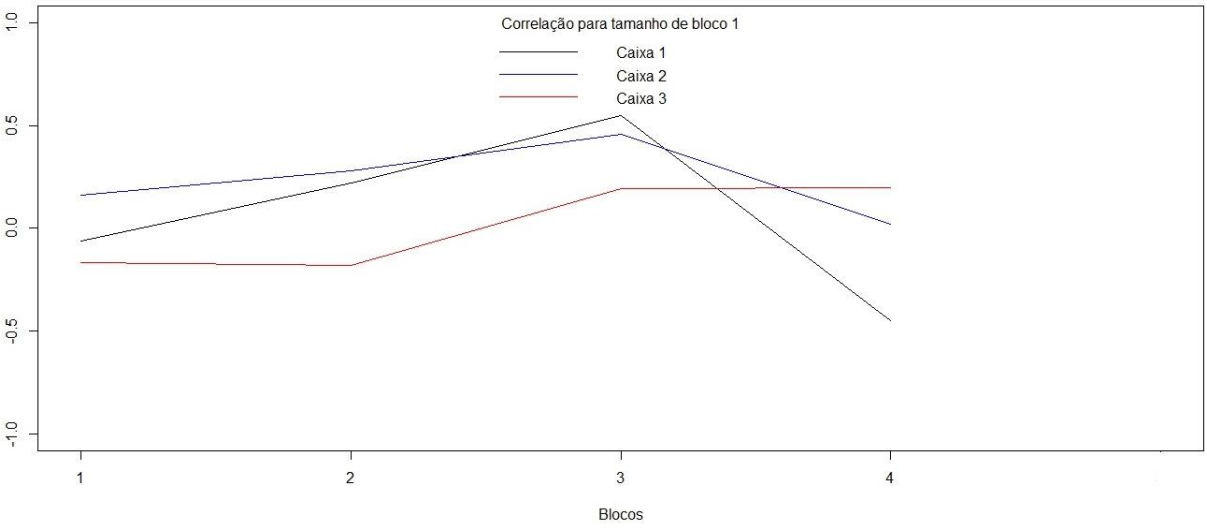

Figura 47: Correlações das distâncias por entropias com tamanho de bloco 1 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 100.000 e igual probabilidade de ocorrência das bases

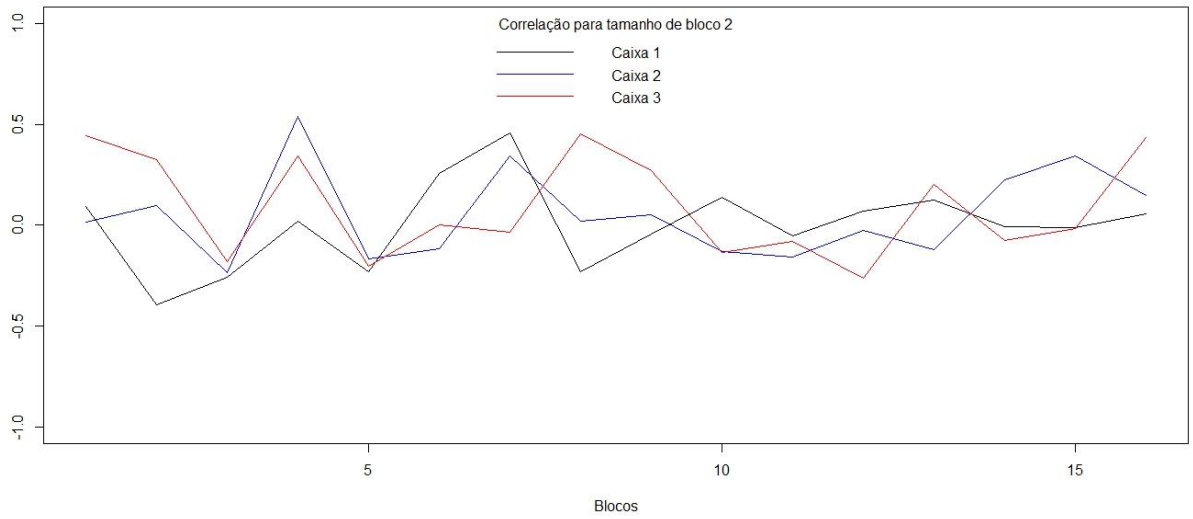

Figura 48: Correlações das distâncias por entropias com tamanho de bloco 2 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 100.000 e igual probabilidade de ocorrência das bases

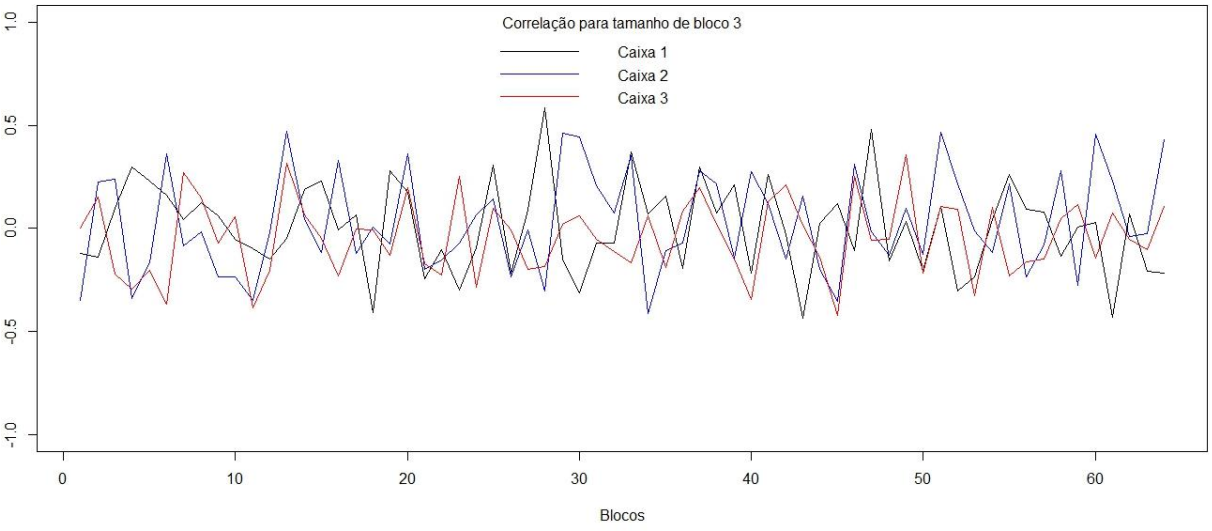

Figura 49: Correlações das distâncias por entropias com tamanho de bloco 3 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 100.000 e igual probabilidade de ocorrência das bases

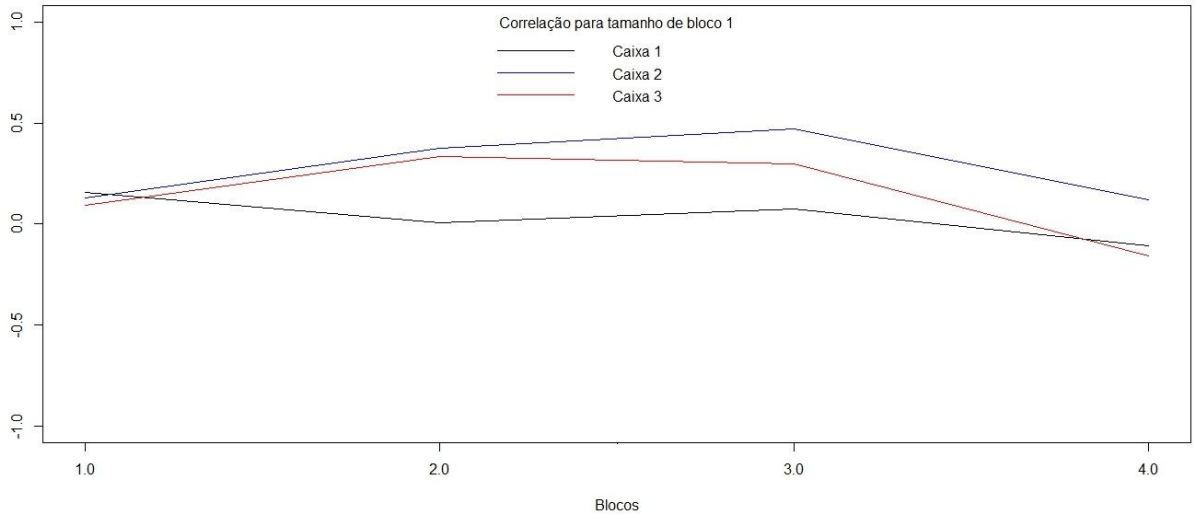

Figura 50: Correlações das distâncias por entropias com tamanho de bloco 1 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 100.000 e diferentes probabilidades de ocorrência das bases

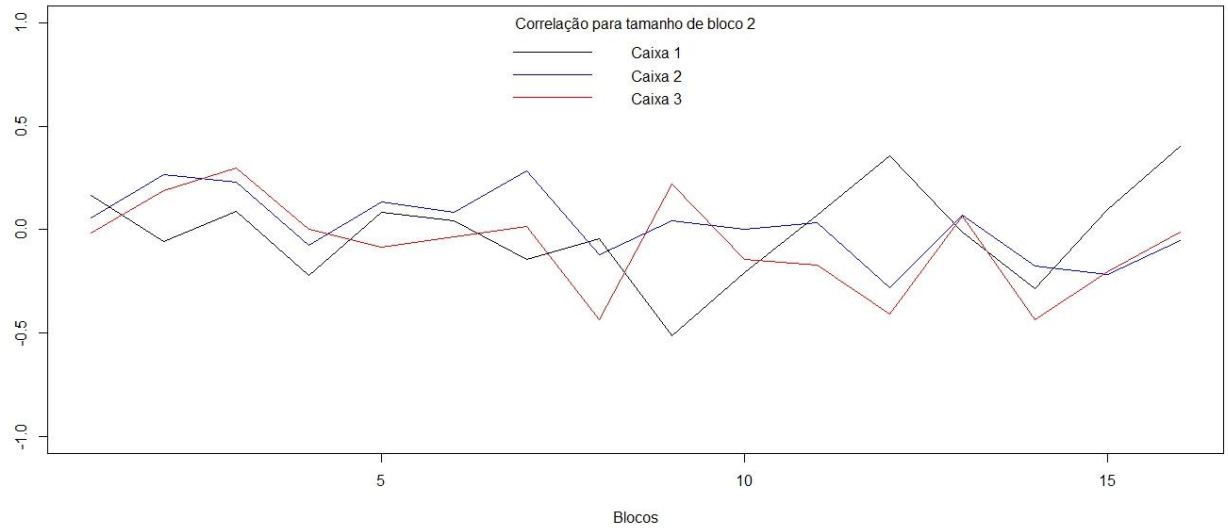

Figura 51: Correlações das distâncias por entropias com tamanho de bloco 2 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 100.000 e diferentes probabilidades de ocorrência das bases

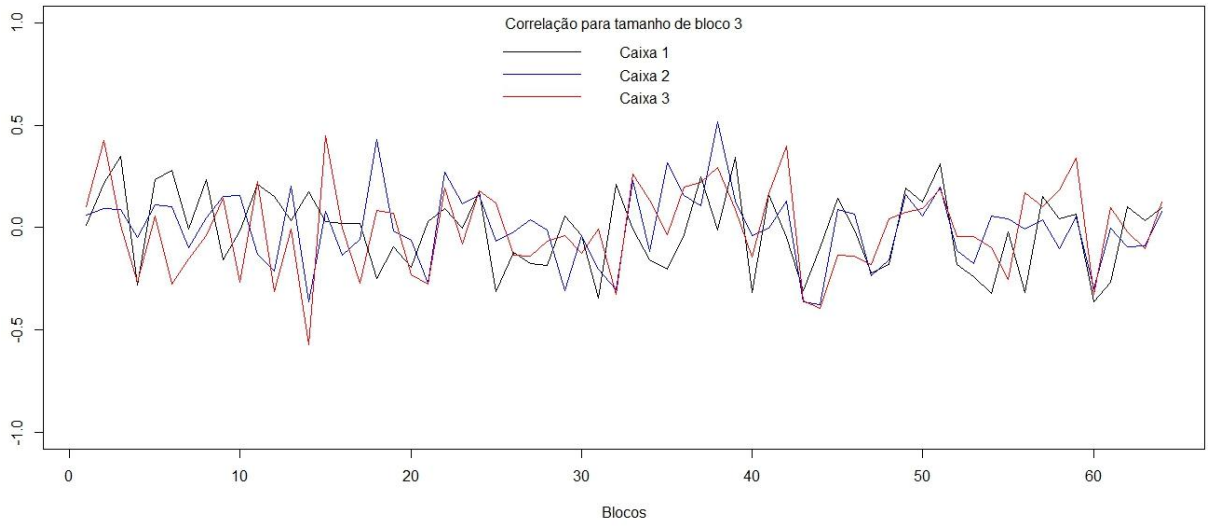

Figura 52: Correlações das distâncias por entropias com tamanho de bloco 3 antes dos eventos e tamanho de caixa de 1 a 3 com as distâncias genéticas para sequências de tamanho 100.000 e diferentes probabilidades de ocorrência das bases

5. CONSIDERAÇÕES FINAIS

Verificou-se que as distâncias por entropia para análise de sequências genéticas têm resultados bem diferentes das distâncias genéticas, exceto quando consideramos blocos, que chamamos nesse estudo de entropia em blocos ou simplesmente entropia condicional.

Utilizando o conceito de entropia condicional observou-se que quanto maior o tamanho das sequências maior é a correlação com as distâncias genéticas. Quando foi analisado o gene BoLA em que o tamanho das sequências variaram de 236 a 393, foi possível identificar uma correlação de 0,5. Quando se analisou as sequências de abelhas sem ferrão *Melipona quinquefasciata* com o tamanho das sequências de 1.869, foi possível identificar uma correlação de 0,61. Quando analisou-se as sequências do cromossomo 5 do *Homo Sapiens* em que o tamanho das sequências variaram de 134.506 a 188.332, identificou-se uma correlação de 0,94. Quando feito as simulações observou-se pouca diferença entre as correlações máxima (0,47 e 0,43) para as sequências de tamanho 400 e 2.000 respectivamente, e assim como no estudos da sequências reais, nas sequências simuladas houve um aumento significativo na correlação para as sequências de tamanho 100.000 que apresentou uma correlação de 0,58.

A maior vantagem do uso da entropia para a análise de sequências é pela sua simplicidade em relação às distâncias genéticas, pois não é necessário fazer os alinhamentos e com isso tem um custo computacional muito menor, ou seja, é possível ter uma análise muita mais rápida.

6. TRABALHOS FUTUROS

Embora ainda tenha muito caminho a ser percorrido, a análise de similaridade de sequências de DNA por entropia apresenta resultados promissores e com muito a ser explorado. Pretende-se pesquisar sobre a identificação do bloco e tamanho de caixa que apresenta a maior correlação com as distâncias genéticas. Com isso, se dará um grande passo para a análise por entropia a partir do momento que não precisar mais dos alinhamentos para efeitos de comparação de resultados.

REFERÊNCIAS BIBLIOGRÁFICAS

BARROS, P.S.N. **Reconhecimento Quântico de Padrões Aplicados à Sequência de DNA.** Recife, 2011. Dissertação de Mestrado em Biometria e Estatística Aplicada, Universidade Federal Rural de Pernambuco.

BUSSAB, W. DE O; MIAZAKI, E. S; ANDRADE, D. **Introdução à análise de agrupamentos.** São Paulo: Associação Brasileira de Estatística, 1990. 105p.

CARPENA, P.; OLIVER, J.L.; HACKENBERG, M.; CORONADO, A.V.; BARTUREN, G. and BERNAOLA-GALVÁN, P. High-level organization of isochores into gigantic superstructures in the human genome, **Physical review**, 83, mar. 2011.

CHARIF, D.; LOBRY, J.R. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis in Structural approaches to sequence evolution: Molecules, networks, populations (U. Bastolla, M. Porto, H.E. Roman and M. Vendruscolo Eds.) **Biological and Medical Physics, Biomedical Engineering.** p. 207-232 (2007).

CLAUDE E. SHANNON. A mathematical theory of communication. **The Bell System Technical Journal**, 27:279–423, 623–656, 1948.

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

EVERITT, B. **Cluster analysis**, London: Heinemann Educational Books, 1974. 136p.

FARRIS, J. S. On the cophenetic correlation coefficient. **Systenatic Biology**, v. 18,p. 279-285, 1969.

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

GRIFFITHS, A. J. F; MILLER J.H; SUZUKI, D.T; LEWONTIN R.C; GELBART W. M. 2000. **An Introduction to Genetic Analysis**. New York: W. H.

IDALINO, R. C. L. **Homologia em Genes Relacionados à Resistência à Mastite em Vacas, Ovelhas e Cabras.** Recife, 2010. Dissertação de Mestrado em Biometria e Estatística Aplicada, Universidade Federal Rural de Pernambuco.

JARI OKSANEN; ROELAND KINDT; PIERRE LEGENDRE; BOB O'HARA; GAVIN L. SIMPSON; PETER SOLYMOS; M. HENRY H. STEVENS; HELENE WAGNER. **Vegan: Community Ecology Package.** 2008.

JONES, S. (1995), **A Linguagem dos Genes, Difusão Cultural** (Tradução de Isabel Mafra), Editora Difusão Cultural.

JUKES, T. H. and C.R.CANTOR. 1969. Evolution of Protein Molecules Mammalian Protein Metabolist. Vol. III. M. N. Munro. **Academic Press**, New York.

KELLY, C. (1994), "A test of Markovian model of DNA evolution", **Biometrics**, 50, 653-664.

KERR, W. E. **Estudos sobre a genética de Melipona.** 1948. 276f. Tese (Doutorado em Genética). Escola Superior de Agricultura Luiz de Queiroz, Piracicaba.

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.

MACHADO, J. A. T.; COSTA A.C.; Quelhas, M.D. Shannon, Rényi and Tsallis entropy analysis of DNA using phase plane, **Nonlinear Analysis: Real World Applications**, 12, 3135-3144 (2011).

MANTEL, B. F. J. The detection of disease clustering and a generalised regression approach. **Cancer Research**, Philadelphia, v. 27, p. 209-220. 1967.

MARIANO-FILHO, J. **Ensaios sobre os meliponidas do Brasil.** 1911, 140 f. Tese (Doutorado em Zoologia). Faculdade de Medicina do Rio de Janeiro, Rio de Janeiro.

MEYER, ANDRÉIA DA SILVA. **Comparação de Coeficientes de Similaridade Usados em Análises de Agrupamento com Dados de Marcadores Moleculares Dominantes.** Piracicaba, 2002. Dissertação de Mestrado em Agronomia, Escola Superior de Agricultura "Luiz Queiroz".

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada.** Belo Horizonte: Ed UFMG, 2005. 297p.

MOURE, J. S. Estudando as abelhas do Brasil (pareceres de sistemática). **Chácaras e Quintais**, São Paulo, v.77, p. 339-341, 1948.

MOURE, J. S. Notas sobre a espécie de Melipona descritas por Lepeletier em 1836. **Revista Brasileira de Biologia.** Rio de Janeiro. v. 35, p. 615-623, 1975.

NCBI, Sequências de DNA de diversos organismos. Disponível em: www.ncbi.nlm.nih.gov. Acesso em: 25 abril 2013.

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.

PIERCE, BENJAMIN A. **Genética: Um Enfoque Conceitual.** Guanabara Koogan S.A., 2004. 758p.

REIS, E. **Estatística multivariada aplicada.** Lisboa: Edições Silabo, 1997. 342p.

RENYI, A. (1976). Some fundamental questions of information theory. Selected Papers of Alfred Renyi, vol. 2, pp. 526–552. **Akademia Kiado**, Budapest.

SCHUWARZ, H. F. The genus Melipona. The type genus of Meliponidae or stingless bees. **Bulletin of the American Museum Natural History**, v. 63, p. 231-459, 1932.

TSALLIS, C . Possible Generalization of Boltzmann-Gibbs Statistics. **J Stat Phys**, v. 52, p. 479-487, 1988.

VIANA, L. S.; MELO, G. A. R. Conservação de abelhas. **Informativo agropecuário.** Aracaju, v. 13, n. 149, p. 23-26, 1987.

VILAÇA, L. F. **Polimorfismo do Gene Bola-DRB3 em Rebanhos Bovinos Leiteiros 5/8 Girolando e Holandês no Estado de Pernambuco.** Garanhuns, 2012. Dissertação de Mestrado em Ciência Animal e Pastagens, Universidade Federal Rural de Pernambuco.

WATERMAN, MICHAEL S. **Introduction Computational Biology.** London: Chapman & Hall, 1995. 431p.

WATSON, J. D.; CRICK, F. H. C. A structure for deoxyribose nucleic acid. **Nature**, v. 171, p. 737-738, 1953

WEIR, B. S. **Genetic data analysis: Methods for discrete population genetic data.** Sunderland, Sinauer Associates, Inc. Publishers, 1990. 377p.

APÊNDICE

Abaixo, são apresentados os algoritmos utilizados nesta dissertação através do software estatístico R 2.15.1

Entropia de Shannon para o gene BoLA

```
library("seqinr")       #carrega o pacote "seqinr"#
dados=read.fasta (file = "dados.fasta")   #ler o arquivo de dados.fasta#
entropia=c()  #cria um vetor nulo#
caixa=1     #tamanho da caixa#
s=c()       #cria um vetor nulo, vai ser o somatório da entropia#
for(i in 1:length(dados)){  #leitura de todas as sequências#
f=count(dados[[i]],caixa) #conta a frequência de cada possível resultado para o determinado tamanho de caixa#
f=f/sum(f)  #calcula a probabilidade de cada possível resultado para o determinado tamanho de caixa#
for(j in 1:length(f)){ #vai percorrer todos os possiveis resultados#
if(f[j]==0){s[j]=0} #Se a probabilidade de um determinado resultado é 0, então nesse determinado somatório receberá 0#
else{
s[j]=f[j]*log(f[j],2)}} #Vai calculando o somatório#
entropia[i]=sum(s)}     #Valor da entropia que é a soma do somatório#
entropia=entropia*(-1)  #Corrigindo o valor da entropia#
entropia=entropia/max(entropia) #Normalizando para um valor entre 0 e 1#
distentropia=dist(entropia) #matriz de distância euclidiana#
banco2=read.table("distjk69.txt",header=T) #leitura da matriz de distância genética (Jukes-Cantor)#
distgen=distentropia #Truque para entender como uma matriz de distâncias#
#esses próximos passos são para preencher a matriz de distância genética com seus respectivos valores#
k=1
n=1
for(j in 1:144){
```

```
for(i in n:144){
distgen[k]=banco2[i,j]
k=k+1}
n=n+1}
library(vegan) #carrega o pacote "vegan"#
mantel(distentropia,distgen) #calcula a correlaçao entre 2 matrizes de similaridades#
summary(entropia) #estatísticas das entropias#
sort(entropia) #ordena as entropias#
agrupamento = hclust(distentropia, method = "average") #Agrupamento pelo método de médias#
plot(agrupamento) #dendograma#
ds=cophenetic(agrupamento) #matriz cofenética#
cor(distentropia,ds) #Coeficiente de correlação cofenético#
sink("shannon1") #Nome do arquivo a ser gravado#
cbind(entropia) #dados que serão gravado#
sink() #Termino da gravação#
grupos=rect.hclust(agrupamento,k=5,border=rainbow(5)) #Quantidade de grupos (k), separado por cores aleatórias#
grupos #impressão dos grupos#
```

Entropia de Rényi para o gene BoLA

```
library("seqinr")
dados=read.fasta (file = "dados.fasta")
entropia=c()
caixa=1
q=0.5
s=c()
for(i in 1:length(dados)){
f=count(dados[[i]],caixa)
f=f/sum(f)
for(j in 1:length(f)){
s[j]=f[j]^q}
entropia[i]=log(sum(s),2)/(1-q)}
entropia=entropia/max(entropia)
```

```
distentropia=dist(entropia)
banco2=read.table("distjk69.txt",header=T)
distgen=distentropia
k=1
n=1
for(j in 1:144){
for(i in n:144){
distgen[k]=banco2[i,j]
k=k+1}
n=n+1}
library(vegan)
mantel(distentropia,distgen)
summary(entropia)
sort(entropia)
agrupamento = hclust(distentropia, method = "average")
plot(agrupamento)
ds=cophenetic(agrupamento)
cor(distentropia,ds)
grupos=rect.hclust(agrupamento,k=5,border=rainbow(5))
grupos
```

Entropia de Tsallis para o gene BoLA

```
library("seqinr")
dados=read.fasta (file = "dados.fasta")
entropia=c()
caixa=1
q=0.5
s=c()
for(i in 1:length(dados)){
f=count(dados[[i]],caixa)
f=f/sum(f)
for(j in 1:length(f)){
s[j]=f[j]^q}
entropia[i]=(1-sum(s))/(q-1)}
```

```
entropia=entropia/max(entropia)
distentropia=dist(entropia)
banco2=read.table("distjk69.txt",header=T)
distgen=distentropia
k=1
n=1
for(j in 1:144){
for(i in n:144){
distgen[k]=banco2[i,j]
k=k+1}
n=n+1}
library(vegan)
mantel(distentropia,distgen)
summary(entropia)
sort(entropia)
agrupamento = hclust(distentropia, method = "average")
plot(agrupamento)
ds=cophenetic(agrupamento)
cor(distentropia,ds)
grupos=rect.hclust(agrupamento,k=5,border=rainbow(5))
grupos
```

Entropia condicional (bloco antes dos eventos) gene BoLA

```
dados=read.fasta (file = "dados.fasta")   #ler o arquivo de dados.fasta#
entropia=c()  #cria um vetor nulo#
library(vegan) #carrega o pacote "vegan"#
a=function(x,y){
co=c()
caixa=x
word=y
dif=caixa-word
for(k in 1:(4^word)){
s=c()       #cria um vetor nulo, vai ser o somatório da entropia#
for(i in 1:length(dados)){  #leitura de todas as sequências#
```

```
f=count(dados[[i]],caixa) #conta a frequência de cada possível resultado para o determinado tamanho de caixa#
f=f[(((4^dif)*k)-((4^dif)-1)):((4^dif)*k)]
if(sum(f)==0){
for(w in 1:length(f)){
f[w]=0}}
else{
f=f/sum(f)}  #calcula a probabilidade de cada possível resultado para o determinado tamanho de caixa#
for(j in 1:length(f)){ #vai percorrer todos os possiveis resultados#
if(f[j]==0){s[j]=0} #Se a probabilidade de um determinado resultado é 0, então nesse determinado somatório receberá 0#
else{
s[j]=f[j]*log(f[j],2)}} #Vai calculando o somatório#
entropia[i]=sum(s)}     #Valor da entropia que é a soma do somatório#
entropia=entropia*(-1)  #Corrigindo o valor da entropia#
entropia=entropia/max(entropia) #Normalizando para um valor entre 0 e 1#
distentropia=dist(entropia) #matriz de distância euclidiana#
co[k]=cor(distentropia,distgen)}
mantel(distentropia,distgen)
agrupamento = hclust(distentropia, method = "average") #Agrupamento pelo método de médias#
plot(agrupamento) #dendograma#
ds=cophenetic(agrupamento) #matriz cofenética#
cor(distentropia,ds) #Coeficiente de correlação cofenético#
grupos=rect.hclust(agrupamento,k=5,border=rainbow(5)) #Quantidade de grupos (k), separado por cores aleatórias#
grupos #impressão dos grupos#
```

Entropia condicional (bloco depois dos eventos) gene BoLA

```
dados=read.fasta (file = "dados.fasta")   #ler o arquivo de dados.fasta#
library(vegan) #carrega o pacote "vegan"#
entropia=c()  #cria um vetor nulo#
```

```
g=c()
f=c()
caixa=3
word=2
dif=caixa-word
for(k in 1:(4^word)){
s=c()       #cria um vetor nulo, vai ser o somatório da entropia#
for(i in 1:length(dados)){  #leitura de todas as sequências#
f=count(dados[[i]],caixa) #conta a frequência de cada possível resultado para o determinado tamanho de caixa#
f=f[(((4^dif)*k)-((4^dif)-1)):((4^dif)*k)]
if(sum(f)==0){
for(w in 1:length(f)){
f[w]=0}}
else{
f=f/sum(f)}  #calcula a probabilidade de cada possível resultado para o determinado tamanho de caixa#
for(j in 1:length(f)){ #vai percorrer todos os possiveis resultados#
if(f[j]==0){s[j]=0} #Se a probabilidade de um determinado resultado é 0, então nesse determinado somatório receberá 0#
else{
s[j]=f[j]*log(f[j],2)}} #Vai calculando o somatório#
entropia[i]=sum(s)}     #Valor da entropia que é a soma do somatório#
entropia=entropia*(-1)  #Corrigindo o valor da entropia#
entropia=entropia/max(entropia) #Normalizando para um valor entre 0 e 1#
distentropia=dist(entropia) #matriz de distância euclidiana#
print(k)
print(mantel(distentropia,distgen))}
agrupamento = hclust(distentropia1, method = "average") #Agrupamento pelo método de médias#
plot(agrupamento) #dendograma#
ds=cophenetic(agrupamento) #matriz cofenética#
cor(distentropia,ds) #Coeficiente de correlação cofenético#
```

```
grupos=rect.hclust(agrupamento,k=5,border=rainbow(5)) #Quantidade de grupos (k), separado por cores aleatórias#
grupos #impressão dos grupos#
```

Entropia condicional (bloco antes dos eventos) Cromossomo 5

```
library(seqinr)
library("seqinr")       #carrega o pacote "seqinr"#
dados=read.fasta (file = "dados.fas")   #ler o arquivo de dados.fasta#
entropia=c()  #cria um vetor nulo#
caixa=6     #tamanho da caixa#
s=c()       #cria um vetor nulo, vai ser o somatório da entropia#
for(i in 1:length(dados)){  #leitura de todas as sequências#
f=count(dados[[i]],caixa) #conta a frequência de cada possível resultado para o determinado tamanho de caixa#
f=f/sum(f)  #calcula a probabilidade de cada possível resultado para o determinado tamanho de caixa#
for(j in 1:length(f)){ #vai percorrer todos os possiveis resultados#
if(f[j]==0){s[j]=0} #Se a probabilidade de um determinado resultado é 0, então nesse determinado somatório receberá 0#
else{
s[j]=f[j]*log(f[j],2)}} #Vai calculando o somatório#
entropia[i]=sum(s)}     #Valor da entropia que é a soma do somatório#
entropia=entropia*(-1)  #Corrigindo o valor da entropia#
entropia=entropia/max(entropia) #Normalizando para um valor entre 0 e 1#
distentropia=dist(entropia) #matriz de distância euclidiana#
gen=c(0.925,0.993,0.976,0.915,0.999)
entropia1=c(entropia[1],entropia[2])
entropia2=c(entropia[3],entropia[4])
entropia3=c(entropia[5],entropia[6])
entropia4=c(entropia[7],entropia[8])
entropia5=c(entropia[9],entropia[10])
distentropia=c(dist(entropia1),dist(entropia2),dist(entropia3),dist(entropia4),dist(entropia5))
cor(gen,distentropia)
```

```
co=c()
caixa=6
word=3
dif=caixa-word
for(k in 1:(4^word)){
s=c()       #cria um vetor nulo, vai ser o somatório da entropia#
for(i in 1:length(dados)){  #leitura de todas as sequências#
f=count(dados[[i]],caixa) #conta a frequência de cada possível resultado para o determinado tamanho de caixa#
f=f[(((4^dif)*k)-((4^dif)-1)):((4^dif)*k)]
if(sum(f)==0){
for(w in 1:length(f)){
f[w]=0}}
else{
f=f/sum(f)}  #calcula a probabilidade de cada possível resultado para o determinado tamanho de caixa#
for(j in 1:length(f)){ #vai percorrer todos os possiveis resultados#
if(f[j]==0){s[j]=0} #Se a probabilidade de um determinado resultado é 0, então nesse determinado somatório receberá 0#
else{
s[j]=f[j]*log(f[j],2)}} #Vai calculando o somatório#
entropia[i]=sum(s)}     #Valor da entropia que é a soma do somatório#
entropia=entropia*(-1)  #Corrigindo o valor da entropia#
entropia=entropia/max(entropia) #Normalizando para um valor entre 0 e 1#
entropia1=c(entropia[1],entropia[2])
entropia2=c(entropia[3],entropia[4])
entropia3=c(entropia[5],entropia[6])
entropia4=c(entropia[7],entropia[8])
entropia5=c(entropia[9],entropia[10])
distentropia=c(dist(entropia1),dist(entropia2),dist(entropia3),dist(entropia4),dist(entropia5))
co[k]=cor(gen,distentropia)}
```