MEDIDAS DE ENTROPIA COM A UTILIZAÇÃO DO BANCO DE DADOS DE FAMÍLIAS DE PROTEÍNAS (PFAM)

Wellington de Souza Vieira

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Rubem Pinto Mondaini

Rio de Janeiro
Dezembro de 2022

MEDIDAS DE ENTROPIA COM A UTILIZAÇÃO DO BANCO DE DADOS DE FAMÍLIAS DE PROTEÍNAS (PFAM)

Wellington de Souza Vieira

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Rubem Pinto Mondaini, D.Sc.

Prof. Argimiro Resende Secchi, D.Sc.

Prof. Fernanda Duarte Vilela Reis de Oliveira, D.Sc.

Dr. Simão Coutinho de Albuquerque Neto, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
DEZEMBRO DE 2022

Vieira, Wellington de Souza

Medidas de Entropia com a Utilização do Banco de Dados de Famílias de Proteínas (Pfam)/Wellington de Souza Vieira. – Rio de Janeiro: UFRJ/COPPE, 2022.

XXVI, 164 p.: il.; 29, 7cm.

Orientador: Rubem Pinto Mondaini

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2022.

Referências Bibliográficas: p. 156 – 164.

1. Medidas de Entropia. 2. Domínios de Proteínas. 3. Classificação de Proteínas. I. Pinto Mondaini, Rubem. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Às três mulheres da minha vida:
minha mãe Maria da Graça,
minha esposa Fabiana e minha
filha Maria Eduarda.

Agradecimentos

Aos meus pais, por toda ajuda possível, e em especial a minha avó, Ana Mattos Vieira, pois foi quem participou ativamente na minha educação no ensino fundamental. Sou eternamente grato.

Ao meu orientador, Professor Rubem Mondaini, pela oportunidade concedida de ingressar no curso de mestrado, por sua experiência como pesquisador e por sua análise construtiva que me apoiou muito na construção desta dissertação.

Aos amigos com quem convivi durante o mestrado, Edgar, Ramon e Simão, pelo apoio e companheirismo. Ao Simão em especial por toda a ajuda com as muitas dúvidas que tive durante a realização do trabalho, como por todas as sugestões de texto para melhorar várias partes dessa dissertação.

Aos meus amigos de convívio pessoal, em especial ao Luan.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

MEDIDAS DE ENTROPIA COM A UTILIZAÇÃO DO BANCO DE DADOS DE FAMÍLIAS DE PROTEÍNAS (PFAM)

Wellington de Souza Vieira

Dezembro/2022

Orientador: Rubem Pinto Mondaini

Programa: Engenharia de Sistemas e Computação

Com o avanço da tecnologia houve a necessidade de obter um sequenciamento e a catalogação de proteínas, quando identificadas em famílias reunidas em clãs. Os Domínios de Proteínas são representados pela homologia em suas sequências, cuja estrutura e funções podem ser confirmadas por técnicas de Cadeias Ocultas de Markov e Machine Learning. Este trabalho tem como propósito verificar de forma introdutória esse alinhamento através de métodos estáticos alternativos interligados à distribuição de valores de funções (Medidas de Entropia) de variáveis aleatórias (probabilidade de ocorrência de aminoácidos nos domínios de proteínas) que mostram a existência das famílias e clãs extraídos de blocos retangulares (m linhas e n colunas).

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

ENTROPY MEASURES USING THE PROTEIN FAMILIES DATABASE (PFAM)

Wellington de Souza Vieira

December/2022

Advisor: Rubem Pinto Mondaini

Department: Systems Engineering and Computer Science

With the advancement of technology there was the need to obtain a sequencing and cataloging of proteins, when identified in families gathered in clans. Protein Domains are represented by homology in their sequences, whose structure and functions can be confirmed by Hidden Markov Chains and Machine Learning techniques. This work aims to verify in an introductory way this alignment through alternative static methods linked to the distribution of values of functions (Entropy Measures) of random variables (probability of occurrence of amino acids in protein domains) that show the existence of families and clans extracted from rectangular blocks (m rows and n columns).

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

A descoberta da estrutura do DNA foi um grande fato revolucionário que mudou a vida dos seres humanos. Houve uma grande evolução da vida em termos de diversidade de espécies, propriedades e atributos específicos de um espécime e funcionalidade de novas habilidades, assim como a capacidade de reprodução e transmissão de características hereditárias. Estas estão diretamente relacionadas à evolução da complexidade e estrutura de unidades microscópicas denominadas proteínas. Novas estruturas estáveis de proteínas podem acarretar tanto em vantagens (curas e/ou tratamento) quanto em desvantagens (doença e/ou morte) para as populações das diferentes espécies, a molécula do DNA tem a estrutura de uma dupla hélice, uma descoberta que daria novos rumos à ciência [9].

O aparecimento e a evolução das proteínas deram-se a partir de pequenas cadeias de aminoácidos que progressivamente aumentaram de tamanho e/ou se uniram dando origem a diversos polipeptídeos capazes de realizar diferentes funções celulares. As proteínas são formadas por uma ou mais estruturas independentemente estáveis, denominadas domínios.

Diversos grupos da comunidade científica vêm identificando e armazenando informações sobre as proteínas em bancos de dados de proteínas ao longo das últimas décadas. Um exemplo é o Pfam, uma base de dados de famílias de proteínas. Os domínios cujas funções e sequências de aminoácidos são identificados por homologia, são aglutinados em famílias. A classificação em famílias implica que os domínios podem possuir algum tipo de relação evolutiva mútua: ou se originam de um ancestral em comum, ou alguns destes são oriundos de outros da mesma família, que por sua vez tiveram sua origem a partir de mais outros também da mesma família, como um grande ramo de uma árvore genealógica [3].

A evolução tecnológica proporcionou inúmeros avanços na área biológica. Uma grande descoberta nas pesquisas genéticas foi obtida por conta do Projeto do Genoma Humano, o programa criado para desvendar a sequência completa do DNA humano. Uma nova fase nas pesquisas genéticas iniciou-se com o reconhecimento das informações inerentes ao genoma humano, o qual tornou-se factível com a contribuição de métodos computacionais.

Essa junção entre a Biologia e a Computação, proporcionou o surgimento de novas técnicas de estudos da Biologia Molecular. Diversos pesquisadores subdividiram os estudos de Biologia Molecular em dois importantes segmentos: a Bioinformática e a Biologia Computacional [10]. A Bioinformática, analisa os problemas referentes à produção de grandes massas de dados, em sua abordagem de como tratar e armazenar os dados. A Biologia Computacional, relata a compreensão e a análise dos dados genômicos criados. Atualmente, a maior adversidade na Biologia Molecular é analisar e entender o grande volume de dados de sequências e de estruturas de DNA e de proteínas que são geradas pelos projetos em elaboração pela área.

Devido à imposição e busca por uma melhor gestão e armazenamento de domínios de proteínas, que eram armazenados em sistemas de arquivos antes do surgimento do Banco de Dados Biológicos(BDB) na segunda metade da década de 1980 [11], foram criados e implementados sistemas específicos para manipular essas informações, juntamente com o uso de bancos de dados tradicionais. Os bancos de dados biológicos contêm não apenas informações inerentes a proteínas, mas também contemplam os alinhamentos de sequências de proteínas e suas pontuações (*scores*).

Na atualidade existem muitas variações de bancos de dados biológicos disponíveis na Internet e muitas ferramentas utilizadas para consultar os conjuntos de dados. Por exemplo, o GenBank que é um banco de dados de bases de nucleotídeos que faz parte de uma rede de colaboração junto com o European Molecular Biology Laboratory (EMBL) e o DNA DataBank of Japan (DDBJ). Há outros como o PDB e o Pfam, que são bancos de dados de proteínas disponíveis na Web para pesquisa, possibilitando o acesso de pesquisadores das mais diversas áreas diretas e indiretas, auxiliando portanto no diagnóstico e tratamento de doenças.

A quantidade de dados biológicos disponíveis em várias fontes de dados vem crescendo de forma bastante acelerada, o que demanda não apenas novas tecnologias para a consulta e manipução dos dados gerados, mas também o desenvolvimento de novos métodos e técnicas de análise que sejam eficientes e eficazes na manipulação das grandes massas de dados. Assim, a adversidade para tratar esses volumes de

dados não podem ser resolvidos apenas com computadores mais velozes e com uma capacidade maior de tratamento dos processos, mas através do desenvolvimento de estruturas de dados e de algoritmos mais inteligentes e sólidos que analisem a complexidade da estrutura dos bancos de dados referentes às proteínas que estão disponíveis na internet.

O banco de dados do Pfam não é tão diferenciado dos bancos de dados tradicionais existentes no mercado de Tecnologia da Informação (TI). O seu propósito é armazenar grandes massas de dados geradas pela comunidade científica, bem como as sequências de proteínas e as classificações definidas a partir da caracterização de famílias de proteínas (conjuntos de sequências de proteínas que possuem um determinando grau de homologia entre si) e superfamílias, também denominadas de clãs.

Existem diversos fundamentos, expostos na literatura [12], utilizados pela comunidade científica para classificar os domínios de proteínas em famílias e clãs, além do reconhecimento de novas proteínas e de regras de construção de novas sequências. Atualmente, os bancos de dados de proteínas são criados e mantidos de acordo com esse padrão de classificação.

1.2 Objetivos

O presente trabalho apresenta os resultados de um método de averiguação da catalogação de famílias de domínios de proteínas em clãs na base de dados Pfam. A identificação de uma ancestralidade auxilia no reconhecimento de um processo evolutivo e consequentemente na identificação das estruturas estáveis, informações necessárias para o passo posterior de determinar novas proteínas com funções específicas para auxiliar, por exemplo, no combate de determinadas doenças, ou seja, estabelecer um processo de engenharia genética [3].

O principal objetivo é analisar regiões específicas destas famílias extraídas do banco de dados Pfam na versão 27.0, com o propósito de verificar essa classificação dos bancos de dados de proteínas. Cálculos envolvendo distribuições de probabilidades simples e conjunta dos aminoácidos nas famílias pertencentes a clãs do banco de dados Pfam foram realizados. Há várias medidas de entropia utilizadas para melhor descreverem o comportamento do bancos de dados Pfam, por exemplo, a de Sharma-Mittal [13] e a de Havrda-Charvat [14] que foram utilizadas neste trabalho minuciosamente, no capítulo 5 desta pesquisa. Então foi possível otimizar a precisão das atuais famílias no banco de dados analisado, através das entropias citadas neste parágrafo, assim

como fazer a avaliação de novos agrupamentos nos domínios de famílias que puderam ser analisados.

Para a realização dos cálculos de entropia, foi utilizado um conjunto de 166 clãs com um total de 2557 famílias. Esse conjunto foi obtido após a adoção de duas restrições impostas ao banco de dados Pfam. A primeira restrição é em relação à construção de blocos representativos de 80 linhas por 80 colunas (um total de 6400 aminoácidos) por família. Somente as famílias pertencentes a clãs que possam ser representadas por esses blocos 80×80 são consideradas, as que não atendem a restrição são descartadas. A segunda restrição, feita após a primeira, se refere ao número mínimo de famílias de um clã para que este seja considerado para testes estatísticos. Adotamos o valor mínimo de cinco famílias por clã. Desta forma, somente os clãs que possuam um mínimo de cinco famílias representadas por blocos 80×80 foram utilizados para o cálculo de entropia.

1.3 Composição do Trabalho

Este trabalho está dividido em sete capítulos. O primeiro capítulo apresenta a motivação, os objetivos e a estruturação do trabalho. Os demais capítulos desta dissertação estão estruturados conforme os parágrafos abaixo.

O capítulo 2 aborda a identificação da molécula de DNA, juntamente à definição de sua estrutura, funcionalidade e composição.

O capítulo 3 relata o que são proteínas, suas funções e estruturas. Descreve também a classificação de proteínas em famílias e clãs com especificação de domínios.

O capítulo 4 menciona o surgimento das bases de dados de proteínas com generalidade. Faz uma introdução ao Protein Family Database (Pfam) e sua classificação dos dados na versão 27 e finaliza apresentando o procedimento de extração das famílias de proteínas.

O capítulo 5 mostra um estudo introdutório de probalidades com a construção do espaço de probabilidades e a definição das "janelas $m \times n$". Tratamos também da entropia de Havrda-Charvat e seus histogramas correspondentes. Realizamos uma análise comparativa dos histogramas da entropias Havrda-Charvat (HC) com outras janelas de trabalhos anteriores realizados realizados pelo grupo de pesquisa.

O capítulo 6 apresenta os desafios encontrados no trabalho de pesquisa descrito nessa dissertação. Também é relatado o uso de ferramentas computacionais e operacionais como apoio para a resolução de problemas matemáticos.

No capítulo 7 são colocadas as considerações finais deste trabalho e apresenta propostas para trabalhos futuros.

Capítulo 2

A Molécula de DNA. De sua descoberta à determinação de sua composição e estrutura

Em 1953, os pesquisadores Francis Crick e James D. Watson revolucionaram a bioquímica ao decifrar a estrutura do DNA, que nada mais é que um composto orgânico que possui moléculas com as instruções genéticas que orquestram o desenvolvimento principal de todos os seres vivos e de alguns vírus. A descoberta foi um divisor de águas não apenas na área científica, mas também para compreensão da própria base da vida. Com essa descoberta o número de pesquisas na área cresceu muito consideravelmente, o que gerou uma explosão na bioquímica que transformou a ciência. Após nove anos de sua descoberta Crick e Watson foram agraciados com o Prêmio Nobel de Medicina [9].

Um passo muito significante para a compreensão do funcionamento dos genes foi a identificação de sua natureza química, o que ocorreu no início da década de 1950, quando se descobriu que os genes são constituídos por DNA. A sigla DNA, que significa ácido desoxirribonucleico, se tornou amplamente reconhecida nas últimas cinco décadas devido a inúmeros progressos científicos, entre os quais: a determinação de sua estrutura molecular, a descoberta do código genético, a descoberta de como podemos controlar o funcionamento e desenvolvimento de análises e a manipulação de técnicas; por esse motivo foram abertos novos campos e centros de pesquisas e tecnologias para estudar o DNA.

Considerando que o DNA é o guia de instrução das espécies, de onde surgiu a vida? Para uns, de uma força vital emanada de Deus. Para outros, a vida ainda era algo incompreensível pela ciência, e ainda para outros, a vida se perpetuava devido a um código secreto, ainda inexplicável, porém capaz de causar tamanho

fascínio. Na busca de explicar o inexplicável, a ciência por meio de estudos científicos iniciados por Erwin Schrödinger, que formulara a hipótese da vida ter tido seu princípio a partir de estruturas portadoras de informações genéticas denominadas proteínas (macromoléculas formadas por micromoléculas, denominadas aminoácidos, que, por sua vez, aparecem associadas em sequência como os elos de uma corrente), que poderiam ser o ponto crucial para decifrar as informações que sustentavam a diversidade da vida.

A proposição do modelo de dupla hélice da estrutura do DNA foi um acontecimento extremamente importante para a Genética, porque promoveu a disseminação e o desenvolvimento da Biologia Molecular. Esse termo foi proposto por Warren Weaver, o então diretor da divisão da Divisão de Ciências Naturais da Fundação Rockfeller, em um de seus relatórios publicados na revista Science em 1938. No relatório o termo descrevia como os fenômenos biológicos podem ser compreendidos fundamentalmente pelo conhecimento das estruturas das moléculas e das interações destas, e gradualmente foi sendo utilizado para designar mais especificamente as pesquisas relacionadas aos genes [15].

Inúmeros exemplares de livros já foram elaborados e escritos sobre a história do modelo da dupla hélice do DNA, além de muitas centenas de artigos em periódicos internacionais de grande impacto e de revistas científicas renomadas. Também não é novidade que alguns desses autores conduziram particularmente a evolução da Biologia Molecular, assim como muitos tiveram participação na construção deste fato científico. É possível também observar que muitos eventos deste episódio histórico foram fundamentais para a Biologia, porque envolveram, além dos conhecimentos biológicos, conhecimentos interdisciplinares de muitas áreas, tais como Química e Física.

A descoberta da molécula de DNA começa no final da década de 1860, com a chegada do médico suíço Friedrich Miescher (1844-1895) à Universidade de Tübingem, na Alemanha. Nesta época surgiram as ideias a respeito das origens e funções das células, pois a teoria da geração espontânea fora derrubada pelos experimentos diligentes conduzidos por Francesco Redi em 1668 e por Louis Pasteur em 1862. Paralelamente a estes fatos, a teoria celular se tornava um dos pilares da Biologia [9].

Com diversos experimentos, Miescher adquiriu inesperadamente algo diferente de todas as substâncias proteicas já identificadas e classificadas até então. Com grande maestria ele descobriu que tal substância se concentra no núcleo das células (parte estrutural da célula ainda pouco estudada na época). Os resultados de suas análises

mostraram que as quantidades relativas dos elementos de hidrogênio (H), carbono (C), oxigênio (O) e nitrogênio (N) presentes nesta nova substância eram divergentes das encontradas nas proteínas e nela continha o elemento fósforo (P), ausente em moléculas de proteínas. A esta inovadora substância denominou-a nucleína, pelo fato de estar concentrada no núcleo das células [16].

Muitos cientistas tinham hipóteses diferentes das de Miescher, que só foram abandonadas por volta de 1889, quando Richard Altmann (1852–1900), obteve preparações purificadas de nucleína, sem nenhuma contaminação por proteínas, e que por ter um caráter ácido, em vez de se chamar nucleína, passou a ser chamada de ácido nucleico [17].

Um grande pesquisador que também merece seu destaque quanto à descoberta do ácido nucleico é Albrecht Kossel (1853–1927) que, dentre os outros produtos da degradação do ácido nucleico, detectou duas bases nitrogenadas, a adenina e a guanina. Em 1893, identificou a timina pela degeneração da célula do timo e em seguida a citosina. No ano de 1894, em união com outros pesquisadores, percebeu que os ácidos nucleicos continham pentose, um açúcar com cinco átomos de carbono [18].

Por volta de 1909, Phoebis Levine e Walter Jacobs determinaram a organização das moléculas de fosfato, de pentose e de bases nitrogenadas no ácido nucleico. Estes três componentes estão interligados entre si formando uma unidade essencial, o nucleotídeo. Com ácidos nucleicos e proteínas no núcleo, Levine e muitos de seus assistentes estavam convencidos de que estas complexas e abundantes moléculas de proteínas, e não o DNA, armazenavam todas as informações genéticas nos cromossomos, sendo atribuída como função do DNA simplesmente manter ligadas as moléculas de proteínas. No ano de 1930, Levine e seus assistentes classificaram dois ácidos nucleicos: o RNA (ácido ribonucleico) e o DNA (ácido desoxirribonucleico) [19].

A função do DNA na concepção de Levine foi corrigida por pesquisas de um bacteriologista, Frederick Griffith (1877–1941) que, em virtude de um grande fenômeno descoberto através da transformação bacteriana, iniciou todo o processo de identificação do DNA como material hereditário em 1928. Suas pesquisas foram conduzidas utilizando a bactéria *Diplococcus pneumoniae*, atualmente conhecida como *Streptococcus pneumoniae*, causadora da pneumonia em seres humanos e em outros mamíferos, o que o levou a descobrir que certa substância não identificada presente nas células de uma variedade de pneumococos mortos era capaz de penetrar em inúmeros pneumococos diferentes e vivos, e ainda fazer com que transmitissem as características

hereditárias da variedade morta à prole dos indivíduos vivos [20].

Dentre os muitos experimentos realizados, o principal foi a evidenciação do DNA como material hereditário pelos pesquisadores norte-americanos Alfred Day Hershey e Martha Chase. As pesquisas posteriores realizadas nos Estados Unidos pelo bioquímico Erwin Chargaff (1899–1985) definiram os quatro componente do DNA: adenina (A), citosina (C), guanina (G), e timina (T), denominadas bases nucleotídicas ou nitrogenadas. Em 1950, Chargaff dimensionou as proporções exatas das bases em cada molécula de DNA da seguinte maneira: ao analisar junto com seus colaboradores o DNA de várias espécies, verificou que a quantidade de timina era sempre igual à de adenina, e a de citosina era igual à de guanina. Esta descoberta gerou um grande indício de que para realizar a construção do modelo de dupla hélice do DNA, deveria ser com duas cadeias de bases atreladas: timina com adenina e citosina com guanina. Essas semelhanças são identificadas como Razões de Chargaff, e se tornou o ponto crucial para a descoberta da estrutura da molécula do DNA [9].

No final da década de 40, alguns indícios preconizavam que o DNA deveria ser a substância da qual eram constituídos os genes [9]. Consequentemente, muitos cientistas voltaram suas atenções e esforços a compreender os estudos das moléculas dessa substância, na tentativa de comprovar os mínimos detalhes da estrutura química do material genético que revelavam os segredos da hereditariedade. Entretanto, desconheciam a exata estrutura do DNA, como realizava suas funções e como se replicava. Desta forma, identificar as propriedades do DNA tornou-se enfim um objetivo extremamente importante para todo pesquisador que desejasse dar o próximo passo na Ciência.

Outro feito impactante na evidenciação da molécula de DNA primária foi a descoberta dos raios X por Wilhelm Röntgen em 1895. Entre grandes e renomados cientistas, um deles teve uma posição destacada no mundo da Ciência: Lawrence Bragg, que se interessou em compreender e desvendar a estrutura das moléculas inspirado nos trabalhos de seu pai, que 40 anos antes foi o co-criador da Cristalografia por raios X, uma técnica pela qual a estrutura dos cristais podem ser representadas em placas fotográficas especiais, quase como um retrato da difração em cristais. Méritos também para o cientista Linus Pauling. Sua grande descoberta foi que as unidades de aminoácidos nas proteínas estão dispostas de tal maneira que fazem surgir estruturas secundárias extremamente ordenadas e periódicas: as hélices alfa e as folhas beta.

Tentando estruturar e aplicar estudos que comprovassem as recentes descobertas nas diversas áreas científicas, dois cientistas iniciaram suas investigações sobre o DNA na esperança de desvendar o segredo da vida. Um deles é o pesquisador Francis Harry Compton Crick, nascido em Northampton, na Inglaterra e que desde muito jovem já demonstrava interesse por assuntos científicos. Formado em Física, teve como uma das influências para entrar no mundo da Biologia a leitura do livro "What is Life?" (em português: "O que é a Vida?") do físico Erwin Schrödinger, em que os mistérios da Genética são abordados pela Mecânica Quântica. O outro é James Dewey Watson, que aos vinte e dois anos, após ter concluído o doutorado pela Universidade de Indiana em Bloomington, aceitou uma bolsa de estudos do governo norte-americano para trabalhar em Copenhague, sob a orientação do bioquímico dinamarquês pioneiro da investigação em respiração celular Herman Kalckar.

Precisamente no ano de 1951, Francis Crick conhece James Watson, ambos engajados e fascinados pela misteriosa molécula do DNA, se convenceram de que era ela a chave para explicar o segredo da vida. Devido às suas pesquisas, identificaram a estrutura molecular da dupla hélice do DNA, em que uma molécula representada por dois filamentos formados por diversos nucleotídeos é torcida em hélice no espaço, ligados um filamento ao outro pelas bases nitrogenadas. A conexão entre as bases é realizada por meio de pontes de hidrogênio (ligações que se formam quando um hidrogênio que está conectado por uma ligação covalente a um átomo mais eletronegativo se aproxima de outro átomo eletronegativo, como oxigênio ou nitrogênio). Se em um dos filamentos da molécula de DNA houver a sequência AATTCCAGT, por exemplo, no outro filamento a sequência será TTAAGGTCA. Os dois filamentos não são então análogos entre si, mas sim complementares.

Linus Pauling, após a identificação das estruturas secundárias das proteínas, começou a estudar a estrutura da molécula de DNA e, ainda que sem grandes evidências, sugeria que esta fosse helicoidal. Com a notícia de que Pauling estava próximo de determinar a estrutura do DNA, Watson e Crick se empenharam em suas investigações. Munidos das imagens do DNA utilizando a difração de raios X obtidas por Maurice Wilkins e Rosalind Franklin, do King's College, e com o conhecimento das Razões de Chargaff, conseguiram enfim, no dia 28 de fevereiro de 1953, montar um modelo em que as bases nucleotídicas estavam dispostas no interior das hélices e ligadas corretamente.

2.1 Definição da estrutura do DNA

Na seção anterior relatamos a descoberta da molécula de DNA, um marco na história para a Ciência. O DNA é encontrado no núcleo das células de todos os seres vivos e possui todo o material genético de um organismo. É composto por dois filamentos (de nucleotídeos) que se entrelaçam em forma de espiral (dupla hélice). Agregados aos nucleotídeos, a estrutura do DNA é composta por mais três elementos:

1. **Bases Nitrogenadas** – adenina (A), timina (T), citosina (C) e guanina (G);
2. **Pentose** – açúcar composto por cinco átomos de carbono;
3. **Fosfato** – um radical de ácido fosfórico.

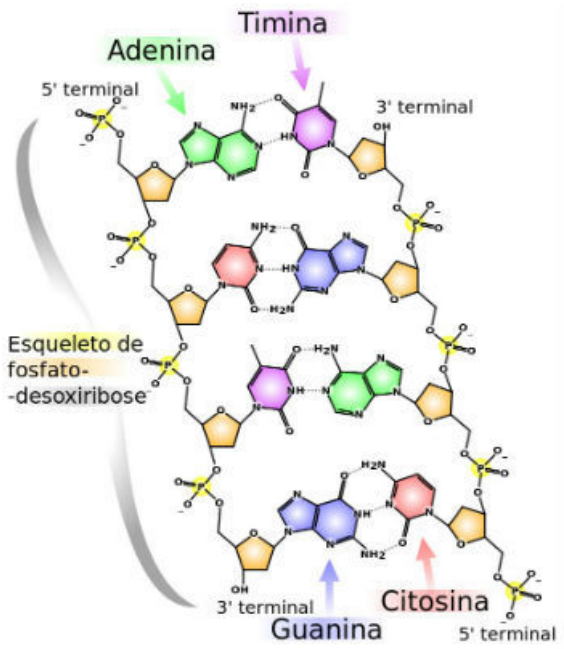

Figura 2.1: Estrutura da molécula do DNA. Fonte: `https://br.pinterest.com/geneticauva2015/estrutura-do-dna/`

Os dois filamentos que compõem o DNA envolvem-se um sobre o outro, interligando-se através de pontes de hidrogênio que se formam entre os pares de bases nitrogenadas dos nucleotídeos: adenina com timina (A-T) e citosina com guanina (C-G). A polaridade estrutural de um filamento é oposta a do outra filamento, portanto, uma encontra-se na direção 3' - 5' e a outra, 5' - 3', conforme os carbonos da ribose. Assim, se um lado apresenta 5' - ACTG - 3' a sua correspondente complementar e oposta é 3' - TGAC - 5' (Figura 2.1). Formam-se três pontes de Hidrogênio entre C-G e duas entre A-T. Os pares de bases formam pilhas entre os dois esqueletos açúcar-fosfato. As interações de empilhamento e as pontes de hidrogênio estabilizam estereoquimicamente a dupla fita.

Os ácidos nucleicos estão relacionados à transmissão das características de hereditariedade, e também comandam e gerenciam todas as atividades das células. São duas categorias de ácidos nucleicos: ácido desoxirribonucleico (DNA) e o ácido ribonucleico (RNA). O DNA e o RNA podem ser distinguidos por sua composição química e pelas bases nitrogenadas. A citosina, guanina e adenina ocorrem em todos os ácidos nucléicos. Enquanto a timina ocorre apenas no DNA e a uracila ocorre somente no RNA.

O DNA está compactado no núcleo celular. Diversas formas de vida do planeta possuem material genético codificado em sequências de bases nitrogenadas do DNA. Nas regiões do DNA que constituem os genes, as sequências de nucleotídeos atuam como um código: determinadas sequências representam um determinado aminoácido. Assim, as letras do DNA codificam uma sequência de aminoácidos, sendo peculiar a uma proteína [21]. Os genes são unidades de referência hereditária que compõem os cromossomos, constituídos por sequências especiais de centenas ou milhares de pares de bases nitrogenadas (A-T ou C-G). São eles que especificam tanto as características próprias da espécies, quanto as características próprias de cada indivíduo. Os genes são responsáveis por especificar as sequências de aminoácidos que servem como base principal para a síntese de proteínas celulares. Essas proteínas, em geral, vistas como enzimas, agem na estrutura e nas funções metabólicas das células e, consequentemente, no funcionamento de todo o organismo. As informações genéticas estão registradas nos cromossomos, sendo constituídos de estruturas do DNA.

O Genoma é o grande condutor de toda a informação hereditária codificada no DNA de um organismo, ou no RNA, no caso da maioria dos vírus. É o conjunto de todos os genes que define a espécie. O sequenciamento de DNA ou genoma é o método que auxilia na determinação da ordem em que as bases nitrogenadas ocorrem no DNA. Sequenciar um genoma consiste em determinar a ordem em que as

informações, ou seja, os genes, estão empregados na sequência completa de DNA, o que permite obter informações sobre a linha evolutiva dos organismos, podendo atrair a criação de novos métodos para diagnosticar doenças ou formular medicamentos e vacinas.

Capítulo 3

Proteínas – Origem, função e estrutura

O termo proteína foi criado no ano de 1838 pelo químico sueco Jöns Jacob Berzelius para descrever um tipo em particular de macromoléculas compostas por uma cadeia linear de aminoácidos, e que existem em grande variedade em organismos vivos [22]. O termo é proveniente do grego proteios e significa "a mais importante" [23]. As proteínas são polímeros de aminoácidos decorrentes da tradução das informações genéticas contida no DNA das células e são os compostos orgânicos mais abundantes de matéria viva da célula.

As proteínas são extremamente complexas e importantes para evolução de todos os seres vivos [24]. Em decorrência de inúmeras variações de funções, as proteínas têm como principal característica estarem envolvidas em quase todos os fenômenos biológicos, como produção de energia, defesa imunológica, contração muscular, atividade neuroquímica e reprodução [25]. Muitas proteínas são responsáveis por incentivar reações químicas, fornecendo uma rigidez estrutural à célula, controlam o fluxo do material através da membrana, regulam as concentrações dos metabólitos que atuam como um aparelho capaz de detectar pontos, chaves que produzem movimentos e controlam a função genética [26]. Dentre as peculiaridades inerentes às proteínas estão: a catálise enzimática, o transporte, o armazenamento de moléculas e íons, sustentação mecânica, proteção imunitária (anticorpos), geração e transmissão dos impulsos nervosos e no controle do crescimento da célula [27].

Os aminoácidos são pequenas moléculas que constituem as proteínas. As proteínas têm sua formação nos ribossomos das células através do processo de tradução do RNA mensageiro (RNAm), molécula portadora da sequência de nucleotídeos transcritos do DNA a ser sintetizado. Os ribossomos são considerados uma espécie de fábrica de proteínas das células. São encontrados livres no citoplasma tanto nas células

eucariontes como nas procariontes. O ribossomo só se faz funcional quando suas duas subunidades estão unidas ao Aminoacil e ao Peptidil (sítios de ligação de RNAt, RNA transportador), para realizarem a criação das proteínas. Existem mais de 200 aminoácidos na natureza, no entanto, apenas 20 formam as proteínas.

A discriminação entre as proteínas está relacionada à quantidade, tipo e sequência dos aminoácidos. A ligação entre um ou mais aminoácidos é denominada ligação peptídica e ocorre entre o grupamento carboxila de um aminoácido e o grupamento amina do outro, com liberação de uma molécula de água no processo.

Os aminoácidos são classificados em duas categorias:

- Naturais: Produzidos pelo próprio organismo. São aminoácidos naturais: arginina, glicina, alanina, serina, cisteína, tirosina, ácidos aspártico, ácido glutâmico, asparagina, glutamina e prolina.
- Essenciais: Ingeridos através da alimentação. São aminoácidos essenciais: histidina, lisina, triptofano, fenilalanina, treonina, valina, metionina, leucina e isoleucina.

A fórmula do aminoácido tem sua definição **"R-CH(NH2)-COOH"** um grupo amina (NH_2) um grupo carboxila (COOH) conectados a um átomo de carbono conhecido como carbono alfa (C_α), que ainda se conecta a um átomo de hidrogênio (H) e a uma cadeia lateral (R) que diferencia os tipos de aminoácidos. A Figura 3.1 abaixo ajuda compreender a fórmula.

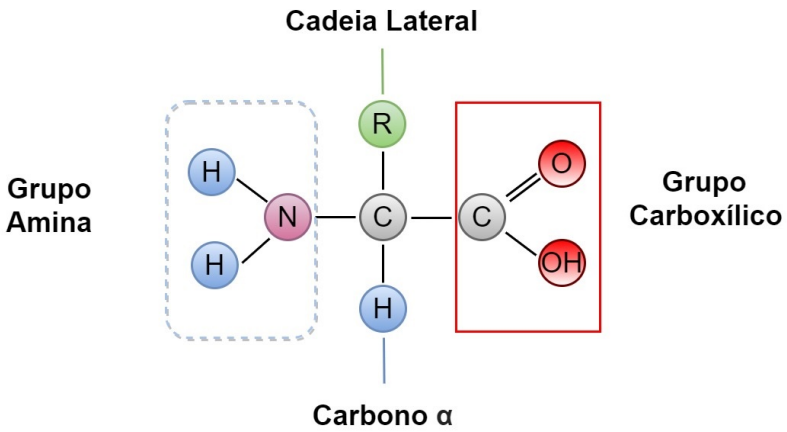

Figura 3.1: Estrutura de um aminoácido.

Os níveis de organização molecular de uma proteína são uma das formas que podem ser utilizadas para classificá-la. Se a conformação da proteína é alterada, a mesma torna-se inativa. Estas alterações podem ocorrer por mudanças de pH, altas temperaturas e outros fatores. As proteínas têm suas estruturas analisadas em quatro níveis: estrutura primária, secundária, terciária e quaternária. As estruturas primárias são constituídas por vários aminoácidos interligados por ligações peptídicas, onde cada aminoácido identificado nessas ligações são denominados resíduos. As estruturas secundárias são formadas por folhas beta ou hélices alfa. As estruturas terciárias são representadas pela junção das estruturas anteriores e, por fim, as estruturas quaternárias são um complexo de moléculas proteicas de diversas estruturas terciárias aglutinadas que se interligam mutuamente [28].

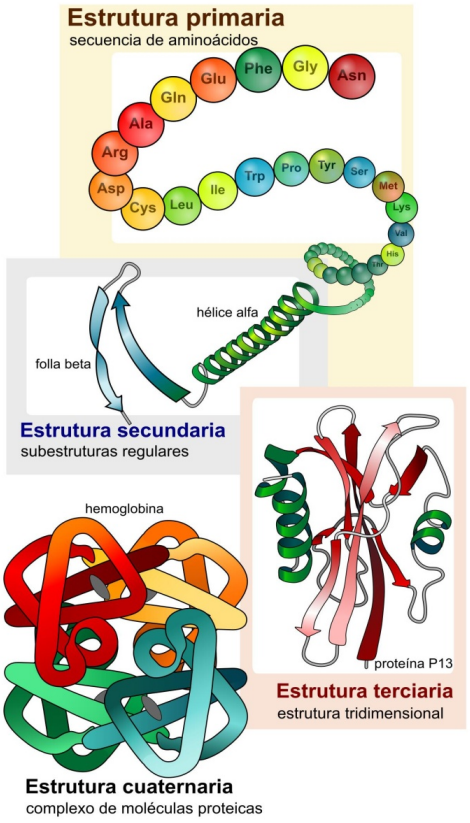

Figura 3.2: Estrutura de uma proteína. Fonte: `https://commons.wikimedia.org/wiki/File:Main_protein_structure_levels_gl.svg`

As proteínas podem apresentar uma grande variedade entre os organismos. Suas composições devem ser completamente determinadas pela molécula de DNA: longas cadeias peptídicas formadas por cerca de vinte tipos diferentes de aminoácidos que podem ser consideradas como "palavras" longas com base em um alfabeto de 20 letras [29] podendo conter sequências de 4.500 aminoácidos [30]. A Tabela 3.1 abaixo lista os 20 tipos de aminoácidos e os respectivos códigos de três letras e de uma letra.

Tabela 3.1: Aminoácidos representados por códigos (letras)

Nome	Sigla de 3 letras	Sigla de 1 letra	Polaridade
Alanina	Ala	A	Apolar
Cisteína	Cys	C	Apolar
Ácido aspártico	Asp	D	Negativa
Ácido glutâmico	Glu	E	Negativa
Fenilalanina	Phe	F	Apolar
Glicenina	Gly	G	Apolar
Histidina	His	H	Positiva
Isoleucina	Ile	I	Apolar
Lisina	Lys	K	Positiva
Leucina	Leu	L	Apolar
Metionina	Met	M	Apolar
Asparagina	Asn	N	Neutra
Prolina	Pro	P	Apolar
Glutamina	Gln	Q	Neutra
Arginina	Arg	R	Positiva
Serina	Ser	S	Neutra
Treonina	Thr	T	Neutra
Valina	Val	V	Apolar
Triptofano	Trp	W	Apolar
Tirosina	Tyr	Y	Neutra

O surgimento e evolução das proteínas teve seu início com pequenas cadeias de aminoácidos que gradativamente aumentaram de tamanho e/ou se agruparam dando origem a estruturas mais estáveis capazes de realizar determinadas funções. As proteínas são formadas por uma ou mais destas estruturas independentemente estáveis, denominadas domínios. A identificação de domínios que ocorrem em uma proteína pode fornecer informações significantes sobre suas funções. Por esta razão biólogos classificam domínios de proteínas em famílias e importam-se sobre a confiabilidade de sua classificação [31].

Um domínio proteico é uma região identificada de uma proteína onde há uma grande densidade de átomos: é neste local onde se sucedem inúmeras dobras. Uma cadeia polipeptídica pode possuir um ou mais domínios e, por sua vez, uma proteína pode ser composta por uma ou mais cadeias de polipeptídeos. Uma proteína formada por mais de uma cadeia polipeptídica pode, inclusive, ter um só domínio compartilhado pelas cadeias de polipeptídios. Um domínio proteico pode ser funcional caso seja uma unidade modular da proteína que realiza uma função bioquímica determinada, e estrutural quando se refere a um componente estável da estrutura.

3.1 Enovelamento (Folding) de proteínas

O enovelamento ou dobramento de proteínas (do termo inglês *folding*) é uma espécie de mecanismo de formação do arranjo tridimensional da proteína, um método pelo qual as proteínas buscam a sua conformação funcional nativa. Como visto anteriormente, a sequência de resíduos de aminoácidos determina a estrutura tridimensional e esta última, a função da proteína. A técnica da síntese da proteína no ribossomo determina uma cadeia linear de resíduos de aminoácidos sem estrutura tridimensional específica [32]. O enovelamento proteico depende das interações intermoleculares da proteína com o meio, sendo este geralmente aquoso. Assim, o processo de enovelamento para uma mesma proteína é diferente, estando ela isolada em solução ou em ambiente celular, onde há altas concentrações de outras macromoléculas. Dentro da célula o processo de enovelamento pode ser ajudado por proteínas auxiliares denominadas chaperonas. As proteínas podem não alcançar sua conformação nativa estável e funcional devido à ocorrência de mutações ou a efeitos distintos no meio que inibam o enovelamento.

A função proteica se mantém conservada nos domínios proteicos. Os domínios possuem a capacidade de realizar dobramentos independentes e são aproximadamente conservados entre as espécies. Um objetivo importante no estudo de proteínas é a identificação e classificação de domínios e famílias de domínios. Alguns algoritmos

surgem na literatura científica para obter informações sobre domínios a partir da estrutura tridimensional de proteínas, mas muitos desses resultados divergem com os definidos por biólogos especialistas [33].

Embora seja extremamente complexo, a técnica de enovelamento pode ocorrer em questões de segundos ou até milésimos de segundos. Na década de 1960 foram realizados muito estudos que abordavam o problema do enovelamento proteico com objetivo de ampliar o entendimento sobre como a informação contida na sequência de aminoácidos leva a proteína a aproximar-se da sua conformação nativa [34]. O total entendimento dos processos moleculares envolvidos é ainda um desafio para os pesquisadores. Até o momento não foi identificado nada que pudesse ser um fato determinístico tão essencial que direcione a sequência de aminoácidos até o estado nativo de enovelamento da proteína.

Capítulo 4

Bases de dados Biológicos

Com o avanço de tecnologias atuantes em diversos níveis da sociedade, na área da Biologia não seria diferente, o que vem proporcionando um impacto considerável. Com uma alta demanda de procedimentos eficientes e eficazes inerentes à área, o termo bioinformática se refere a uma ciência que utiliza métodos matemáticos e computacionais para analisar e solucionar questões relevantes a partir de um complexo conjunto de dados biológicos. Em decorrência dessa junção, é possível a realização de diversas pesquisas nessa área que incluam o desenvolvimento de técnicas/métodos para armazenamento, recuperação e análise de dados biológicos.

Os dados biológicos são considerados dados ou medidas coletadas a partir de inúmeras fontes biológicas. São armazenados em banco de dados biológicos (BDBs) através de arquivos em sua forma bruta ou em sistemas de gerenciamento de bancos de dados (SGBDs).

O primeiro banco de dados biológico surgiu em 1956 com o sequenciamento da insulina e por volta da década de 60, outras sequências de proteínas foram determinadas tornando posteriormente possível o surgimento de outros banco de dados, como o Protein Data Bank (PDB) [35], criado em 1971 pelo Laboratório Nacional de Brookhaven. No começo eram armazenadas apenas sete estruturas de macromoléculas no PDB, sendo outras acrescentadas nos anos subsequentes. A integração dos computadores nas ciências vêm crescendo nas últimas décadas. Com os dados vêm os desafios computacionais para analisar, interpretar, visualizar e integrar informações.

O maior desafio atualmente na Biologia Molecular é analisar e entender o grande volume de dados de sequências e de estruturas de proteínas oriundas de projetos atualmente em desenvolvimento na área biológica. Ao longo das últimas décadas surgiram sistemas especializados capazes de manipular inúmeras informações em

conjunto com os banco de dados usados para o armazenamento e o gerenciamento de proteínas, dentre outras informações extremamente sensíveis e de grande valor para os estudos.

Na atualidade existe uma grande variedade de bancos de dados biológicos disponíveis na Internet com diversos métodos de acesso que fornecem seus conjuntos ou subconjuntos de dados, tais como GenBank, mantido pela instituição chamada National Center for Biotechnology Information (NCBI), EMBL, PDB, ZINC, UNIPROT, SRA e Pfam, dentre outros, em suas maioria disponíveis na Web para pesquisa, possibilitando o acesso de inúmeros pesquisadores das mais diversas áreas de atuação.

A base de dados PDB armazena estruturas obtidas por cristalográfica ou por ressonância magnética nuclear (RMN). Outras estrururas moleculares tambem estão disponíveis. As estruturas são identificadas através de um código alfanumérico de quatro caracteres. O acesso aos dados do PDB pode ser feito pelo espelho www.rcsb.org.

O ZINC é uma base de dados de estrutura de moléculas, pouco conhecida, mas que tem as suas vantagens por ser extremamente pequena, podendo oferecer pesquisas completas em banco de dados, resultando em melhores estimativas das respostas verdadeiras, com o objetivo de fazer com que essas consultas de longa duração sejam redirecionadas automaticamente para o modo de lote, possibilitando novas ferramentas de análise mais intuitivas. O ZINC contém mais de 230 milhões de compostos disponíveis em um formato de 3D prontos para serem utilizados gratuitamente através do endereço: https://zinc.docking.org/ [36].

O Universal Protein Resource (UniProt) é uma base de dados de sequências de proteínas e de suas funções. É subdividido em duas bases de dados: Swiss-Prot. e o TrEMBL. A diferença entre elas é que o Swiss-Port possui proteínas que são manualmente anotadas e revisadas com informações em caráter experimental, já o TrEMBL foi criado originalmente porque os dados de sequência estavam sendo gerados em um ritmo que excedia a capacidade do Swiss-Port de acompanhar. O acesso ao Uniprot pode ser feito por meio do link:https://www.uniprot.org/ [37].

GenBank é uma base de dados de sequência genética do National Center for Biotechnology Information (NCBI). O banco de dados do GenBank é projetado para fornecer e incentivar o acesso na comunidade científica às informações mais atualizadas e abrangentes sobre a sequência de DNA. Portanto, o NCBI não impõe restrições ao uso ou distribuição dos dados do GenBank. Para acessá-lo temos dois

caminhos, o acesso via um protocolo de comunicação conhecido como FTP, que é atualizado a cada dois meses, e a versão mais recente está disponível em seu próprio site: www.ncbi.nlm.nih.gov/genbank [38].

Sequence Read Archive (SRA) é uma base de dados que armazena dados brutos de sequenciamento e informações de alinhamento para aprimorar a reprodutibilidade e facilitação de novas descobertas por meio de análise de dados. É considerado como um dos maiores repositórios disponíveis para o público que trabalha com sequenciamento de alto rendimento. O formato de dados recomendado para arquivos submetidos à SRA é o formato BAM, capaz de armazenar leituras alinhadas e não alinhadas. Atualmente encontra-se em vários provedores de nuvem e servidores do NCBI, e pode ser acessado pelo espelho: www.ncbi.nlm.nih.gov/sra [39].

4.1 Uso de Modelos Ocultos de Markov (HMM) no alinhamento de sequências de proteínas

A análise computacional é cada vez mais importante para inferir as funções e estruturas das proteínas porque a velocidade do sequenciamento de DNA há muito ultrapassou a taxa onde a função biológica das sequências pode ser elucidada experimentalmente [40].

Técnicas de comparações de sequência em pares, como BLAST e FASTA, geralmente consideram que todas as posições de aminoácidos são igualmente importantes, embora uma quantidade abundante de posições dessas informações específicas geralmente estão disponíveis para uma proteína ou família de proteínas. Alinhamentos múltiplos de famílias de sequências de proteínas indicam resíduos que são mais conservados do que outros, e os pontos em que as inserções e as exclusões são mais frequentes. Estruturas de informações tridimensionais (3D), permitem que ambientes estruturais possam ser considerados ao marcar os resíduos alinhados e permite que inserções e exclusões sejam esperadas mais frequentemente em *loops* na superfície em elementos de estrutura secundária no núcleo. Um "perfil" (definido como um consenso de modelo de estrutura primária consistindo em pontuações de resíduos específicos nas posições da sequência e penalidades de inserção ou exclusão de resíduos) é um passo intuitivo além dos métodos de alinhamento de sequências emparelhadas [41].

A ideia principal é que um Modelo Oculto de Markov (HMM, do inglês Hidden Markov Model) é um modelo finito que descreve uma distribuição de probabilidade sobre um número infinito de possíveis sequências. O HMM é composto por um número de estados, que podem corresponder a posições em uma estrutura 3D ou colunas de um alinhamento múltiplo. Cada estado "emite" símbolos (resíduos) conforme a probabilidade de emissão de símbolos, e os estados estão interligados por probabilidades de transição de estado.

O problema com os perfis é que eles são modelos complexos quando trabalham com muitos parâmetros livres. Métodos de perfil usando HMMs, introduzidos das últimas décadas, vêm se mostrando mais eficazes para reconhecimento homólogo e podem ser construídos a partir de uma vasta gama de sequências [41].

Os perfis baseados em HMM resolveram muitos dos problemas associados à forma padrão de análise. HMMs fornecem uma teoria consistente para pontuação de inserções e exclusões, e uma estrutura consistente para combinar informações

estruturais e de sequência. O alinhamento de múltiplas sequências baseadas em HMM está melhorando rapidamente. O reconhecimento homólogo baseado em HMM já é suficientemente poderoso para comparar favoravelmente com métodos de enovelamento de proteínas que são mais complicados, como o enovelamento inverso (dado uma conformação alvo, encontrar a sequência de aminoácidos que melhor se ajusta a ela).

A análise de estruturas a partir das sequências usando HMM é um problema complicado, sendo um melhor tratamento considerar como um problema de inferência estatística usando modelos probabilísticos completos. É importante ter em mente que os perfis HMM são um caso muito especial de abordagens de HMM e que os métodos de HMMs estão sendo usados para uma variedade de problemas biológicos [41], tendo a utilidade e o alcance do HMM nas aplicações oriundas da biologia estrutural vem crescendo nas últimas décadas.

4.2 Introdução ao Protein Family Database (Pfam)

O Pfam é um banco de dados de domínio de proteínas, família de domínios e clãs altamente selecionados, cada família definida por duas técnicas de alinhamento e um perfil HMM. Os perfis HMM das famílias são construídos a partir de um conjunto de sequências representativas alinhadas, selecionadas pelo comitê biológico [42]. Um alinhamento de sementes de alta qualidade é essencial, pois fornece a base para as frequências de aminoácidos específicos da posição, intervalos e parâmetros de comprimento no perfil HMM [43]. Mediante um processo de verificação de similaridade entre as sequências de aminoácidos de domínios distintos, estes podem ser aglutinados em famílias, de forma que é esperado que elas reúnam indivíduos que progrediram de algum ancestral em comum. A homogeneidade entre as sequências dos membros de uma mesma família não acarreta necessariamente que estes apresentem funções similares [44].

O objetivo do banco de dados Pfam é fornecer uma classificação completa e precisa de famílias com seus respectivos domínios de proteínas [45]. Originalmente, a razão por trás da criação do banco de dados era ter um recurso semiautomático para trabalhar as informações sobre famílias de proteínas conhecidas para elevar a eficiência da anotação de genomas. A classificação Pfam de famílias de proteínas foi amplamente adotada pelos biólogos por sua abrangente cobertura de proteínas e convenções de nomenclatura sensíveis. Surgiu na comunidade científica uma

diversidade de grupos empenhados na identificação de sequências de aminoácidos e em sua subsequente armazenação em grandes bancos de dados, gerando uma espécie de catalogação [46].

A versão 18.0 do Pfam introduziu o conceito de clãs [47]: agrupamentos de duas ou mais famílias, como superfamílias. Os domínios de um clã apresentam semelhanças em suas sequências de aminoácidos, mas não o suficiente para classificar todos como pertencentes a uma mesma família. A classificação em clãs auxilia na identificação de funções e estruturas de determinadas famílias cujas características eram desconhecidas, graças à associação com famílias cujas informações são conhecidas [48].

No presente trabalho foram adotadas como objeto de estudo e pesquisa as famílias agrupadas em clãs na versão 27.0 do Pfam. Os arquivos do banco de dados podem ser encontrados no site do Instituto de Bioinformática Europeu (European Bioinformatics Institute – EMBL-EBI) [49], que disponibiliza publicamente dados para a comunidade científica através de um conjunto de serviços e ferramentas, executa buscas básicas e provê treinamento profissional em bioinformática. A organização das famílias armazenadas na versão 27.0 do Pfam é dada da seguinte forma:

- Pfam-A: Composto por famílias de proteínas com alta qualidade, organizadas "manualmente" com um critério para inserção e (algumas delas) agrupadas em clãs;
- Pfam-B: Composto por famílias de proteínas automaticamente adicionadas, produzidas a partir de clusters de sequências que atualmente não são cobertas pelas entradas do Pfam-A.

Atualmente o Pfam está na versão 35.0 e contém um total de 19.632 famílias e 657 clãs. Em relação à versão anterior, foram introduzidas 460 novas famílias, eliminadas 7 famílias e originaram-se 12 novos clãs. Os proteomas de referência do UniProt aumentaram 7% em relação à versão 34.0, contendo atualmente 61 milhões de sequências. Das sequências que estão nos proteomas de referência, 74,5% têm pelo menos uma correspondência Pfam e 48,7% de todos os resíduos que se enquadram em uma família Pfam. O Pfam na versão atual gerou uma enormidade de modelos estruturais para novas famílias em seu *pipeline* que está integrado ao InterPro, que fornece análise funcional de proteínas, classificando-as em famílias e prevendo domínios e sítios importantes [50].

Prever a função de uma proteína a partir de sua sequência de aminoácidos bruta é uma etapa crítica para compreender a relação entre o genótipo e o fenótipo. À medida que o custo do sequenciamento de DNA cai e os projetos de sequenciamento metagenômico florescem, ferramentas rápidas e eficientes que anotam quadros de

leitura abertos com função desempenharão um papel central na exploração desses dados [51, 52]. Para a classificação das proteínas dessa forma, a InterPro usa modelos preditivos, conhecidos como assinaturas, fornecidos por vários bancos de dados diferentes (chamados bancos de dados membros) que compõem o consórcio InterPro, combinando as assinaturas de proteínas desses bancos de dados membros em um único recurso pesquisável, capitalizando seus pontos fortes individuais para produzir um poderoso banco de dados integrado junto com uma ferramenta de diagnóstico.

Uma equipe de pesquisadores da gigante empresa de TI, Google Research, utilizou métodos de aprendizagem profunda (deep learning) no alinhamento de sequências em conjunto com o Pfam HMMER (software usado pelo Pfam para a construção dos alinhamentos), encontrando muitas correspondências adicionais com o UniProtKB (UniProt KnowledgeBase). Os alinhamentos obtidos por este processo são disponibilizados desde a versão 34.0 no arquivo Pfam-N. Acredita-se que este método tem potencial para ajudar a identificar proteínas que catalisam novas reações, projetar novas proteínas que se ligam a alvos microbianos específicos ou construir moléculas que aceleram os avanços na biotecnologia. A prática atual para a previsão funcional de uma nova sequência de proteína envolve o alinhamento em um grande banco de dados de sequências anotadas usando algoritmos como BLASTp [53], ou perfis HMM construídos a partir de famílias de sequências alinhadas, como as fornecidas pelo Pfam [54, 55].

O Pfam-N anota 6,8 milhões de regiões de proteína (Figura 4.1) em 11.438 famílias Pfam. Essas regiões incluem quase 1,8 milhões de sequências completas de proteínas do proteomas de referência do UniProtKB que anteriormente não tinham correspondência no Pfam, uma melhoria de 4,2% em relação aos 42,5 milhões anotados no arquivo Pfam-A da versão 34.0. Observaram também que, entre as sequências que recebem sua primeira anotação Pfam, 360 são sequências humanas [50].

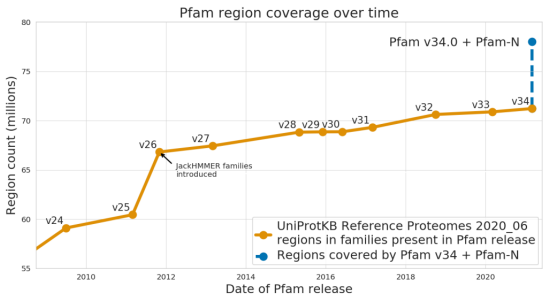

Figura 4.1: Referência de Proteomas no UniProtKB. Fonte: `https://xfam.wordpress.com/2021/03/24/google-research-team-bring-deep-learning-to-pfam/`

A Tabela 4.1 abaixo, descreve a constituição do Pfam desde a versão 18.0, em que foram incluídos os primeiros clãs, apresentando o ano de lançamento da versão do banco, a quantidade de famílias existentes, o número de famílias organizadas em clãs e a quantidade de clãs existentes em cada versão [46].

Tabela 4.1: Evolução do Banco de Proteínas Pfam.

Versão	Ano	Nº de Famílias	Nº de Famílias em Clãs	Nº de Clãs
18.0	2005	7973	1181	172
19.0	2005	8183	1399	205
20.0	2006	8296	1560	239
21.0	2006	8957	1683	262
22.0	2007	9318	1815	283
23.0	2008	10340	2016	303
24.0	2009	11912	3132	423
25.0	2011	12273	3439	458
26.0	2011	13672	4243	499
27.0	2013	14831	4563	515
28.0	2015	16230	4939	541
29.0	2015	16295	5282	559
30.0	2016	16306	5423	595
31.0	2017	16712	5996	604
32.0	2018	17929	7001	628
33.1	2020	18259	7331	635
34.0	2021	19179	8251	645
35.0	2021	19632	8704	657

Entre cada nova versão disponibilizada do Pfam podem ocorrer alterações como, por exemplo: inclusão e exclusão de domínios; criação ou exclusão de famílias, podendo os domínios que formavam a família extinta distribuídos entre uma ou mais famílias; criação ou exclusão de clãs, podendo as famílias que formavam o clã serem aglutinadas a um ou mais clãs. O motivo para tantas alterações deve-se ao fato dos resultados computacionais serem continuamente examinados por diversos biólogos especialistas que registram suas opiniões sobre as famílias e clãs [3].

Assim como as famílias e clãs, os alinhamentos sementes das famílias também estão em constante evolução. Para explicar sua formação, precisamos antes introduzir como se deu o processo de criação do Pfam e como ele se desenvolve a cada nova versão disponibilizada. Com a identificação de domínios e a percepção de que eles se repetiam em diferentes tipos de proteínas com algumas alterações em sua sequência de aminoácidos, um processo de verificação e que comprovasse a ligação entre as diferentes sequências, foi introduzido. O método de alinhamento consiste na busca do

resultado mais adequado de uma correlação mais assertiva entre os aminoácidos das sequências sobrepostas. Para tal, lacunas ou aberturas (*gaps*) podem ser acrescidas às sequências. A motivação para a utilização destas lacunas se deve ao fato de que durante o processo de evolução, o que anteriormente era um único domínio, deu origem a diversos outros domínios através da inclusão e/ou eliminação de resíduos, além da alteração de alguns resíduos da sequência original [3].

O alinhamento foi, inicialmente, efetuado entre sequências já conhecidas como sendo biologicamente associadas e, em seguida, novos indivíduos suspeitos foram incorporados às famílias presentes através de um processo de verificação. Todo o método de verificar a estrutura de uma proteína, determinar sua composição química, distinguir os domínios, alinhá-los com outros previamente identificados, conferir os resultados da correspondência de aminoácidos e, muitas vezes, retornar a uma destas etapas, era muito demorado, não havendo perspectivas de quando todos os domínios estariam devidamente catalogados [3].

No início dos anos 90, os primeiros trabalhos incluindo a utilização de HMM na modelagem de alinhamentos de múltiplas sequências de aminoácidos começaram a despontar [56, 57], o que influenciou e estimulou inúmeros grupos a optarem pelos métodos com modelos probabilísticos, e na segunda metade da mesma década, precisamente no ano de 1997, a primeira versão do PFAM foi concebida [58]. O uso de HMM viabilizou um grande impulso na velocidade de identificação dos domínios [3].

As sementes foram criadas pela seleção feita por especialistas de sequências consideradas mais significantes para a caracterização das famílias. As mais pertinentes não são aquelas mais idênticas entre si, mas sim aquelas que, sendo por pressuposição, reconhecidas como sendo pertinentes à mesma família, resultam em uma maior diversidade. Após a distinção dos elementos, em seguida ocorre o processo de alinhamento. Com as sequências alinhadas são construídos os perfis de modelos HMM, sendo justamente as sementes que depois serão utilizadas para a identificação dos domínios nas sequências inseridas em versões subsequentes. Quando uma sequência não é identificada pelo software como sendo pertinente a uma dada família, mas é estabelecida como tal pelo critério dos especialistas, ela é então inserida na família. Efetuada a incorporação, todas as sequências cadastradas na família passam novamente pelo processo de análise para checar se a sua "associação" com a semente permanece significativa. Se a "pontuação" (*score*) diminuir consideravelmente para uma determinada sequência, pode ocorrer desta ser também incluída à semente, dando reinício ao processo de verificação, ou pode ela ser eliminada da família. Para

não haver falsos positivos ou falsos negativos na identificação de sequências, a semente deve ser criada com o máximo de qualidade possível [3].

Na Figura 4.2 temos um esquema de caracterização de domínios em proteínas com sua organização em famílias. Para as 12 proteínas fictícias são identificados um total de sete diferentes domínios. Os domínios semelhantes são agrupados em famílias e devidamente rotulados com informações que identifiquem sua proveniência, sua ordenação na sequência de aminoácidos da proteína da qual faz parte e sua família [3].

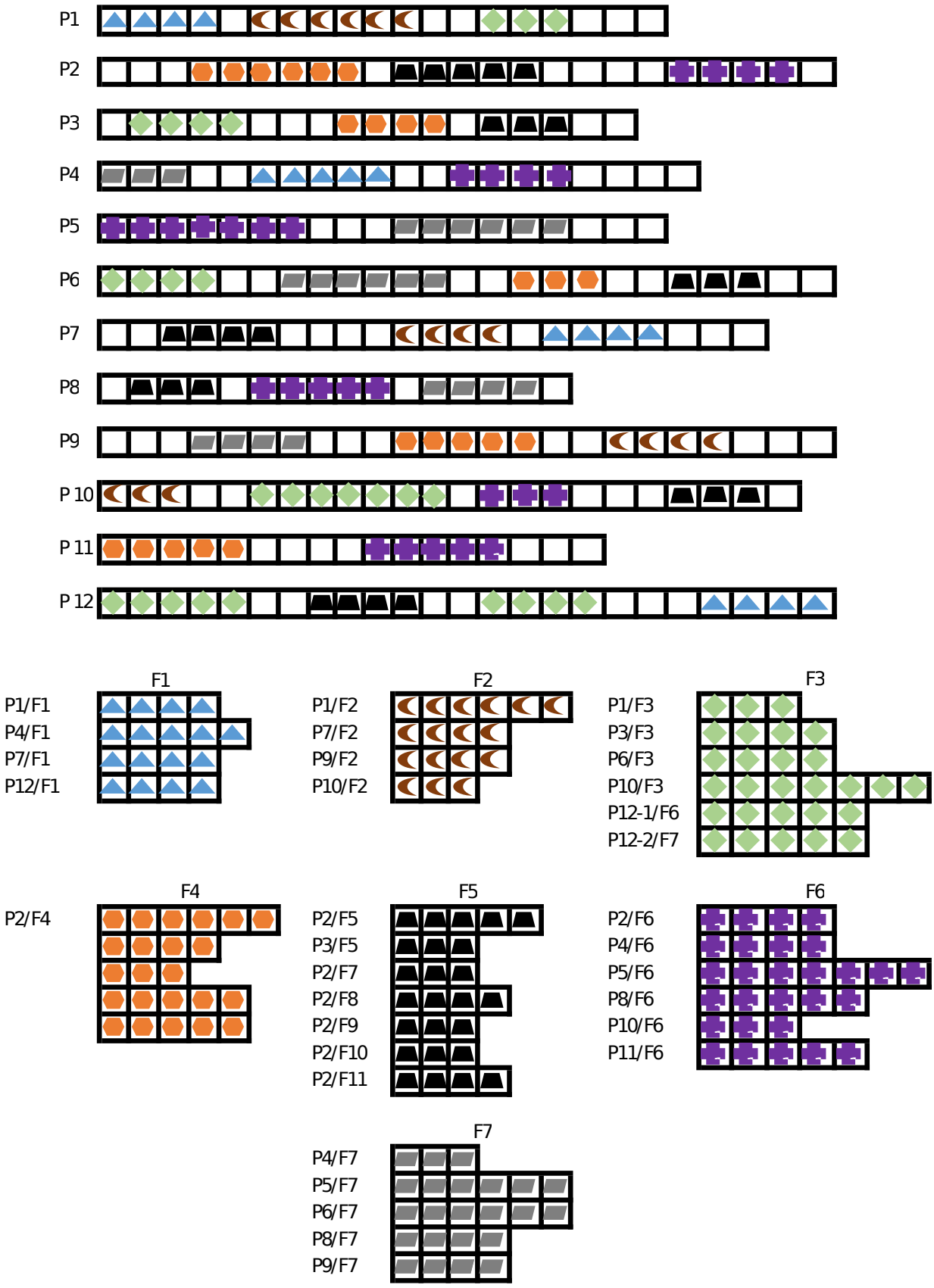

Figura 4.2: Exemplo de reconhecimento de domínio em proteínas e geração de famílias de domínios [1–3].

O EMBL-EBI disponibiliza uma grande quantidade de ferramentas e serviços utilitários, como, por exemplo: consulta por semelhanças de sequências, visualização de sequências de proteínas, consulta a famílias de proteínas, dentre outros. Para obter o arquivo Pfam-A.fasta.gz que contém os dados das famílias utilizados neste trabalho, o acesso deve ser feito através da URL ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam27.0/, utilizando o protocolo de transferência FTP.

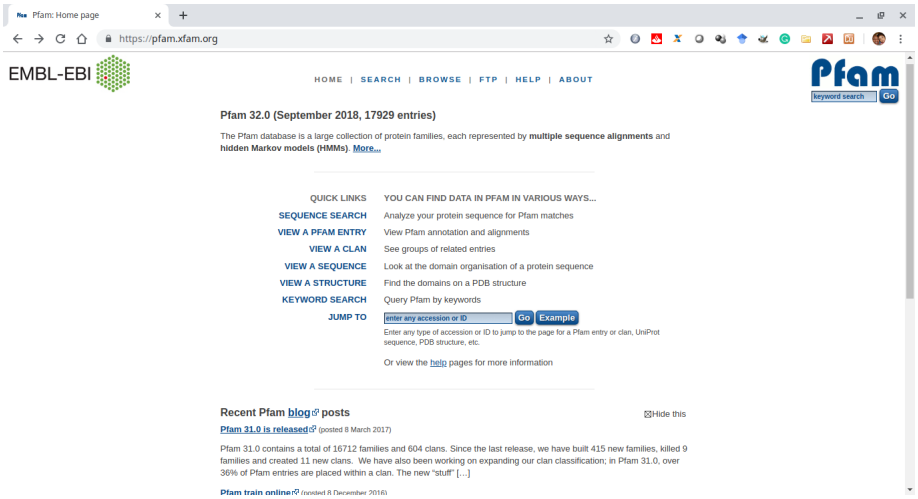

Figura 4.3: Página inicial e oficial do Pfam. Fonte: `http://pfam.xfam.org/`

Na Figura 4.3 é exibida a tela inicial do site oficial do Pfam. As Figuras 4.4 e 4.5 representam o procedimento para encontrar os arquivos da versão 27.0 do Pfam. Para a execução do trabalho, precisamos dos arquivos Pfam-A.fasta.gz, já mencionado anteriormente, e Pfam-A.clans.tsv.gz, que contém informações sobre todas as famílias contidas no Pfam-A. Em particular extraímos do arquivo Pfam-A.clans.tsv.gz a listagem das famílias pertencentes a um clã específico.

Os arquivos Pfam-A.fasta.gz e Pfam-A.clans.tsv.gz, após serem descompactados, são convertidos do formato Fasta e tsv, respectivamente, para arquivos no formato txt. O (formato) Fasta é muito utilizado na Bioinformática para representar sequências de nucleotídeos ou de peptídeos, nas quais nucleotídeos ou aminoácidos são caracterizados pelo uso dos códigos de uma única letra. Este formato origina-se do pacote de (software) Fasta, utilizado para o alinhamento de sequências de proteínas e de DNA. Foi inicialmente descrito como FastaP (Fasta Pearson), desenvolvido por David J. Lipman e William R. Pearson [59].

Um arquivo do tipo Fasta é constituído por diversas sequências. Cada sequência é introduzida por um cabeçalho que, por sua vez, é diferenciado dos outros dados da

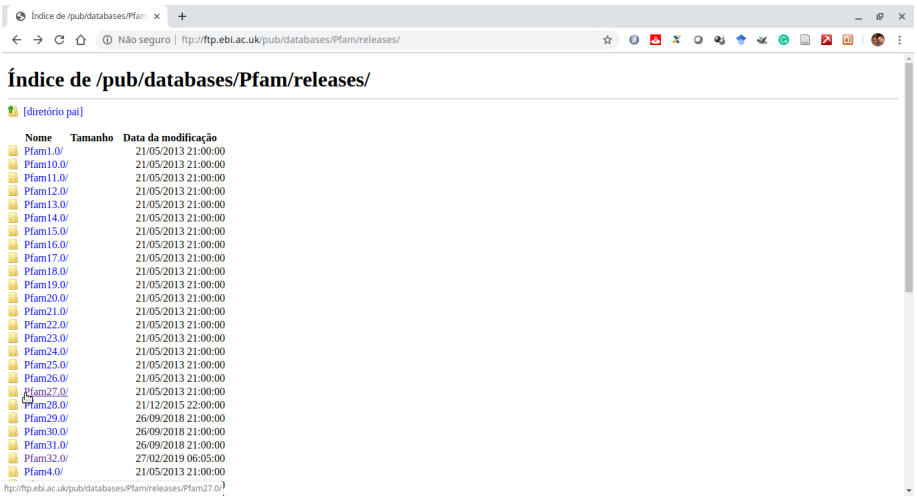

Figura 4.4: Página das versões do PFAM. Fonte: `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/`

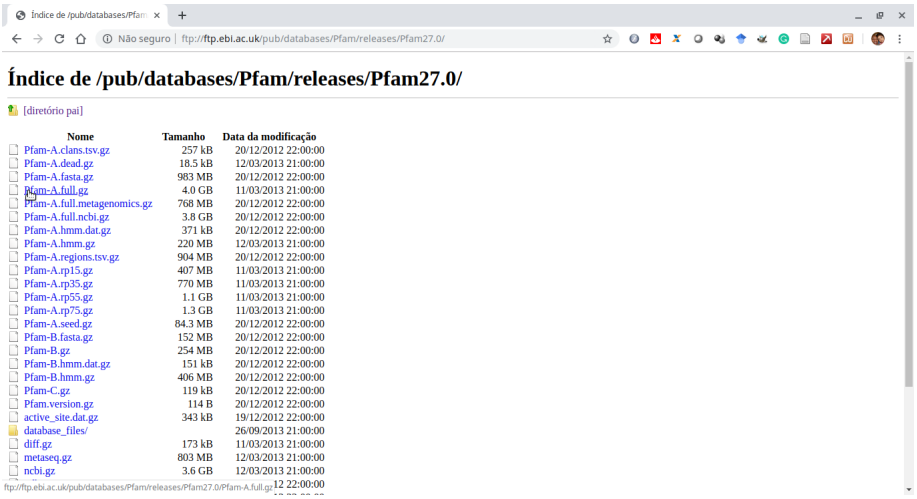

Figura 4.5: Página da versão 27.0 do PFAM. Fonte: `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam27.0/`

sequência pela presença do símbolo “maior que” (>) no início da linha. Contudo, de forma menos habitual, o cabeçalho pode ser iniciado com o símbolo de “ponto e vírgula” (;). As linhas iniciadas com “;” são tratadas como comentários. Um arquivo Fasta pode conter diversos cabeçalhos, e aqueles que possuem podem ser chamados de “Multifasta”. Após a linha de cabeçalho, é listada a sequência de elementos, representados através do código de uma letra. Frequentemente, a sequência é finalizada com o uso de asterisco (*). Recomenda-se que todas as linhas de texto tenham menos de 80 caracteres de comprimento.

```
>sequenceID-001 description
AAGTAGGAATAATATCTTATCATTATAGATAAAAACCTTCTGAATTTGCTTAGTGTGTAT
ACGACTAGACATATATCAGCTCGCCGATTATTTGGATTATTCCCTG
>sequenceID-002 description
CAGTAAAGAGTGGATGTAAGAACCGTCCGATCTACCAGATGTGATAGAGGTTGCCAGTAC
AAAAATTGCATAATAATTGATTAATCCTTTAATATTGTTTAGAATATATCCGTCAGATAA
TCCTAAAAATAACGATATGATGGCGGAAATCGTC
>sequenceID-003 description
CTTCAATTACCCTGCTGACGCGAGATACCTTATGCATCGAAGGTAAAGCGATGAATTTAT
CCAAGGTTTTAATTTG
```

Figura 4.6: Representação de um arquivo com a sequência Fasta.

```
>sequenceID-001 description
AAGTAGGAATAATATCTTATCATTATAGATAAAAACCTTCTGAATTTGCTTAGTGTGTAT
ACGACTAGACATATATCAGCTCGCCGATTATTTGGATTATTCCCTG
>sequenceID-002 description
CAGTAAAGAGTGGATGTAAGAACCGTCCGATCTACCAGATGTGATAGAGGTTGCCAGTAC
AAAAATTGCATAATAATTGATTAATCCTTTAATATTGTTTAGAATATATCCGTCAGATAA
TCCTAAAAATAACGATATGATGGCGGAAATCGTC
>sequenceID-003 description
CTTCAATTACCCTGCTGACGCGAGATACCTTATGCATCGAAGGTAAAGCGATGAATTTAT
CCAAGGTTTTAATTTG
```

Figura 4.7: Representação de um arquivo com a sequência Multi-Fasta.

Espera-se que as sequências sejam representadas nos códigos de aminoácidos e ácidos nucleicos IUB/IUPAC padrão, com as seguintes exceções: letras minúsculas são aceitas e mapeadas em maiúsculas; um único hífen ou travessão pode ser usado para representar uma lacuna de comprimento indeterminado (*gap*); e nas sequências de aminoácidos, "U" e "*" são letras aceitáveis.

Para realizar o procedimento de extração de blocos representativos de famílias pertencentes à clãs, utilizamos um script, extractCLAN.sh, construído especificamente para este propósito. Inicialmente utilizado para a construção de blocos 100 × 100 [60], pode ter seu código adaptado para o número de linhas e colunas desejado. Ao rodarmos o script em um terminal, é pedido que informemos o clã a ser investigado, por exemplo o clã CL0023. O primeiro passo seguido pelo script é verificar a existência de famílias pertencentes ao clã selecionado no arquivo Pfam-A.clans.txt. É criado um arquivo txt, por exemplo CL0023.txt, onde são incluídos os códigos de identificação das famílias classificadas como pertencentes ao clã encontradas no arquivo Pfam-A.clans.txt (para o clã CL0023, por exemplo, temos PF00054, PF00139, PF00337 etc.). Caso não seja encontrada nenhuma família, é informado que o clã não existe e o processo é finalizado. Existindo o clã, o script segue para a segunda etapa que consiste na construção de um arquivo por família contendo todas as sequências de domínio pertencentes à família, listadas no arquivo Pfam-A.txt. Assim, para o clã CL0023, por exemplo, temos um arquivo para a família PF00054, um para PF00139 etc. O terceiro passo consiste no primeiro recorte a ser feito nos arquivos das famílias para a construção dos blocos representativos. Por exemplo, vamos supor nosso caso específico de blocos 80 × 80: as sequências de domínios que contenham 80 ou mais aminoácidos têm os seus 80 primeiros aminoácidos preservados e o restante é descartado; os domínios que possuam menos de 80 aminoácidos são inteiramente descartados. Em seguida é verificado se a família contém um mínimo de 80 domínios com 80 aminoácidos e, caso possua, o arquivo contendo 80 domínios com 80 aminoácidos é construído utilizando os 80 primeiros domínios listados. Após finalizado o processo de construção dos blocos representativos das famílias pertencentes ao clã selecionado, é informado o número de famílias que contêm tais blocos e, caso esse número não seja zero, é perguntado se desejamos manter os arquivos com os recortes ou se desejamos descartá-los. Optamos por manter apenas os clãs que contenham um mínimo de cinco famílias que possam ser representadas por blocos 80 × 80 como uma restrição adicional para o cálculo estatístico.

Capítulo 5

Estudo de Probabilidades e a construção do espaço de probabilidades

A teoria das probabilidades começou a ser desenvolvida em meados do século 17 por Blaise Pascal e Pierre de Fermat para responder uma dúvida do *Chevalier de Méré* (Antoine Gombaud) sobre o cálculo de chances de vitória em jogos de azar. Desde então mostrou-se de grande importância em diversas áreas do conhecimento, auxiliando na explicação de fenômenos naturais e no processo de tomada de decisões. Tem aplicações na Física, no cálculo de seguros e pensões, na previsão do tempo, no estudo da eficácia de medicamentos, no mercado financeiro etc.

Os fenômenos cujos efeitos em cada tentativa são diferentes, que não acompanham as leis causais, são conhecidos como experimentos aleatórios ou probabilísticos. O modelo probabilístico é utilizado no estudo de regularidades, que consiste na verificação dos resultados obtidos na replicação de um determinado experimento por diversas vezes. Também é aplicado no estudo de experimentos com muitos eventos aleatórios, com diversas incertezas quanto à ocorrência de um fato, ou a experimentos cujas consequências futuras são totalmente inexploradas e imprevisíveis [61].

Um experimento aleatório (E) é um fato ou fenômeno, dos quais os resultados eventuais só podem ser ponderados quando as ocorrências são coletivamente consideradas. O espaço dos resultados representa o conjunto de todos os possíveis resultados de um experimento. O espaço dos resultados é representado por uma letra maiúscula do alfabeto, geralmente S, e o seu número de componentes por N [62]. Usualmente estamos interessados em uma parte do espaço dos resultados, um subconjunto do espaço dos resultados do experimento, intitulado de evento.

A probabilidade pode ser definida de diversas formas, sendo especialmente importantes as definições clássica (frequencista) e Bayesiana [63]. A definição Bayesiana de probabilidade parte de uma distribuição a priori baseada em expectativas dos resultados possíveis, o que acaba introduzindo uma subjetividade ao processo, que é atualizada com resultados de observações utilizando o teorema de Bayes. A definição clássica de probabilidade, que é a mais utilizada no meio científico, é dada pela proporção entre resultados favoráveis e resultados possíveis em um espaço dos resultados previamente definido:

$$\text{Probabilidade de um evento} = \frac{\text{resultados favoráveis}}{\text{resultados possíveis}} \tag{5.1}$$

Os resultados favoráveis correspondem ao número total de ocorrências de um determinado evento, enquanto que os resultados possíveis correspondem ao espaço dos resultados. Os eventos podem ser decompostos em eventos simples e eventos compostos. Quando temos dois eventos simples distintos, que não podem ocorrer ao mesmo momento, por um anular o outro, dizemos que esses eventos são mutuamente exclusivos. Já os eventos compostos podem ocorrer simultaneamente, não são exclusivos e podem assumir valores diferentes [64].

A equação (5.1) pode ser reescrita de uma forma mais simplificada:

$$P(A) = \frac{n(A)}{n(S)}, \tag{5.2}$$

onde: $P(A)$ é a probabilidade de um evento A ($A \subset S$) ocorrer no espaço dos resultados, $n(A)$ é o número de componentes correspondentes ao evento A e $n(S)$ é o número total de casos possíveis ($n(S) = N$). Como A é um subconjunto de S no espaço de probabilidades do experimento, temos então que:

$$0 \leq n(A) \leq N. \tag{5.3}$$

Das equações (5.2) e (5.3) temos que:

$$0 \leq P(A) \leq 1. \tag{5.4}$$

Caso $P(A) = 1$, significa que o evento A sempre acontece, enquanto que $P(A) = 0$ implica que o evento A nunca acontece.

A probabilidade condicional baseia-se em associar um dado evento em decorrência do acontecimento de outro evento. Por exemplo, o acontecimento do evento A sabendo-se que ocorreu o evento B. Tal probabilidade condicional pode ser expressa por:

$$P(A|B) = \frac{P(A,B)}{P(B)}. \tag{5.5}$$

De forma similar, a probabilidade condicional do acontecimento do evento B dada a ocorrência prévia do evento A pode ser expressa por:

$$P(B|A) = \frac{P(B,A)}{p(A)} \tag{5.6}$$

Com o intuito de averiguar a distribuição de aminoácidos nas famílias de domínios de proteínas por coluna e por pares de colunas, restringiu-se o espaço de probabilidades a clãs compostos por famílias que abrangem blocos representativos de m linhas (domínios) por n colunas (aminoácidos) [3, 4, 60]. Ou seja, blocos $m \times n$ são então recortados das famílias de domínios projetadas na Figura 4.2 do capítulo anterior. No presente trabalho, continuamos com o uso do Pfam na versão 27.0 e adotamos blocos de 80 linhas × 80 colunas. Uma restrição adicional adotada, foi a restrição à utilização de clãs que contenham ao menos cinco famílias que possam ser representadas pelo bloco 80 × 80.

A Figura 5.1 abaixo representa como são construídos os blocos, com a remoção dos domínios (linhas) que são formados por menos do que n aminoácidos e a exclusão dos aminoácidos excedentes quando um domínio tem mais do que os n aminoácidos. Uma família que não contenha um mínimo de m domínios com n aminoácidos é rejeitada, e um clã que não possua um mínimo de cinco famílias que possuam os blocos $m \times n$ também não é manipulado na criação do espaço de probabilidades. Dada uma coluna j, onde $j = 1, 2,..., n$, é verificada a probabilidade de ocorrência do aminoácido a, de um total de prováveis 20 aminoácidos (a = A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y) [3].

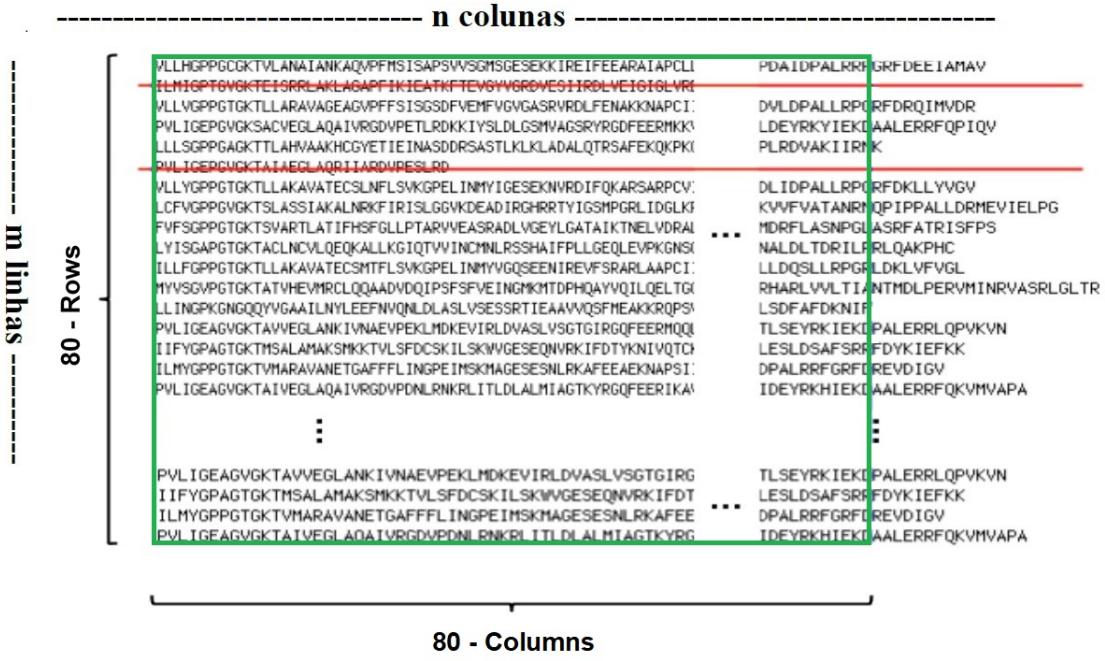

Figura 5.1: Blocos ($m \times n$) de aminoácidos representativos das famílias (em verde). Domínios com menos de n colunas são descartados (em vermelho) e os com mais de n colunas tem o excesso ($n + 1$ em diante) removido (em azul). Fonte: [4–7]

Em trabalhos anteriores, foram também utilizadas outras duas regiões (janelas) na análise das famílias de domínios de proteínas [2, 4, 60], representadas por blocos de aminoácidos do banco de dados Pfam, com as seguintes dimensões:

1. 100 linhas × 200 colunas;
2. 100 linhas × 100 colunas.

Para o bloco de 100 linhas × 200 colunas, do total de 4563 famílias distribuídas entre os 515 clãs encontrados na versão 27.0 do Pfam, apenas 1441 famílias distribuídas em 267 clãs atendem à restrição. Limitando a análise a clãs que contenham um mínimo de cinco famílias, restam 1069 famílias distribuídas entre 68 clãs (Tabela 5.1).

Restrições	Nº de famílias pertencentes a clãs	Nº de clãs
Nenhuma	4563	515
Blocos 100 × 200 (um bloco por família)	1441	267
Clãs com 5 ou mais famílias	1069	68

Tabela 5.1: Restrições inseridas para análise nos blocos 100 × 200.

Para o bloco de 100 linhas × 100 colunas, do total de 4563 famílias distribuídas entre os 515 clãs na versão 27.0 do Pfam, 2525 famílias distribuids entre 393 clãs. Limitando a análise a clãs que contenham um mínimo de cinco famílias, restam 2180 famílias distribuídas entre 146 clãs (Tabela 5.2).

Restrições	Nº de famílias pertencentes a clãs	Nº de clãs
Nenhuma	4563	515
Blocos 100 × 100 (um bloco por família)	2525	393
Clãs com 5 ou mais famílias	2180	146

Tabela 5.2: Restrições inseridas para análise nos blocos 100 × 100.

Conforme descrito no capítulo anterior, neste trabalho adotamos um bloco de 80 linhas $\times$ 80 colunas. Com esta escolha o número de clãs é reduzido a 223, contendo um total de 3223 famílias. Destes 223 clãs, 166 contêm cinco ou mais famílias, resultando em uma redução para um total de 2557 famílias.

Restrições	Nº de famílias pertencentes a clãs	Nº de clãs
Nenhuma	4563	515
Blocos 80 $\times$ 80 (um bloco por família)	3223	223
Clãs com 5 ou mais famílias	2557	166

Tabela 5.3: Restrições inseridas para análise, nos blocos 80 $\times$ 80.

5.1 Construção do espaço de probabilidades

A heterogeneidade de domínios de proteínas de uma determinada família deve ser aferida para garantir que a classificação das famílias em clãs foi realizada de forma adequada em relação às funções biológicas. Uma forma de medir a diversidade das famílias de proteínas, se dá pela verificação da distribuição de aminoácidos através do cálculo de medidas de entropia [4].

O conceito de entropia foi introduzido nas aplicações da Termodinâmica, mas é genérico o suficiente para desvendar os processos físicos que ocorrem nos sistemas, desde as galáxias até a biosfera da Terra, tendo diversas aplicações nas ciências matemáticas e biológicas. Antes de tudo, devemos notar que a inexistência de uma ordem natural ou métrica subjacente dos símbolos utilizados para representar os aminoácidos, deve ser contornada para favorecer os cálculos estatísticos mais convencionais das sequências biológicas [65].

Dando continuidade aos trabalhos realizados [4], utilizamos medidas de entropia alternativas na análise de sequências de domínios de proteínas. Devemos observar que a distribuição de domínios pode ser vista como um aspecto altamente complexo na evolução biológica. Grande parte dos domínios proteicos conhecidos são unidades estruturais discretas com 20 a 400 aminoácidos. No entanto, sabe-se que pouco menos de um terço dos domínios são descontínuos. Os domínios armazenados em bancos de dados são geralmente agrupados em famílias que, por sua vez, são aglutinadas em clãs. Com o armazenamento melhora a precisão da inferência funcional, bem como o agrupamento de novos membros. Foram utilizadas nesse trabalho, assim como em trabalhos feitos anteriormente, as sequências de domínio armazenadas na versão 27.0 do banco de dados Pfam (março de 2013) [43]. Propomos a definição do cenário para identificarmos a relevância biológica do conceito de clã usando métodos baseados em medidas de entropia. Resultados estatísticos detalhados com base nesses métodos serão demonstrados nos capítulos seguintes.

O banco de dados Pfam é organizado por um modelo probabilístico que representa a estrutura de aminoácidos através de modelos ocultos de Markov (HMM) [66, 67]. Perfis HMMs são modelos probabilísticos usados para a inferência estatística por homologia [65, 68], construída a partir de um conjunto já alinhado de sequências representativas das famílias definidas pelo perfil. Um alinhamento de semente de alta qualidade é extremamente essencial, visto que fornece a base para as frequências específicas tais como posição do aminoácido, vãos ou lacunas (*gaps*), e os parâmetros para o comprimento no perfil HMM. Os HMMs são perfis construídos, mantidos e utilizados através do pacote de software HMMER (http://hmmer.janelia.org),

projetado para detectar homólogos remotos da forma mais sensível possível, contando com a força de seus modelos de probabilidade subjacentes. No passado, essa força tinha um certo custo computacional muito significativo, mas a partir do novo projeto HMMER3, o HMMER agora é essencialmente tão rápido quanto o BLAST. Na versão 27, o banco de dados Pfam possui 14×10^3 famílias e o número de domínios neles varia de dezenas a milhares [44]. Por vezes, um único perfil HMM pode não detectar todas as homologias de uma família diversificada, portanto, várias entradas podem ser construídas para representar diferentes sequências de famílias em superfamílias (clãs).

As correlações estatísticas nas sequências de DNA não devem surpreender, visto que diversos cromossomos são sistemas um tanto complexos que envolvem inúmeras escalas diferentes. As sequências de DNA estão repletas de características em escalas pequena, intermediária e grande: as curtas distâncias, há um forte sinal de periodicidade de 3 bases nas regiões codificadoras de proteínas, ausente nas regiões não codificantes e uma base mais fraca, porém universal [69].

A análise estatística foi adotada para realizarmos testes de classificação das famílias de domínios de proteínas em clãs. Para organizar o espaço de probabilidades e realizar a análise estatística, consideramos o domínio proteico a ser representado por blocos de aminoácidos. Procuramos no banco de dados por famílias pertencentes a clãs que possam ser representadas por blocos de tamanhos especificados previamente. Cada bloco possui m linhas e n colunas, com um total de $m \cdot n$ aminoácidos. Um método alternativo para a construção dos blocos, seria considerar que para as m sequências de uma família, com n_k aminoácidos $(k = 1, ..., m)$ cada, para construir um bloco representativo $m \times n$, deveríamos excluir $(n_k - n)$ aminoácidos de cada sequência, onde n é representado por:

$$n = min(n_1, n_2, ..., n_m) \tag{5.7}$$

Seja $n_j(a)$ o número de ocorrência do aminoácido a na j-ésima coluna do bloco, onde a é um tipo de aminoácido de um total de 20 aminoácidos possíveis $(a = 1, \ldots, 20)$, representado pelo código de uma letra por uma das seguintes 20 letras: A,C,D,E,F,G,H,I,K,L,M,N,Q,R,S,T,V,W,Y. A probabilidade de ocorrência desse aminoácido é definida por:

$$p_j(a) = \frac{n_j(a)}{m}, \qquad j = 1, 2, 3, ..., n. \tag{5.8}$$

Um aminoácido a pode não aparecer ou pode aparecer entre 1 a m vezes em uma coluna.

Para ilustrar o cálculo de probabilidades realizado em um bloco representativo, considere a Tabela 5.4 abaixo contendo um bloco 5×4 como exemplo.

	Colunas (n)				
	-	1	2	3	4
Linhas (m)	1	A	R	W	E
	2	C	W	V	K
	3	C	D	W	E
	4	W	W	W	K
	5	A	R	V	K

Tabela 5.4: Exemplo de distribuição dos aminoácidos, em um bloco 5 (linhas) $\times$ 4 (colunas).

A partir do bloco representativo, podemos calcular a distribuição de probabilidades simples nas quatro colunas de acordo com a Equação (5.8):

$$p_1(A) = \frac{2}{5}, \quad p_1(C) = \frac{2}{5}, \quad p_1(W) = \frac{1}{5} \tag{5.9}$$

$$p_2(D) = \frac{1}{5}, \quad p_2(R) = \frac{2}{5}, \quad p_2(W) = \frac{2}{5} \tag{5.10}$$

$$p_3(V) = \frac{2}{5}, \quad p_3(W) = \frac{3}{5} \tag{5.11}$$

$$p_4(E) = \frac{2}{5}, \quad p_4(K) = \frac{3}{5} \tag{5.12}$$

A soma das probabilidades de ocorrência dos aminoácidos em uma coluna deve ser igual a 1:

$$\sum_a p_j(a) = 1, \qquad \forall j. \tag{5.13}$$

Efetuando a soma de probabilidades para a primeira coluna, temos:

$$\sum_{a=\mathrm{A,C,W}} p_j(a) = p_1(A) + p_1(C) + p_1(W) = \frac{2}{5} + \frac{2}{5} + \frac{1}{5} = \frac{5}{5} = 1 \qquad (5.14)$$

Para uma coluna j, as probabilidades $p_j(a)$ podem ser consideradas como componentes de um vetor p_j do espaço de probabilidades:

$$p_j^T = (p_j(A), p_j(C), p_j(D), ..., p_j(V), p_j(W), p_j(Y)) \qquad (5.15)$$

Onde p_j^T, é o vetor transposto do vetor coluna p_j.

Além das probabilidades simples, trabalhamos também com probabilidades conjuntas em pares de colunas. Seja o par de colunas j, k, a ocorrência do par de aminoácidos a, b é dada por:

$$p_{j,k}(a, b) = \frac{n_{j,k}(a, b)}{m}, \qquad \forall a, b. \qquad (5.16)$$

Onde, $n_{j,k}(a, b)$ é o número ocorrências do par de aminoácidos a,b no par de colunas j, k, respectivamente, e m é o número de linhas do bloco. Assim como no caso de probabilidades simples, os aminoácidos a e b podem assumir quaisquer valores dentro do conjunto de 20 aminoácidos: A, C, D, E, F, G, H, I, K, L, M, N, Q, R, S, T, V, W e Y. Como $p_{j,k}(a, b) = p_{j,k}(b, a)$, convencionamos trabalhar com o espaço de probabilidades conjuntas em que $j < k$, tendo então $1 \leq j \leq (n - 1)$ e $2 \leq k \leq n$. Para a probabilidade conjunta temos que:

$$1 \geq p_{j,k}(a, b) \geq 0, \qquad \forall (a, b), \forall (j, k); \qquad (5.17)$$

$$\sum_a \sum_b p_{j,k}(a, b) = 1, \qquad \forall (j, k). \qquad (5.18)$$

Para exemplificar, mostramos abaixo a identificação das probabilidades conjuntas para o bloco (5×4) da Tabela 5.4, em que $j = 1, 2, 3$ e $k = 2, 3, 4$.

Probabilidades conjuntas do par de colunas 1,2:

$$p_{1,2}(A, R) = \frac{2}{5}; \quad p_{1,2}(C, D) = \frac{1}{5}; \quad p_{1,2}(C, W) = \frac{1}{5}; \quad p_{1,2}(W, W) = \frac{1}{5}; \tag{5.19}$$

Probabilidades conjuntas do par de colunas 1,3:

$$p_{1,3}(A, V) = \frac{1}{5}; \;\; p_{1,3}(A, W) = \frac{1}{5}; \;\; p_{1,3}(C, V) = \frac{1}{5}; \;\; p_{1,3}(C, W) = \frac{1}{5} \;\; p_{1,3}(W, W) = \frac{1}{5} \tag{5.20}$$

Probabilidades conjuntas do par de colunas 1,4:

$$p_{1,4}(A, E) = \frac{1}{5}; \;\; p_{1,4}(A, K) = \frac{1}{5}; \;\; p_{1,4}(C, E) = \frac{1}{5}; \;\; p_{1,4}(C, K) = \frac{1}{5}; \;\; p_{1,4}(W, K) = \frac{1}{5} \tag{5.21}$$

Probabilidades conjuntas do par de colunas 2,3:

$$p_{2,3}(D, W) = \frac{1}{5}; \;\; p_{2,3}(R, V) = \frac{1}{5}; \;\; p_{2,3}(R, W) = \frac{1}{5}; \;\; p_{2,3}(W, V) = \frac{1}{5}; \;\; p_{2,3}(W, W) = \frac{1}{5} \tag{5.22}$$

Probabilidades conjuntas do par de colunas 2,4:

$$p_{2,4}(D, E) = \frac{1}{5}; \quad p_{2,4}(R, E) = \frac{1}{5}; \quad p_{2,4}(R, K) = \frac{1}{5}; \quad p_{2,4}(W, K) = \frac{2}{5}; \tag{5.23}$$

Probabilidades conjuntas do par de colunas 3,4:

$$p_{3,4}(V, K) = \frac{2}{5}; \qquad p_{3,4}(W, E) = \frac{2}{5}; \qquad p_{3,4}(W, K) = \frac{1}{5}. \tag{5.24}$$

Observa-se que para cada par de colunas formado com o bloco 5×4 escolhido, ocorreram 3 a 5 pares de aminoácidos distintos. Entretanto, como existem 20 tipos de aminoácidos, temos um total de 400 pares de aminoácidos que precisam ser considerados na construção da matriz de probabilidade conjunta. As probabilidades desses pares que não ocorreram no bloco selecionado são iguais a zero. Por exemplo, $p_{1,2}(A, A) = 0$, $p_{1,2}(A, C) = 0$, $p_{1,2}(Y, W) = 0$, $p_{1,2}(Y, Y) = 0$ etc. Dessa forma,

podemos escrever a matriz $p_{j,k}$ abaixo, contendo todas as probabilidades conjuntas possíveis para as combinações dos 20 tipos de aminoácidos.

$$p_{j,k} = \begin{pmatrix} p_{j,k}(A,A) & p_{j,k}(A,C) & p_{j,k}(A,D) & \dots & p_{j,k}(A,W) & p_{j,k}(A,Y) \\ p_{j,k}(C,A) & p_{j,k}(C,C) & p_{j,k}(C,D) & \dots & p_{j,k}(C,W) & p_{j,k}(C,Y) \\ p_{j,k}(D,A) & p_{j,k}(D,C) & p_{j,k}(D,D) & \dots & p_{j,k}(D,W) & p_{j,k}(D,Y) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{j,k}(W,A) & p_{j,k}(W,C) & p_{j,k}(W,D) & \dots & p_{j,k}(W,W) & p_{j,k}(W,Y) \\ p_{j,k}(Y,A) & p_{j,k}(Y,C) & p_{j,k}(Y,D) & \dots & p_{j,k}(Y,W) & p_{j,k}(Y,Y) \end{pmatrix} \tag{5.25}$$

Após esta apresentação dos conceitos de probabilidades simples e conjunta para a caracterização da distribuição dos aminoácidos no bloco $(m \times n)$, o teorema de Bayes pode então ser expresso como:

$$p_{j,k}(a|b) = \frac{p_{j,k}(b,a)}{p_k(b)} = \frac{p_{k,j}(b|a)p_j(a)}{p_k(b)} \tag{5.26}$$

Onde, pela definição geral do teorema de Bayes, $p_{j,k}(a|b)$ e $p_{k,j}(b|a)$ são as probabilidades condicionais, ou seja, a probabilidade de ocorrência do aminoácido a na coluna j dado que o aminoácido b ocorreu na coluna k e a probabilidade de ocorrência do aminoácido b na coluna k dado que o aminoácido a ocorreu na coluna j, respectivamente [3].

A análise das distribuições de probabilidades dos blocos $(m \times n)$ que representam as famílias de domínios de proteínas são realizadas através da inferência das medidas de entropia que quantificam e qualificam a indeterminação das distribuições das probabilidades. Assim, obteremos um escalar por coluna ou por par de colunas, ao invés de um vetor ou uma matriz de probabilidades, respectivamente. Medidas alternativas de entropia foram utilizadas em trabalhos anteriores fazendo o uso de funções de variáveis aleatórias (probabilidades). As probabilidades de ocorrência de aminoácidos foram verificadas em blocos retangulares, os quais são selecionados de famílias de domínios de proteínas [2–4, 60].

5.2 Noções de Termodinâmica

A Termodinâmica é uma parte da Física que realiza estudos sobre as transferências de energia, buscando o entendimento da relação existente entre calor, energia e trabalho. Verifica assim a conversão da quantidade de calor em trabalho. A Termodinâmica foi, a princípio, elaborada por pesquisadores que visavam uma maneira de aprimorar as máquinas térmicas, no período da Revolução Industrial, aperfeiçoando a eficiência de suas operações [70]. Na área biológica, uma das aplicações da Termodinâmica é o estudo das transições de energia que acontecem nas moléculas ou nos diversos conjuntos de moléculas.

A Primeira Lei da Termodinâmica afirma que a variação de energia em um sistema durante uma transformação é convertida em trabalho e/ou dissipada na forma de calor. Visa o princípio da conservação de energia, um dos princípios mais relevantes da Física. Essa conservação de energia acontece sob as formas de calor e de trabalho. É por isso que a primeira lei é chamada de *Princípio de Conservação de Energia* ou simplesmente *Lei de Conservação de Energia*, o que significa que a energia não pode ser criada nem destruída, mas sim transformada em várias formas. A Primeira Lei da Termodinâmica é expressa pela fórmula abaixo [71]:

$$\triangle U = Q + \tau \tag{5.27}$$

Onde,

- $\triangle U$- Variação de energia interna do sistema (U_{final} - $U_{inicial}$).
- Q - Calor absorvido pelo sistema (ou fornecido pelo sistema).
- τ - Trabalho realizado pelo sistema (ou o trabalho realizado por forças externas no sistema).

Desta forma, a primeira lei pode ser enunciada como: a variação de energia interna de um sistema ($\triangle U$) durante uma transformação é igual a soma do calor absorvido pelo sistema (Q) mais o trabalho realizado pelo sistema (τ).

A Segunda Lei da Termodinâmica trata da espontaneidade de processos físicos e impõe limitações na transformação de energia. O postulado de Lorde Kelvin diz que o calor absorvido por um sistema de uma fonte de calor à mesma temperatura, não é inteiramente convertido em trabalho; enquanto que o postulado de Clausius diz que não existe transformação cujo único resultado seja conduzir calor de um corpo para um segundo corpo que esteja a maior temperatura.

O conceito de Entropia pode ser entendido como a medida do grau de desordem de um sistema, sendo uma medida da não disponibilidade de energia, e diz respeito ao número de estados dinâmicos correspondentes a um dado estado termodinâmico. A nível molecular, a entropia expressa o número total de possíveis maneiras nas quais os átomos de um objeto podem ser arranjados. É uma grandeza relacionada à segunda Lei da Termodinâmica. Para sistemas isolados, após qualquer transformação ocorrida, a entropia final do estado do sistema não pode ser menor que a entropia inicial. O termo "desordem" não deve ser entendido como "bagunça", mas sim como a forma de organização de um sistema. Por exemplo, considere que tenhamos dois potes, um contendo pequenas bolas de cor amarela e o outro contendo o mesmo tipo de bolas, porém de cor vermelha (Figura 5.2). Com cuidado, dispomos as bolas vermelhas do segundo pote por cima das bolas amarelas no primeiro pote, de forma que no final desse processo as bolas no pote estarão separadas pela cor: a parte de baixo do pote contendo apenas bolas amarelas e a de cima apenas bolas vermelhas. Em seguida, agitamos o pote durante um intervalo de tempo de forma que as bolas comecem a se misturar e no final do processo não exista mais a separação inicial. Mesmo que esse processo seja repetido diversas vezes e por longos intervalos de tempo, a probabilidade de que as bolas voltem à organização inicial (separadas por cor) é desprezível. Ou seja, o sistema ordenado (bolas separadas por cor) se tornou um sistema desordenado (bolas misturadas), mas não podemos limitar o conceito de entropia a este exemplo. Podemos verificar que o processo de mistura apresenta

uma orientação natural que corresponde a um aumento da desordem do sistema, indicando um aumento da Entropia. Temos então:
$\triangle S = S_{final} - S_{inicial} > 0$, onde S é a Entropia.

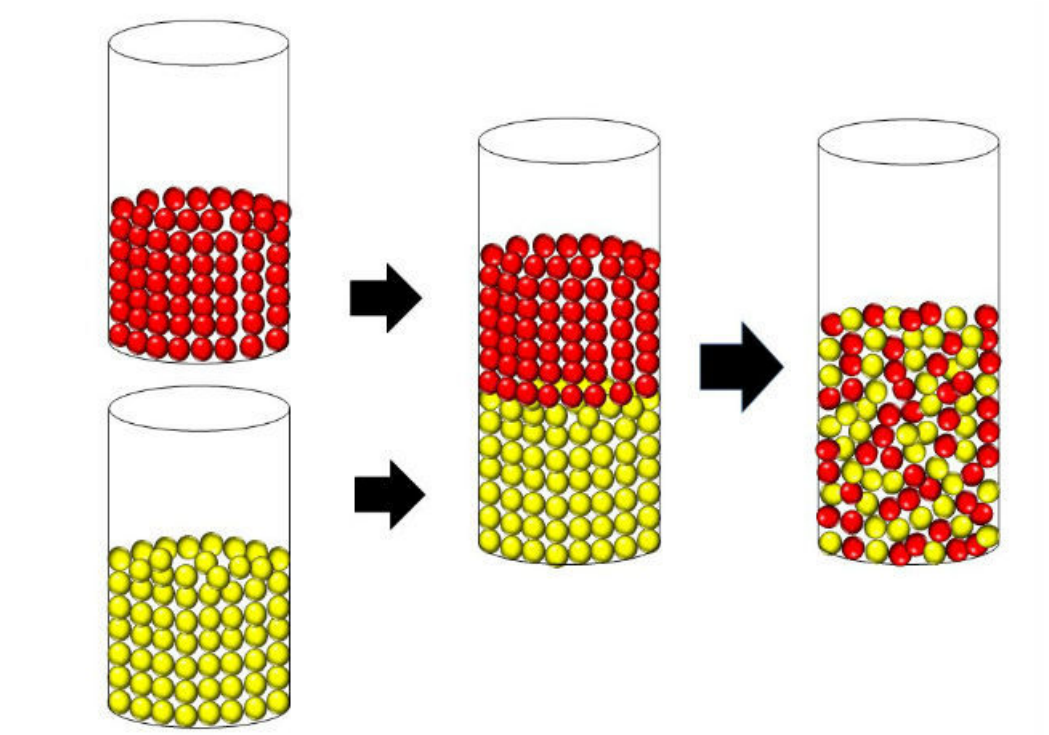

Figura 5.2: Exemplo do processo de mistura de bolas de cores diferentes. Fonte: `https://www.todamateria.com.br/entropia/`

O conceito de Entropia foi introduzido pelo engenheiro e pesquisador francês Nicolas Sadi Carnot. Em suas diversas pesquisas sobre transformação da energia mecânica em energia térmica, verificou que seria impossível a existência de uma máquina térmica com eficiência total. A Primeira Lei da Termodinâmica diz respeito à conservação de energia. Isso quer dizer que nos processos físicos a energia não se perde, ela se converte de um tipo em outro. Por exemplo, uma máquina utiliza energia para realizar trabalho e nesse processo a máquina aquece. Ou seja, a energia mecânica está sendo transformada em energia térmica. A energia térmica não se transforma novamente em energia mecânica (se isso acontecesse a máquina nunca deixaria de funcionar), portanto, o processo é irreversível. Mais tarde, Lorde Kelvin complementou as pesquisas de Carnot sobre a irreversibilidade dos processos termodinâmicos, dando origem às bases da Segunda Lei da Termodinâmica. Rudolf Clausius foi o primeiro a empregar o termo Entropia em 1865. A Entropia é a medida da quantidade de energia térmica que não pode ser revertida em energia mecânica (não pode realizar trabalho), em uma determinada temperatura. A variação de entropia ($\triangle S$) de um sistema devido a uma transformação é obtida pela razão entre o calor transferido e a temperatura do sistema:

$$\triangle S = \frac{\triangle Q}{T} \tag{5.28}$$

Sendo,

- $\triangle S$: variação da entropia
- $\triangle Q$: calor transferido
- T: temperatura

5.3 Medidas de Entropia Sharma-Mittal e Havrda-Charvat

A primeira medida de Entropia a ser apresentada nesta seção é a entropia Sharma-Mittal (SM) [13], que é definida para distribuições discretas de probabilidade simples e conjunta [1, 4, 6, 7, 72], respectivamente, como:

$$(SM)_j(r,s) = -\frac{1}{1-r}\left(1 - \left(\sum_a \left(p_j(a)\right)^s\right)^{\frac{1-r}{1-s}}\right) \tag{5.29}$$

$$(SM)_{jk}(r,s) = -\frac{1}{1-r}\left(1 - \left(\sum_a \sum_b \left(p_{jk}\big(a,b\big)\right)^s\right)^{\frac{1-r}{1-s}}\right) \tag{5.30}$$

Onde, r e s são parâmetros adimensionais e $p_j(a)$ e $p_{j,k}(a,b)$ são variáveis aleatórias (as probabilidades simples e conjunta) de ocorrência de aminoácidos nas colunas j e k.

A partir da entropia Sharma-Mittal, podemos obter outras medidas de entropia de apenas um parâmetro conhecidas no meio científico: Havrda-Charvat [14], Renyi [73], Landsberg-Vedral [74] e Gaussiana "não-extensiva" [75]. Todas essas entropias têm a entropia de Gibbs-Shannon como limite. A entropia Havrda-Charvat é obtida ao fazermos o parâmetro r igual a s nas equações (5.29) e (5.30) para distribuições de probabilidade simples e conjunta, respectivamente [7, 72]:

$$(HC)_j(s) = -\frac{1}{1-s}\left(1 - \left(\sum_a \left(p_j\big(a\big)\right)^s\right)\right) \tag{5.31}$$

$$(HC)_{j,k}(s) = -\frac{1}{1-s}\left(1 - \left(\sum_a \sum_b \left(p_{j,k}\big(a,b\big)\right)^s\right)\right) \tag{5.32}$$

Para exemplificar o cálculo de entropia Havrda-Charvat, consideremos o bloco apresentado na Tabela 5.4. O valor de entropia de probabilidade simples da primeira coluna com o parâmetro s igual a 0.1, é obtido da seguinte forma:

$$\begin{aligned}(HC)_1(0.1) &= \frac{(p_1(A))^{0.1} + (p_1(C))^{0.1} + (p_1(I))^{0.1} + (p_1(T))^{0.1} + (p_1(W))^{0.1} - 1}{1-0.1} \\ &= \frac{5(\frac{1}{5})^{0.1} - 1}{0.9} \approx 3.62\end{aligned}$$

A entropia de Gibbs-Shannon é obtida ao tomarmos o limite de s tendendo a 1 nas equações (5.30) e (5.31) para os casos de probabilidade simples e probabilidade conjunta, respectivamente:

$$\begin{aligned}(GS)_j = \lim_{s\to 1}(HC)_j(s) &= \lim_{s\to 1} -\frac{\left(1 - \sum_a \left(p_j(a)\right)^s\right)}{1-s} = \lim_{s\to 1} -\frac{-\sum_a \left(p_j(a)\right)^s \log\left(p_j(a)\right)}{-1} \\ &= -\sum_a p_j(a)\log\left(p_j(a)\right)\end{aligned} \tag{5.33}$$

$$
\begin{aligned}
(GS)_{j,k} = \lim_{s\to 1}(HC)_{j,k}(s) &= \lim_{s\to 1} -\frac{\left(1 - \left(\sum_a \sum_b \Big(p_{j,k}\big(a,b\big)\Big)^s\right)\right)}{1-s} \\
&= \lim_{s\to 1} -\frac{-\sum_a \sum_b \Big(p_{j,k}\big(a,b\big)\Big)^s \log\Big(p_{j,k}\big(a,b\big)\Big)}{-1} \\
&= -\sum_a \sum_b p_{j,k}(a,b) \log\Big(p_{j,k}\big(a,b\big)\Big) \qquad (5.34)
\end{aligned}
$$

Na próxima seção, apresentamos os histogramas das médias de entropia Havrda-Charvat das famílias classificadas em clãs para valores do parâmetro s igual a 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3. Dado um valor do parâmetro s, as médias de entropia para probabilidade simples e para probabilidade conjunta de um família são definidas, respectivamente, como:

$$
\Big\langle (HC)(s) \Big\rangle = \frac{\sum_{j=1}^{n} (HC)_j(s)}{n} \qquad (5.35)
$$

$$
\Big\langle \big\langle (HC)(s) \big\rangle \Big\rangle = \frac{\sum_{j=1}^{n-1} \Big[\sum_{k=j+1}^{n} (HC)_{j,k}(s) \Big]}{n(n-1)/2} \qquad (5.36)
$$

O termo no denominador da equação (5.35) advém da combinação $\binom{n}{2}$ de todos os pares de coluna possíveis sem repetição. Para os blocos representativos 80×80, temos $n = 80$, o que resulta em $\binom{80}{2} = \frac{80!}{2!(80-2)!} = \frac{80\cdot 79}{2} = 3160$ pares de colunas.

5.4 Histogramas de Médias de Entropia Havrda-Charvat

A fim de verificarmos o comportamento das distribuições das médias de entropia Havrda-Charvat, foram construídos histogramas de densidade para treze valores do parâmetro s, tanto para o caso de distribuições de probabilidade simples quanto para distribuições de probabilidade conjunta. Nas Figuras 5.3, 5.4 e 5.5 abaixo, apresentamos os treze histogramas (um para cada valor do parâmetro s), assim como a melhor curva ajustada ao histograma (curva sólida em preto) e a curva gaussiana aproximada (curva pontilhada em vermelho) construída com o valor médio ($\bar{x}$) e desvio padrão (σ) das médias de entropia Havrda-Charvat de probabilidade

simples. As médias foram calculadas para as 2557 famílias que pertencem aos 166 clãs escolhidos pelo processo descrito no início do capítulo (Tabela 5.3).

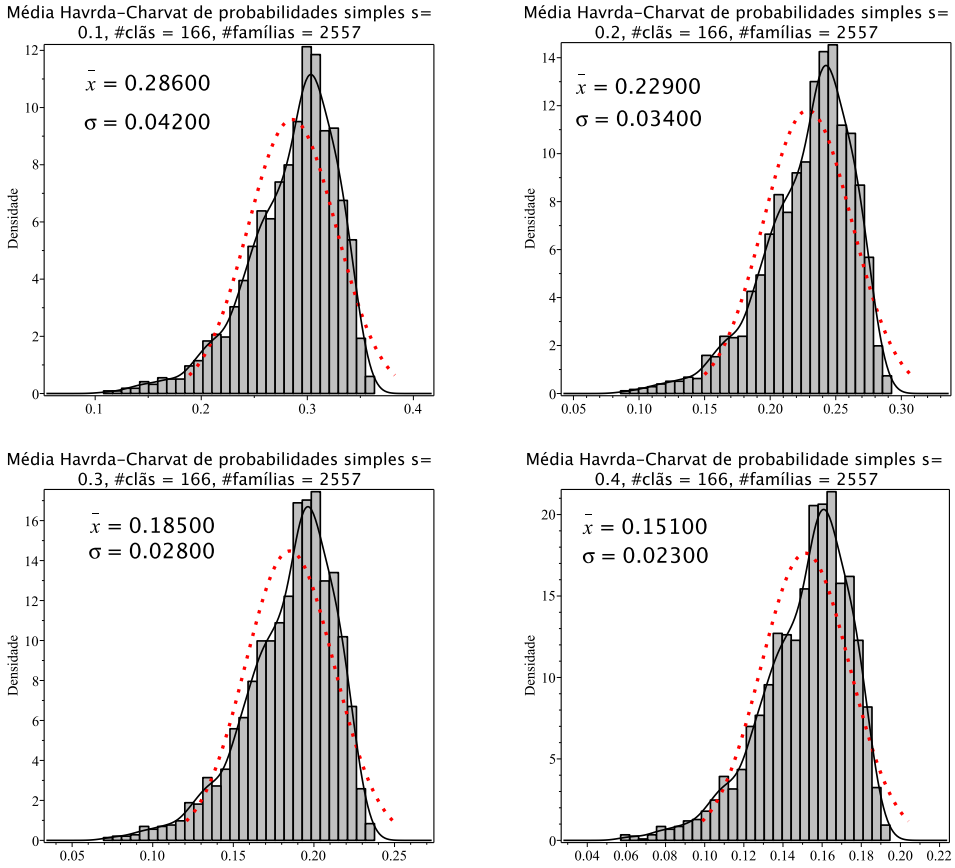

Figura 5.3: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade simples para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

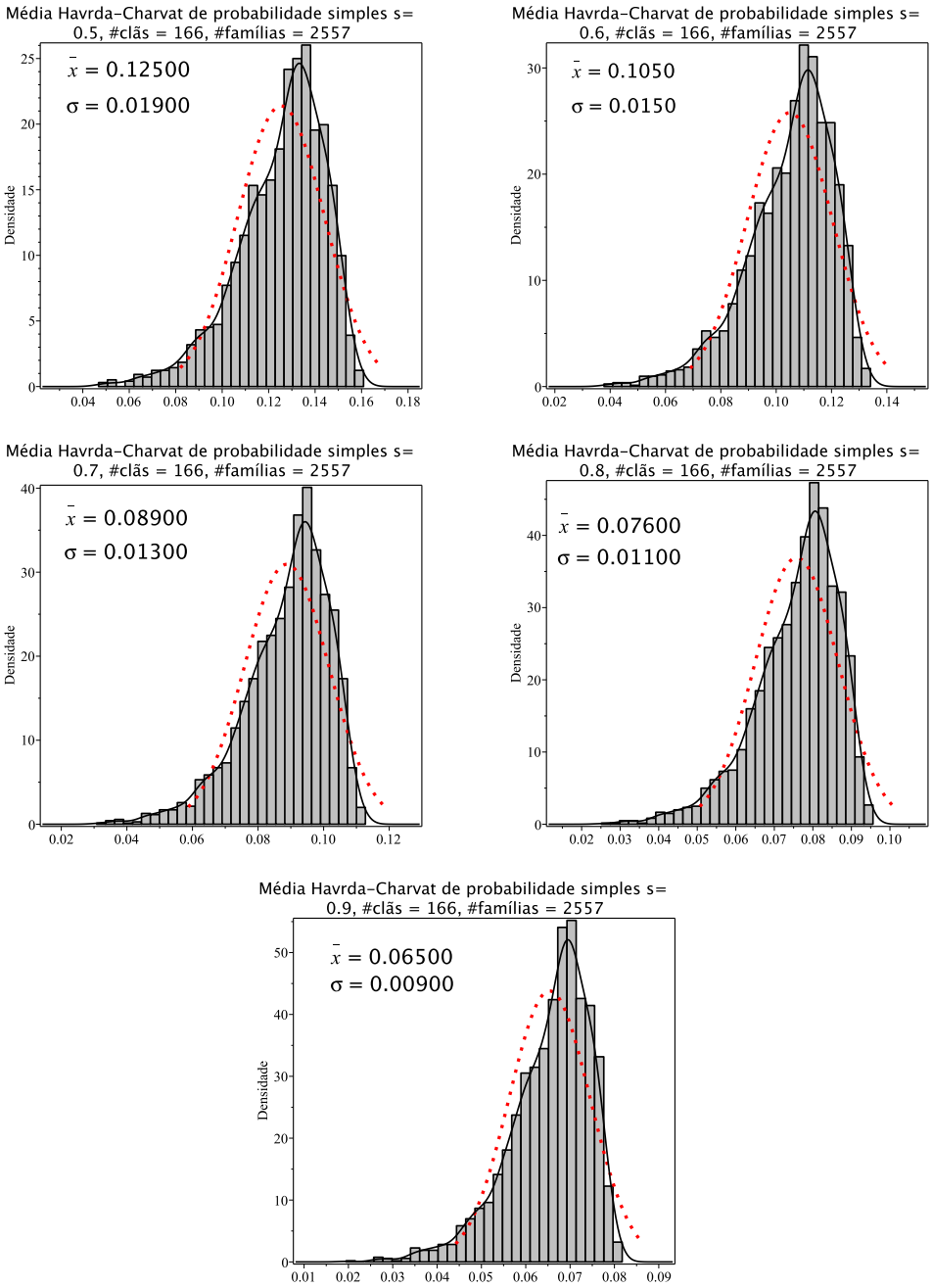

Figura 5.4: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade simples para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

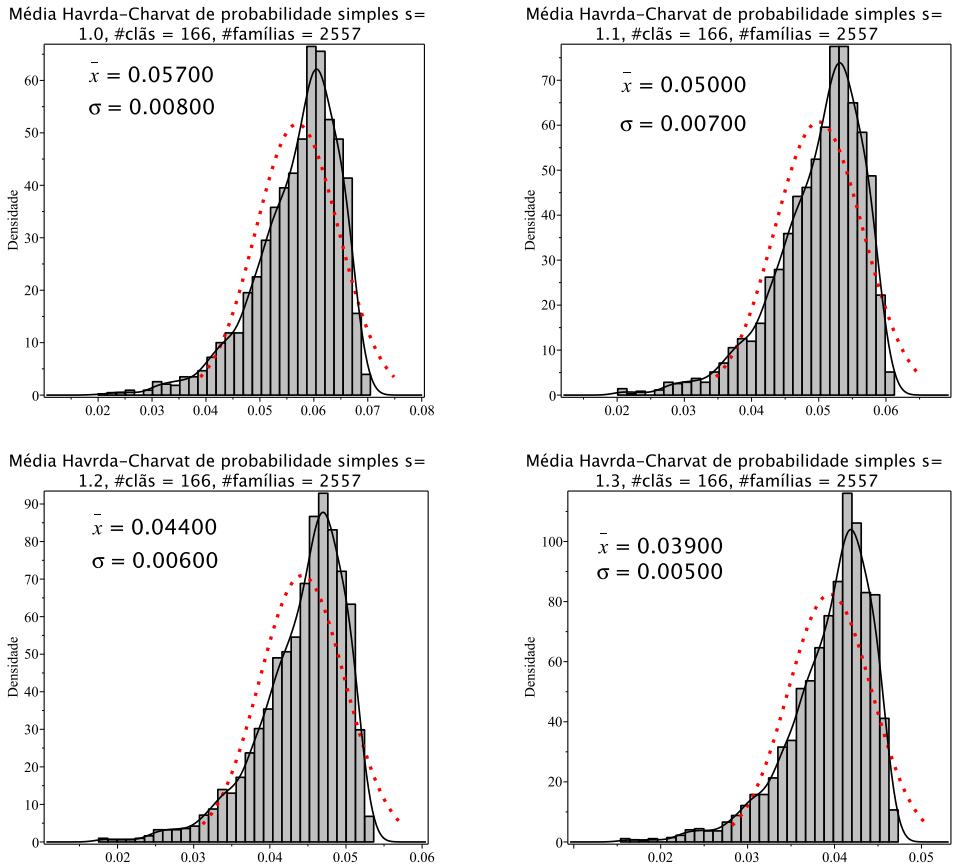

Figura 5.5: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade simples para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

As Figuras 5.6, 5.7 e 5.8 abaixo contêm os treze histogramas das médias de entropia Havrda-Charvat de probabilidade conjunta, assim como a melhor curva ajustada e a gaussiana aproximada para cada valor do parâmetro s. Novamente, as médias foram calculadas para as 2557 famílias pertencentes a 166 clãs.

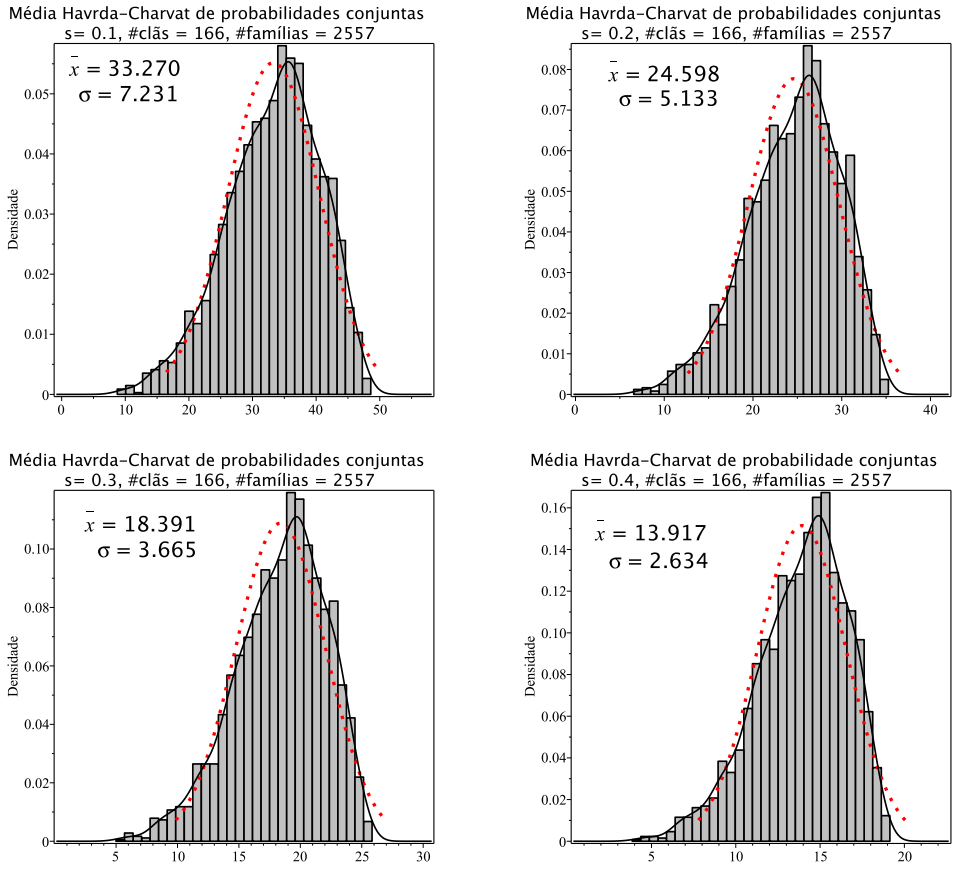

Figura 5.6: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

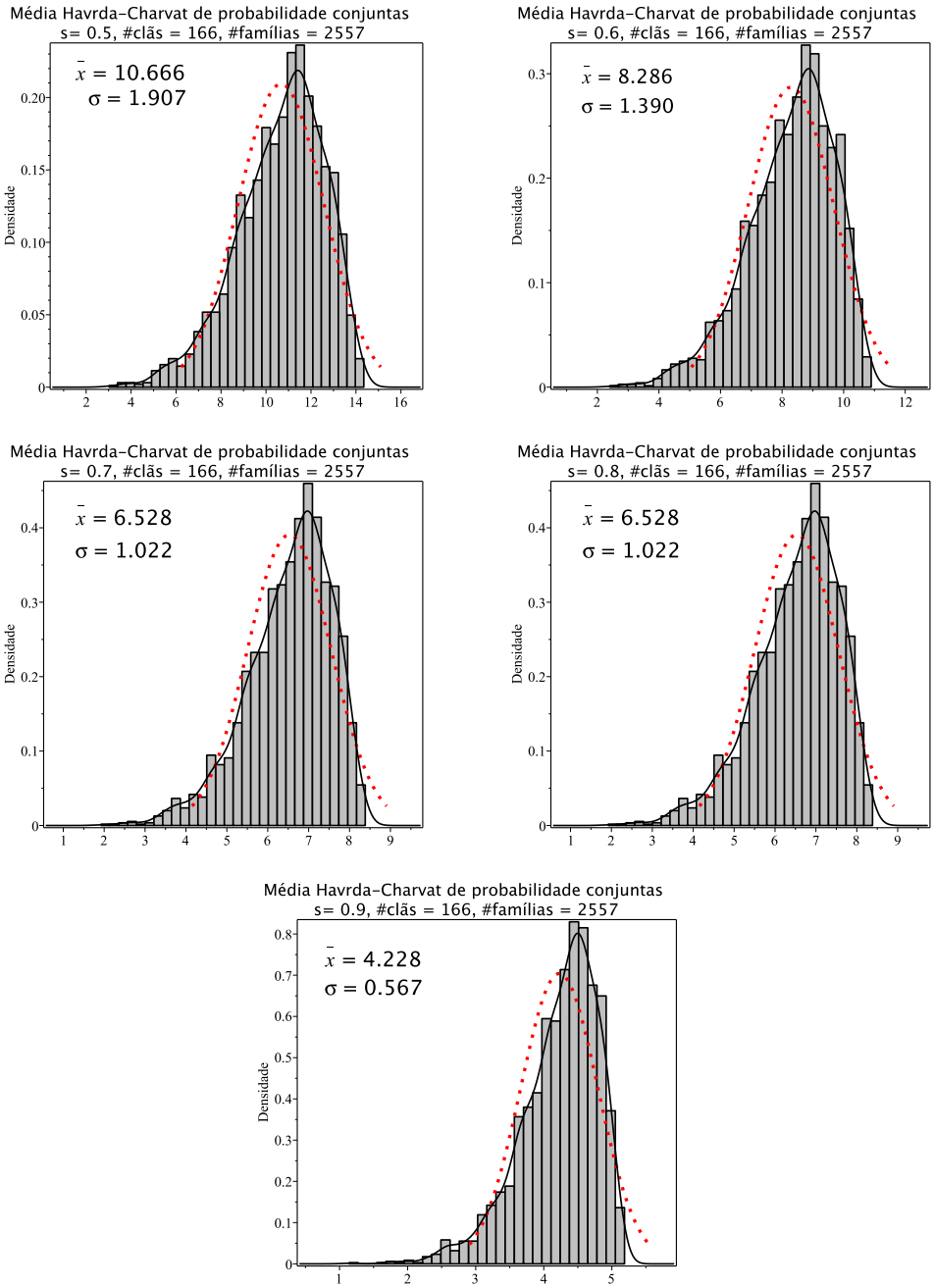

Figura 5.7: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

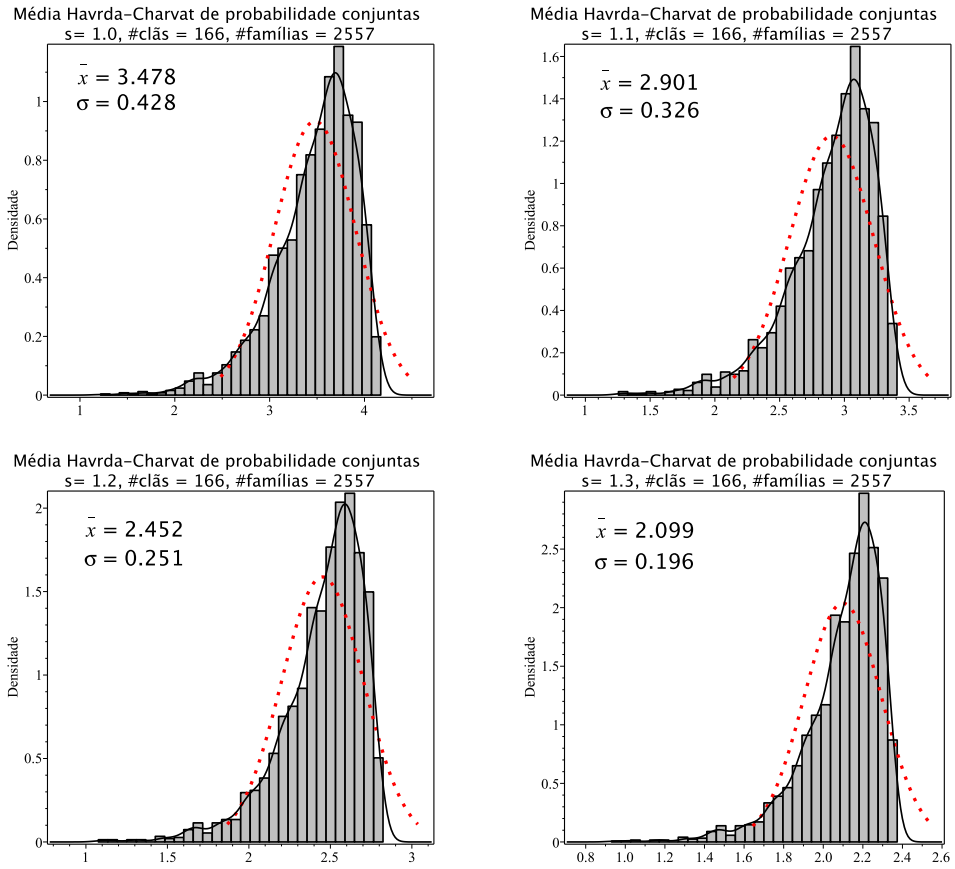

Figura 5.8: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

Nos histogramas das médias de entropia Havrda-Charvat acima, tanto para probabilidades simples quanto para probabilidades conjuntas, podemos perceber que o comportamento dos mesmos em relação aos valores do parâmetro s, sofrem pequenas alterações em sua forma e que não são tão simétricos ao ponto de se aproximarem de uma distribuição gaussiana. O valor médio, $\bar{x}$, e o desvio padrão, σ, das médias diminuem conforme o valor do parâmetro s aumenta, tanto para probabilidade simples quanto para probabilidade conjunta. A diminuição do desvio padrão implica em curvas cada vez mais estreitas e mais altas.

Capítulo 6

Sistemas operacionais e computacionais, linguaguem de programação, desafios e recomendações

O desenvolvimento de máquinas computacionais a partir de meados do século XX permitiu que cálculos complexos envolvendo um grande volume de dados pudessem ser realizados em um tempo muito menor e menos suscetíveis à propagação de erros, desde que se conheça e se verifique todas as etapas de construção de algoritmos e realização dos cálculos. Diversas linguagens de programação e softwares de computação algébrica e/ou numérica foram desenvolvidos com o intuito de otimizar o tempo de processamento e prover ferramentas de visualização dos resultados para a análise dos dados.

Para a realização dos cálculos das probabilidades simples e conjuntas das famílias de domínios de proteínas, em pesquisas realizadas anteriormente, foi utilizado o software de computação algébrica Maple. Sua escolha se deu por ser um "grande livro de cálculo" com diversas funções algébricas predefinidas e implementadas, pela possibilidade de definir funções e procedimentos através da escrita de códigos, e pela simplicidade em gerar gráficos para a visualização. Além disso, o Maple tem uma interface simples e intuitiva, uma documentação bem completa de todas as suas funcionalidades e fóruns oficiais, o que permite que seja rapidamente dominado por pesquisadores e estudantes.

O Maple é uma solução computacional interessante por combinar em um ambiente integrado, facilidades para descrever, visualizar, analisar e resolver problemas matemáticos. Oferece também a alternativa de desenvolvimento de gráficos em duas e

três dimensões para obter uma melhor compreensão e percepção dos dados analisados, sendo muito utilizados nas áreas de matemática, física, engenharias, dentre outras. O uso das características computacionais do Maple nos problemas de difícil solução permite que eles sejam solucionados de maneira fácil e com maior rapidez, ajudando na evolução da pesquisa em desenvolvimento. Além das características e vantagens descritas com o uso do Maple, ele ainda oferece a opção de ser executado em diferentes plataformas. Dessa forma é possível instalar e configurar o Maple em computadores que utilizam os seguintes sistemas operacionais: Windows da Microsoft, IOS da Apple e nas diversas distribuições do Linux.

No entanto, o software Maple se mostrou muito lento para realizar os cálculos estatísticos com longas cadeias de caracteres de aminoácidos extraídas de bancos de dados de proteínas. Para a manipulação de arquivos de texto muito grandes, uma das linguagens de programação mais apropriada é o Perl. O Perl é uma linguagem bastante utilizada por biólogos e por programadores que necessitam realizar tratamento de *strings* (sequências de caracteres como, por exemplo, palavras, frases ou domínios de proteínas).

A linguagem Perl foi utilizada para realizar os cálculos de probabilidade de ocorrência de aminoácidos e de medidas de entropia das famílias de domínios de proteínas contidas no banco de dados Pfam. Muitas operações elementares foram feitas em uma estrutura de array multidimensional.

Para a construção, edição e ajustes nos cálculos, foi escolhido um editor de código-fonte mais robusto do que os que são utilizados no meio comercial/industrial, sendo bem próximo de uma IDE (Integrated Development Environment), o Sublime Text é um editor de código-fonte multiplataforma e *shareware* com uma interface de programação de aplicativos (API) para diversas linguagens. Ele suporta nativamente muitas linguagens de programação e linguagens de marcação. Além disso, funções podem ser adicionadas por usuários com plug-ins, geralmente criados pela comunidade de desenvolvedores e mantidos sob licenças de software livre. Inicialmente, o programa foi pensado para ser uma extensão do *Vim*, um clone do programa editor de textos *Vi* para Unix. O Sublime Text é um editor de texto projetado para ser simples, rápido, flexível e fácil de usar. O seu uso teve uma contribuição muito significativa na construção dos códigos. Com versões para Linux, OS X e Windows, o Sublime Text consegue proporcionar uma experiência idêntica nas três plataformas, mantendo o padrão visual uniforme e um ótimo desempenho.

Algumas adaptações dos métodos utilizados foram feitas para serem usadas com o sistema computacional Perl. Um plug-in foi desenvolvido para auxiliar na compilação e execução (*build*) dos códigos responsáveis pelos cálculos. Isso possibilitou a não utilização do terminal do sistema operacional para realizar as tarefas de execução citadas neste parágrafo. Em uma única tela podemos acompanhar o que foi implementado junto com os resultados da implementação do código.

```
1  {
2      ''cmd'': ['perl', '-w', '$file'],
3      'file_regex': '.* at (.*) line ([0-9]*)',
4      'selector': 'source.perl'
5  }
```

Capítulo 7

Conclusão

Neste trabalho foi considerado que a classificação de famílias de domínios de proteínas em clãs corresponda a algo que de fato ocorre na Natureza, apesar dos métodos de identificação e de reconhecimento de ancestralidade comum ainda não sejam totalmente precisos. Um dos objetivos dos métodos aplicados nesse trabalho, é identificar uma maneira de caracterizar o banco de dados de proteínas através de medidas de entropia, de modo a fornecer uma discussão sólida a ser centrada na otimização de parâmetros que resulte em uma medida de entropia média conveniente para representar todo o banco de dados de proteínas. Porém, é necessário que mais blocos de famílias sejam utilizados. Além dos blocos (100×200) e (100×100), utilizados em pesquisas anteriores realizadas pelo grupo de trabalho, e com a utilização do bloco (80×80), que foi a base desta pesquisa, os testes estatísticos realizados [7] indicaram que não podemos desconsiderar a existência de famílias aglutinadas na composição dos clãs. Os histogramas apresentados no Capítulo 5 mostram que as distribuições das médias de entropia Havrda-Charvat para diferentes valores do parâmetro s, não são muito próximas de uma distribuição normal. Os testes de hipóteses realizados devem então se apoiar na robustez das ferramentas de análise estatística adotadas. Uma forma de tratar os dados obtendo curvas mais próximas de uma distribuição gaussiana é através da adoção do Símbolo de Jaccard [7, 60, 76]. Para trabalhos futuros, as seguintes alterações devem ser investigadas:

1. Aumentar o limite inferior de famílias com blocos (80×80). Um número maior do que cinco famílias deve ser mais adequado para caracterizar um clã;

2. Realizar os testes para as outras medidas de entropia (Renyi, Landsberg-Vedral, Gaussiana "não-extensiva" e Sharma-Mittal) e para o Símbolo de Jaccard associado a elas, que podem ser mais adequadas para lidar com o espaço dos resultados;

3. Utilizar outras versões do banco de dados Pfam para compararmos com a

versão 27.0 que foi utilizada neste trabalho. Desta forma, podemos realizar um estudo temporal, comportamental e evolutivo das famílias de proteínas;

4. Trabalhar com outros blocos representativos como, por exemplo recortes de (80 $\times$ 40) ou (40 $\times$ 40) e compará-los com os obtidos pelos blocos já calculados (100 $\times$ 200), (100 $\times$ 100) e (80 $\times$ 80);

5. Além disso, códigos podem ser otimizados com modificações nas aplicações dos cálculos e com o uso de técnicas de programação paralela.

Capítulo 8

Informações adicionais

8.1 Análise comparativa dos histogramas das médias em diferentes janelas

Foram feitos os histogramas das médias de entropia Havrda-Charvat (HC) em trabalhos anteriores nas janelas (100 × 100) e (100 × 200) [3], selecionando os parâmetros de s iguais a 0.1, 0.5, 0.9 e 1.0 (limite Gibbs-Shannon). Todos os histogramas são de densidade, o que significa que a área de cada barra é proporcional ao número de elementos (valores de entropia dos pares de colunas) presentes no intervalo de valores (largura da barra), de forma que a área total (a soma das áreas de todas as barras) é igual a 1. Assim, com as devidas aproximações e suavizações, obtemos uma curva que se ajusta à distribuição dos dados.

Nas Figuras 8.1 a 8.8 são comparados os histogramas de densidade das distribuições de médias de entropia Havrda-Charvat de probabilidade conjunta construídos a partir dos blocos (80×80) com os construídos a partir dos blocos (100×100) e (100×200). As curvas sólidas em preto são as que melhor se ajustam aos histogramas, enquanto que as curvas pontilhadas correspondem às curvas gaussianas construídas utilizando as médias e os desvios padrão das distribuições, indicados nos histogramas.

Na Figura 8.1 temos no lado esquerdo o histograma construído a partir dos blocos (80 × 80) e no lado direito o histograma construído a partir dos blocos (100 × 100), para o valor de parâmetro s igual a 0.1.

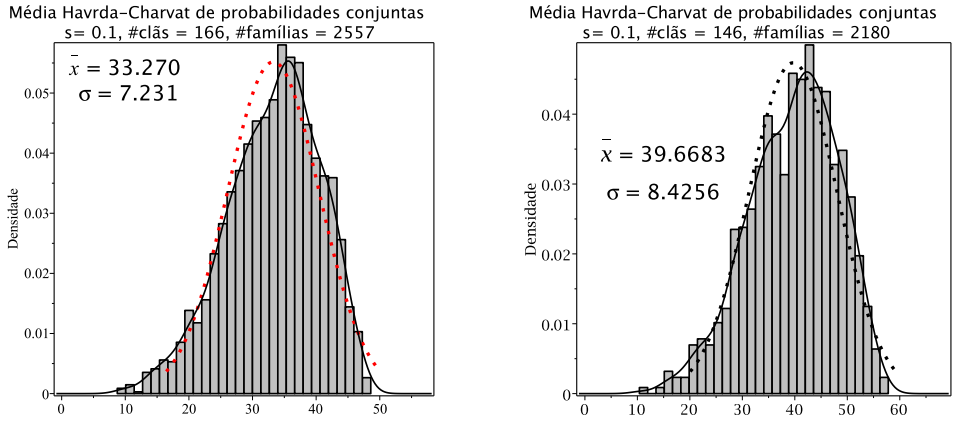

Figura 8.1: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual a 0.1 nas janelas (80×80) e (100×100).

Na Figura 8.2 temos no lado esquerdo o histograma construído a partir dos blocos (80×80) e no lado direito o histograma construído a partir dos blocos (100×200), para o valor de parâmetro s igual a 0.1.

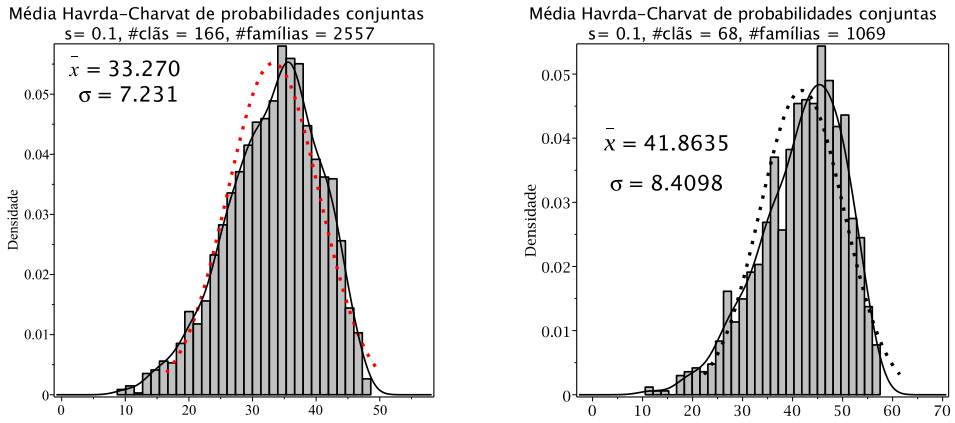

Figura 8.2: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual a 0.1 nas janelas (80×80) e (100×200).

Na Figura 8.3 temos no lado esquerdo o histograma construído a partir dos blocos (80×80) e no lado direito o histograma construído a partir dos blocos (100×100), para o valor de parâmetro s igual a 0.5.

Na Figura 8.4 temos no lado esquerdo o histograma construído a partir dos blocos (80×80) e no lado direito o histograma construído a partir dos blocos (100×200), para o valor de parâmetro s igual a 0.5.

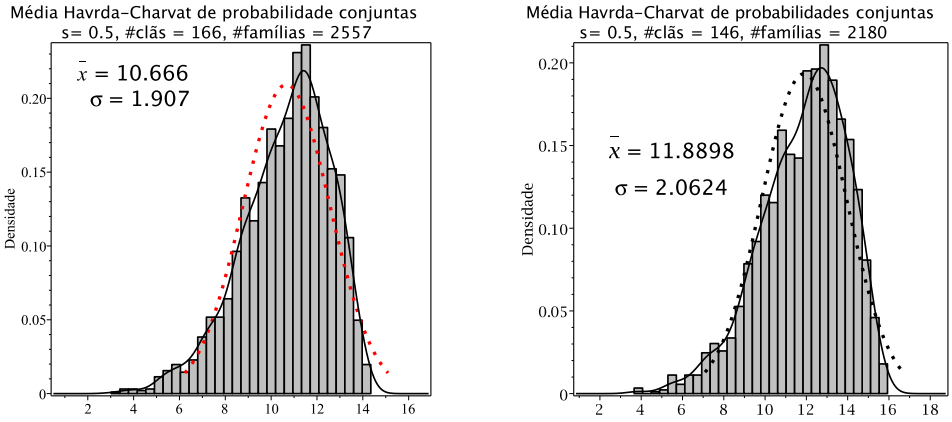

Figura 8.3: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual a 0.5 nas janelas (80×80) e (100×100).

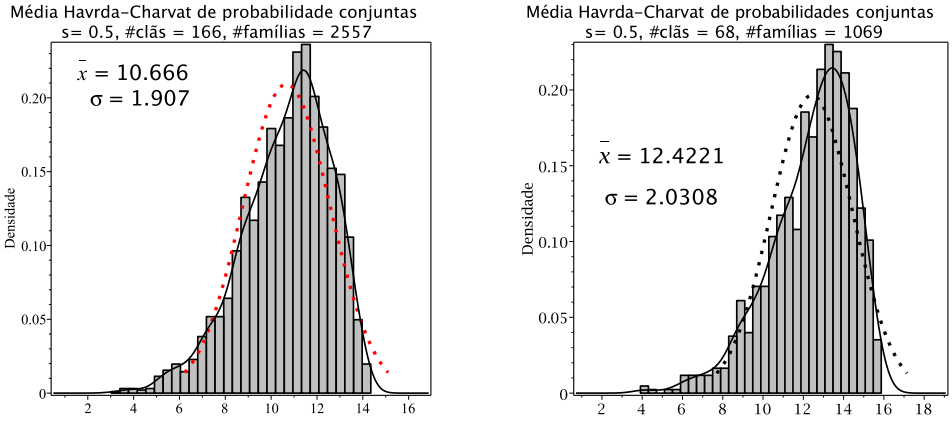

Figura 8.4: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual a 0.5 nas janelas (80×80) e (100×200).

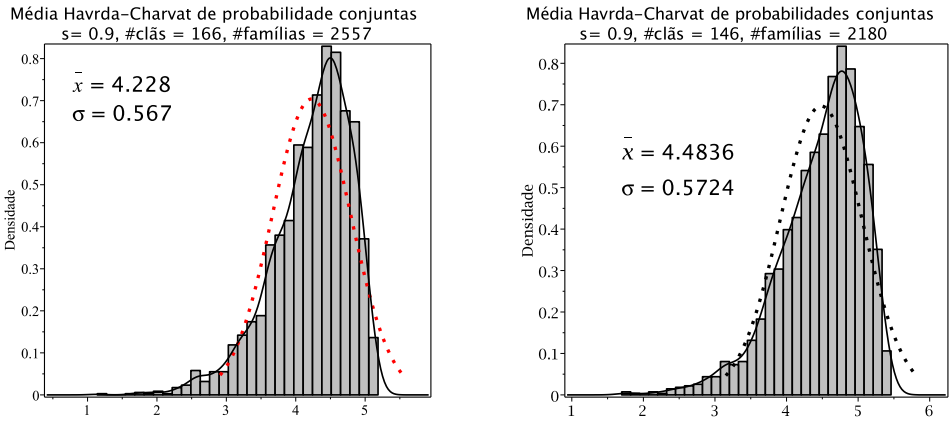

Figura 8.5: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual a 0.9 nas janelas (80×80) e (100×100).

Na Figura 8.5 temos no lado esquerdo o histograma construído a partir dos blocos (80×80) e no lado direito o histograma construído a partir dos blocos (100×100), para o valor de parâmetro s igual a 0.9.

Na Figura 8.6 temos no lado esquerdo o histograma construído a partir dos blocos (80×80) e no lado direito o histograma construído a partir dos blocos (100×200), para o valor de parâmetro s igual a 0.9.

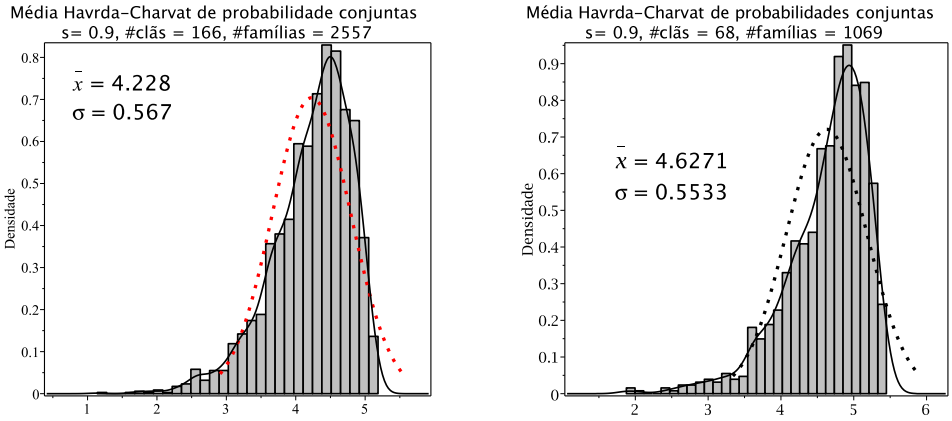

Figura 8.6: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual a 0.9 nas janelas (80×80) e (100×200).

Na Figura 8.7 temos no lado esquerdo o histograma construído a partir dos blocos (80×80) e no lado direito o histograma construído a partir dos blocos (100×100), para o valor de parâmetro s igual a 1.0 (limite Gibbs-Shannon).

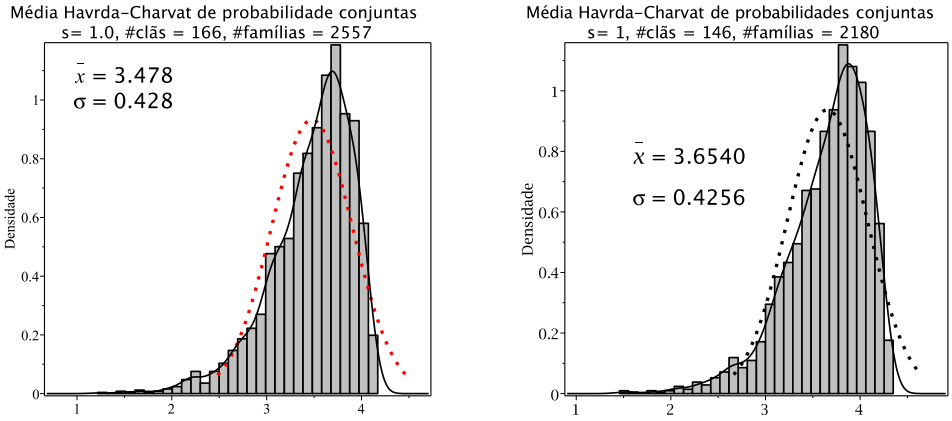

Figura 8.7: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual a 1.0 nas janelas (80×80) e (100×100).

Na Figura 8.8 temos no lado esquerdo o histograma construído a partir dos blocos (80×80) e no lado direito o histograma construído a partir dos blocos (100×200), para o valor de parâmetro s igual a 1.0 (limite Gibbs-Shannon).

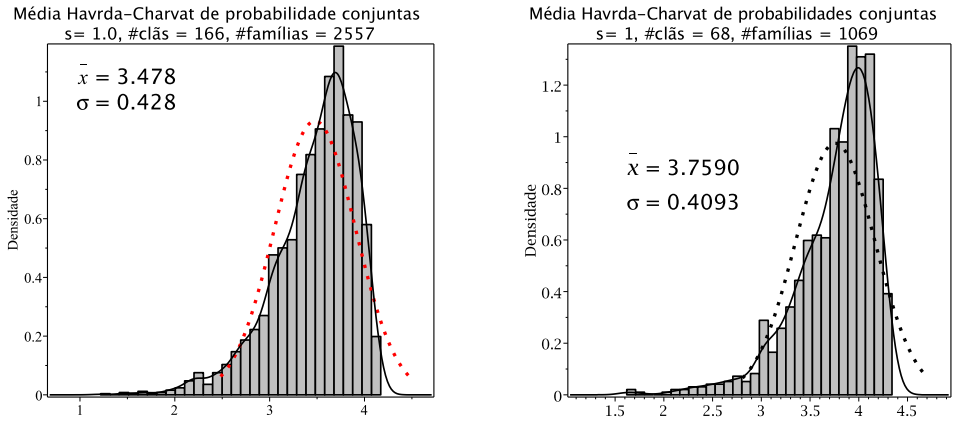

Figura 8.8: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual a 1.0 nas janelas (80×80) e (100×200).

8.2 Medidas de Assimetria e curtose para as distribuições de médias de entropias

As medidas de Assimetria e Curtose, possibilitam, concomitantemente com as medidas de posição e de dispersão, a descrição e compreensão plena das distribuições de frequência. A assimetria é o grau de desvio ou afastamento da simetria de uma distribuição. Em uma distribução simétrica, há igualdade dos valores da média, a mediana e a moda coincidem, num mesmo ponto, de ordenada máxima, havendo um perfeito equilíbrio na distribuição. Quando o equilíbrio não ocorre, isto é, a média($\bar{x}$), a mediana(M_d) e a moda(M_o) recaem em pontos diferentes da distribuição esta será assimétrica; enviesada à direita ou à esquerda[8].

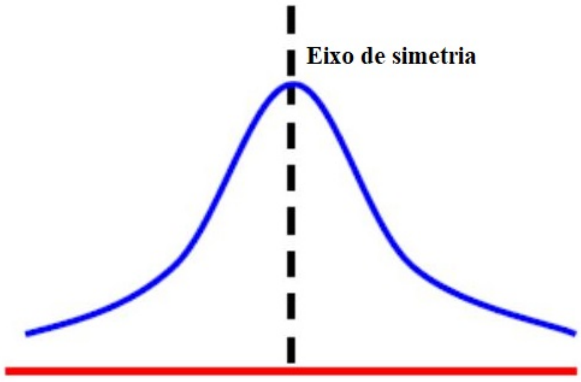

Figura 8.9: Distribução simétrica (média($\bar{x}$)= mediana(M_d)= moda(M_o)).

Em uma distribuição assimétrica positiva, ou assimétrica à direita, quando os valores se concentram na extremidade inferior da escala e se distribuem gradativamente em direção à extremidade superior. Abaixo temos a ilustração uma distribuição assimétrica positiva ou enviesada à direita:

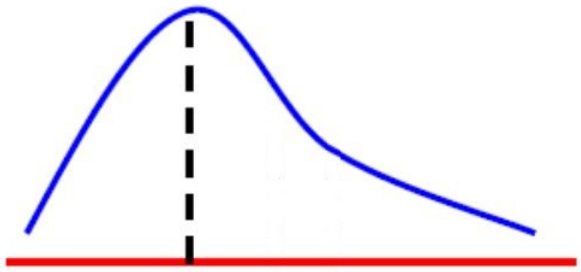

Figura 8.10: Distribução assimétrica positiva (média($\bar{x}$)> mediana(M_d)> moda(M_o)).

Em uma distribuição assimétrica negativa, ou assimétrica à esquerda, quando os valores se concentram na extremidade superior da escala e se distribuem gradativamente em direção à extremidade inferior. Abaixo temos a ilustração uma distribuição assimétrica negativa ou enviesada à esquerda:

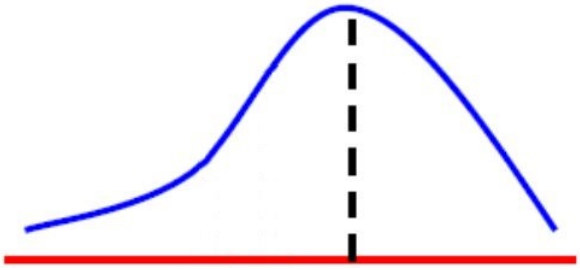

Figura 8.11: Distribução assimétrica negativa (média($\bar{x}$)< mediana(M_d)< moda(M_o)).

Para realizar o coeficiente de assimetria, podemos utilizar:
1^o **Coeficiente de assimetria de Pearson** [77]**:**

$$A_s = \frac{(\bar{x} - M_o)}{s}, \tag{8.1}$$

Em que média($\bar{x}$), moda(M_o) e desvio padrão(s). A interpretação do coeficiente de assimetria é organizada da seguite forma:

Se $A_s = 0$, então a distribução é simétrica.

Se $A_s > 0$, então a distribução é assimétrica positiva.

Se $A_s < 0$, então a distribução é assimétrica negativa.

2^o **Coeficiente de assimetria de Pearson** [77]**:**

$$A_s = \frac{3(\bar{x} - M_d)}{s}, \tag{8.2}$$

em que média($\bar{x}$), mediana(M_d) e desvio padrão(s). A interpretação do coeficiente de assimetria é organizada da seguite forma:

Se $-1 < A_s < 1$, então a distribução é simétrica.

Se $A_s < 1$, então a distribução é assimétrica positiva.

Se $A_s < -1$, então a distribução é assimétrica negativa.

A curtose é o grau de achatamento de uma distribução que mede a concentração ou dispersão dos valores de um conjuto em relação às medidas de tendência central ou normal. A curtose pode ser classificada em três classes:

- Mesocúrtica - Quando apresenta uma curva de frequência idêntica a da distribuição Normal.
- Leptocúrtica - Quando apresenta uma curva de frequência mais fechada (pontiaguda) que a da distribuição Normal.
- Platicúrtica - Quando apresenta uma curva de frequência mais aberta (achatada) que a da distribuição Normal.

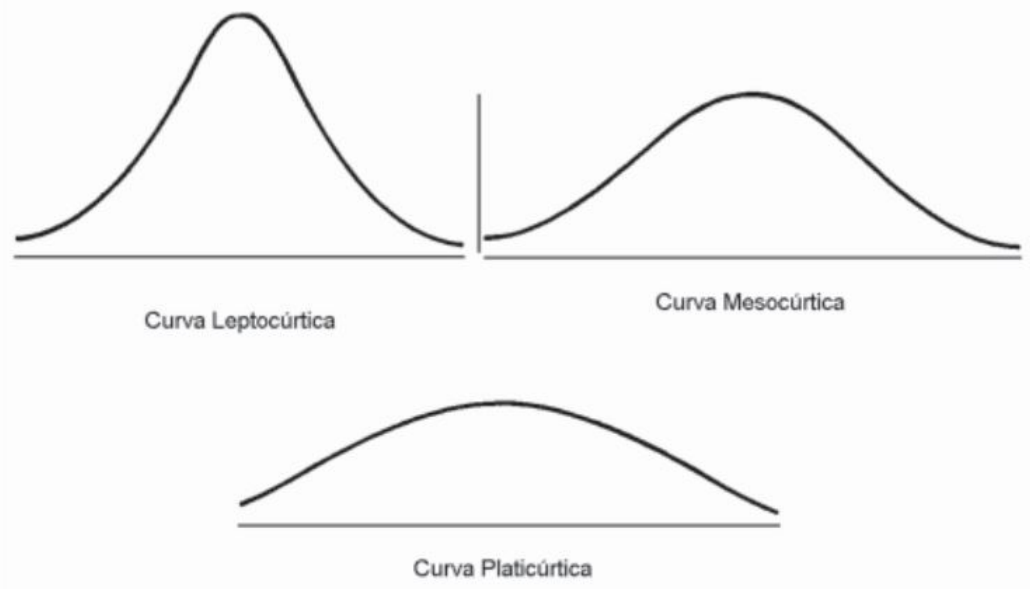

Figura 8.12: Tipos de Curtose [8].

A medida de curtose pode ser calculadas através do coeficiente percentílico de curtose:

$$C = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}, \tag{8.3}$$

Em que:
C = Coeficiente de curtose, $Q_1 = 1^o$ quartil, $Q_3 = 3^o$ quartil, P_{90} = Percentil 90 e P_{90} = Percentil 10.

No que se refere à curva da distribuição Normal, temos:

Se $C < 0.263$, então a curva ou distribução é leptocúrtica (mais afinalada).

Se $C = 0.236$, então a curva ou distribução é mesocúrtica.

Se $C > 0.236$, então a curva ou distribução é platicúrtica (mais achatada).

Foram realizados os cálculos de assimetria e de curtose, assim como o nível de posição central, simetria da distribuição e amplitude de variação para as distribuições de médias de entropia Havrda-Charvat (HC). Os resultados são apresentados na Tabela 8.1.

As Figuras 8.13 a 8.25 abaixo representam os boxplot das médias da entropia de Havrda-Charvat de probabilidade conjunta para as 2557 famílias de 166 clãs para os diferentes valores do parâmetro s (0.1 a 1.3), obtidos nas janelas (80×80). O Boxplot foi utilizado neste trabalho para explorar e comparar as características das distribuições de entropia das famílias em clãs como o nível de posição central, simetria da distribuição e amplitude de variação para as distribuições de médias de entropia Havrda-Charvat (HC).

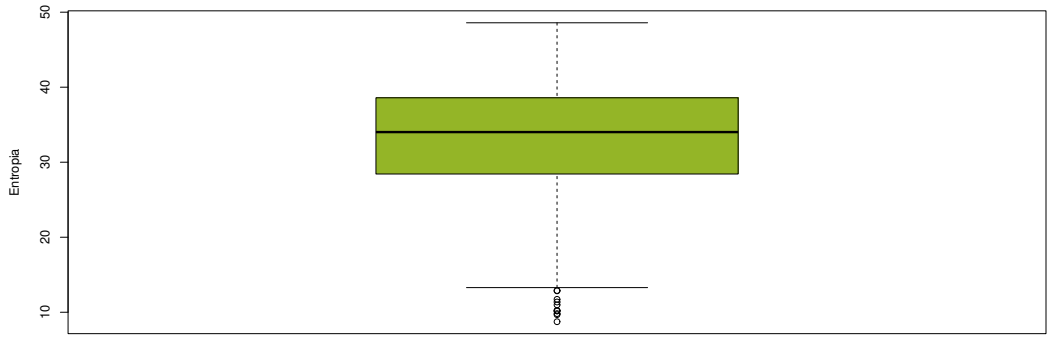

Figura 8.13: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.1, clãs=166, famílias=2557. Fonte: Boxplot construído no software *R*.

Tabela 8.1: Estatística descritiva. Fonte: Geração da tabela no software Bioest 5.0.

	s=0.1	s=0.2	s=0.3	s=0.4	s=0.5	s=0.6	s=0.7	s=0.8	s=0.9	s=1.0	s=1.1	s=1.2	s=1.3
Tamanho da amostra	2557	2557	2557	2557	2557	2557	2557	2557	2557	2557	2557	2557	2557
Mínimo	8.0000	6.0000	4.0000	3.0000	3.0000	2.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
Máximo	48.0000	35.0000	25.0000	19.0000	14.0000	10.0000	8.0000	6.0000	5.0000	4.0000	3.0000	2.0000	2.0000
Amplitude Total	40.0000	29.0000	21.0000	16.0000	11.0000	8.0000	7.0000	5.0000	4.0000	3.0000	2.0000	1.0000	2.0000
Mediana	34.0000	25.0000	18.0000	14.0000	10.0000	8.0000	6.0000	5.0000	4.0000	3.0000	2.0000	2.0000	2.0000
Primeiro Quartil(25%)	28.0000	21.0000	16.0000	12.0000	9.0000	7.0000	5.0000	4.0000	3.0000	3.0000	2.0000	2.0000	2.0000
Terceiro Quartil (75%)	38.0000	28.0000	21.0000	15.0000	12.0000	9.0000	7.0000	5.0000	4.0000	3.0000	3.0000	2.0000	2.0000
Desvio Interquartílico	10.0000	7.0000	5.0000	3.0000	3.0000	2.0000	2.0000	1.0000	1.0000	0.0000	1.0000	0.0000	0.0000
Média Aritmética	32.7704	24.0982	17.8795	13.4145	10.1670	7.7763	6.0192	4.7255	3.7083	2.9257	2.4400	1.9413	1.7603
Desvio Padrão	7.2235	5.1493	3.6840	2.6462	1.9242	1.4197	1.0571	0.8244	0.5948	0.4466	0.5359	0.2350	0.4288
Erro Padrão	0.1428	0.1018	0.0729	0.0523	0.0381	0.0281	0.0209	0.0163	0.0118	0.0088	0.0106	0.0046	0.0085
Assimetria	−0.4415	−0.4875	−0.5366	−0.5916	−0.6400	−0.6627	−0.7622	−0.6318	−1.0362	−0.7000	−0.1568	−3.7584	−1.2497
Curtose	−0.1462	−0.0657	0.0102	0.1248	0.2462	0.3303	0.6684	0.7085	1.6691	3.1821	−1.1533	12.1350	−0.3470

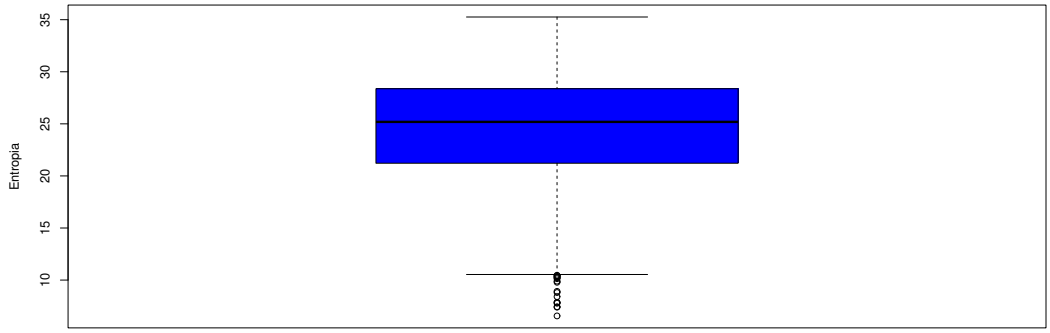

Figura 8.14: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.2, clãs=166, famílias=2557. Fonte: Boxplot construído no software R.

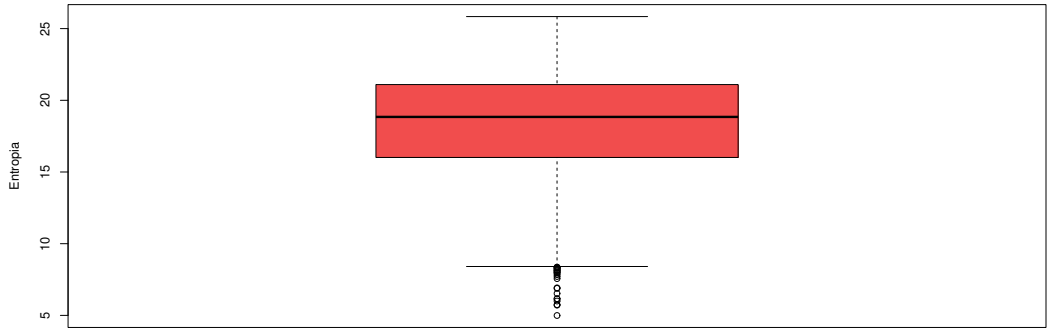

Figura 8.15: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.3, clãs=166, famílias=2557. Fonte: Boxplot construído no software R.

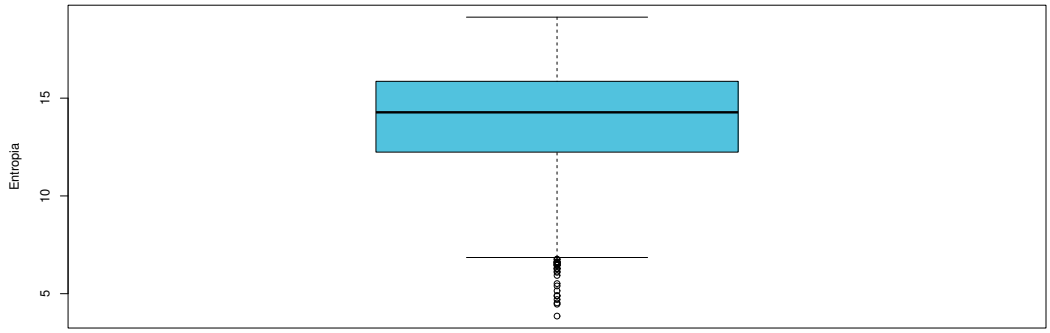

Figura 8.16: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.4, clãs=166, famílias=2557. Fonte: Boxplot construído no software R.

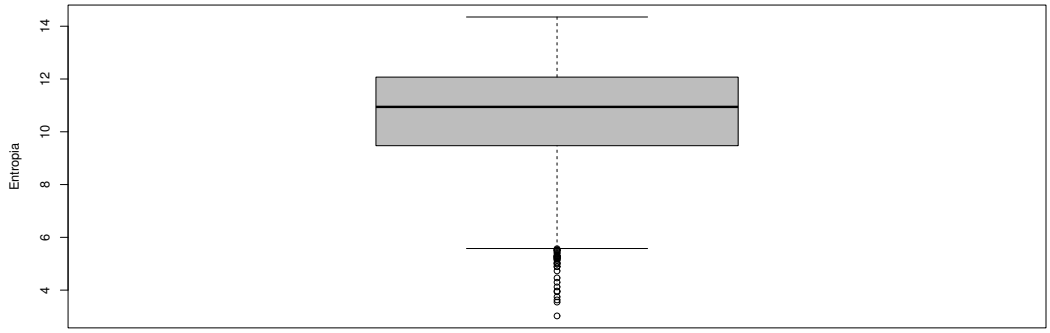

Figura 8.17: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.5, clãs=166, famílias=2557. Fonte: Boxplot construído no software R.

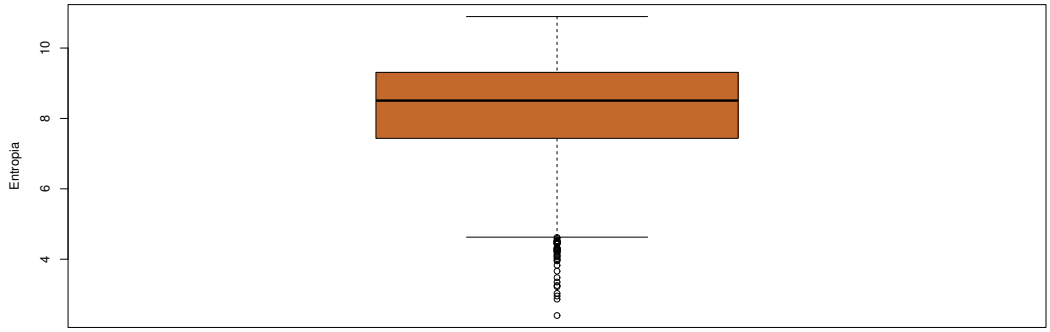

Figura 8.18: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.6, clãs=166, famílias=2557. Fonte: Boxplot construído no software R.

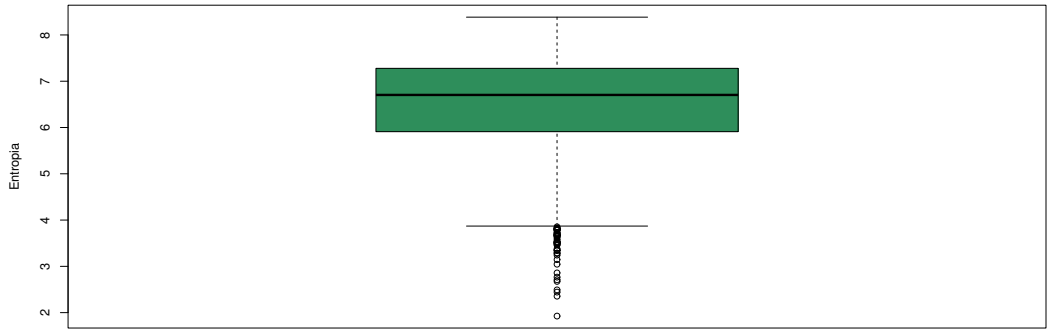

Figura 8.19: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.7, clãs=166, famílias=2557. Fonte: Boxplot construído no software R.

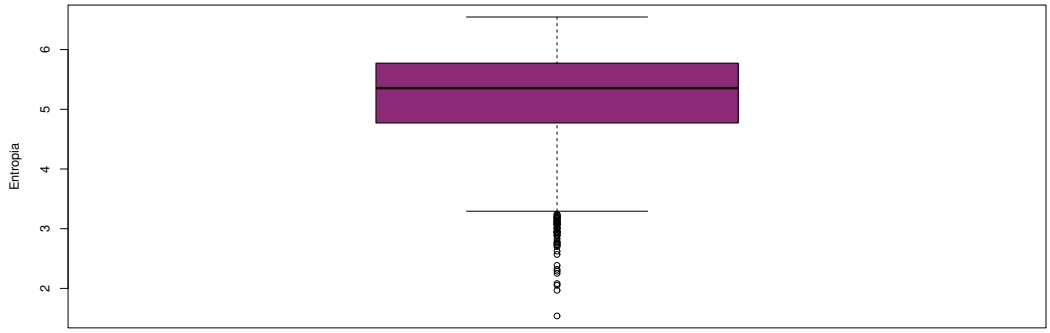

Figura 8.20: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.8, clãs=166, famílias=2557. Fonte: Boxplot construído no software R.

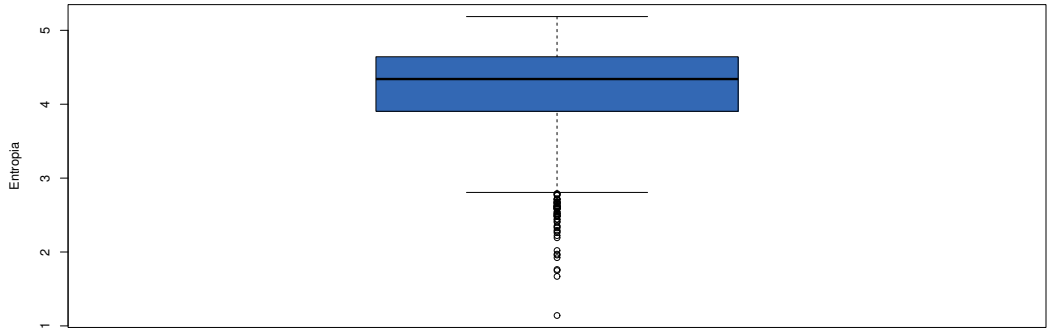

Figura 8.21: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.9, clãs=166, famílias=2557. Fonte: Boxplot construído no software R.

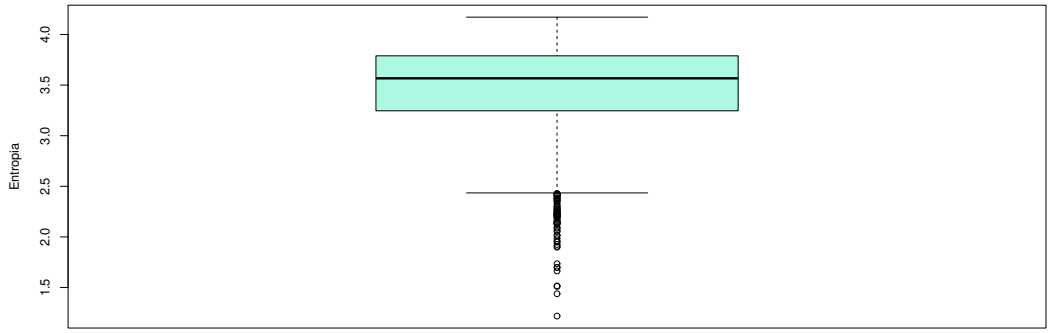

Figura 8.22: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=1.0, clãs=166, famílias=2557. Fonte: Boxplot construído no software R.

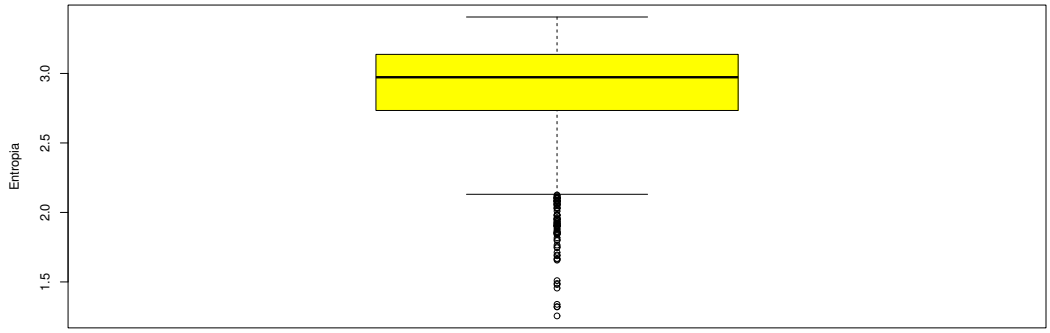

Figura 8.23: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=1.1, clãs=166, famílias=2557. Fonte: Boxplot construído no software R.

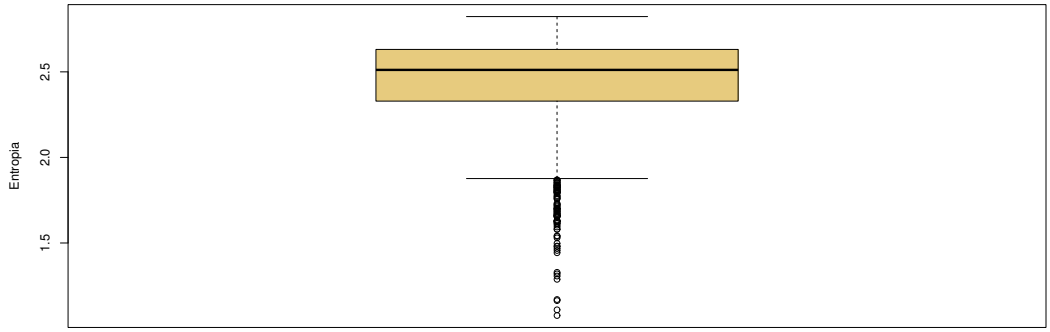

Figura 8.24: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=1.2, clãs=166, famílias=2557. Fonte: Boxplot construído no software R.

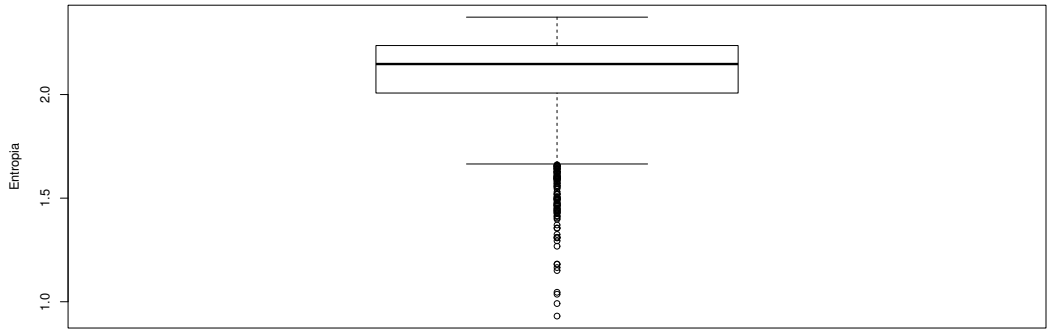

Figura 8.25: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=1.3, clãs=166, famílias=2557. Fonte: Boxplot construído no software R.

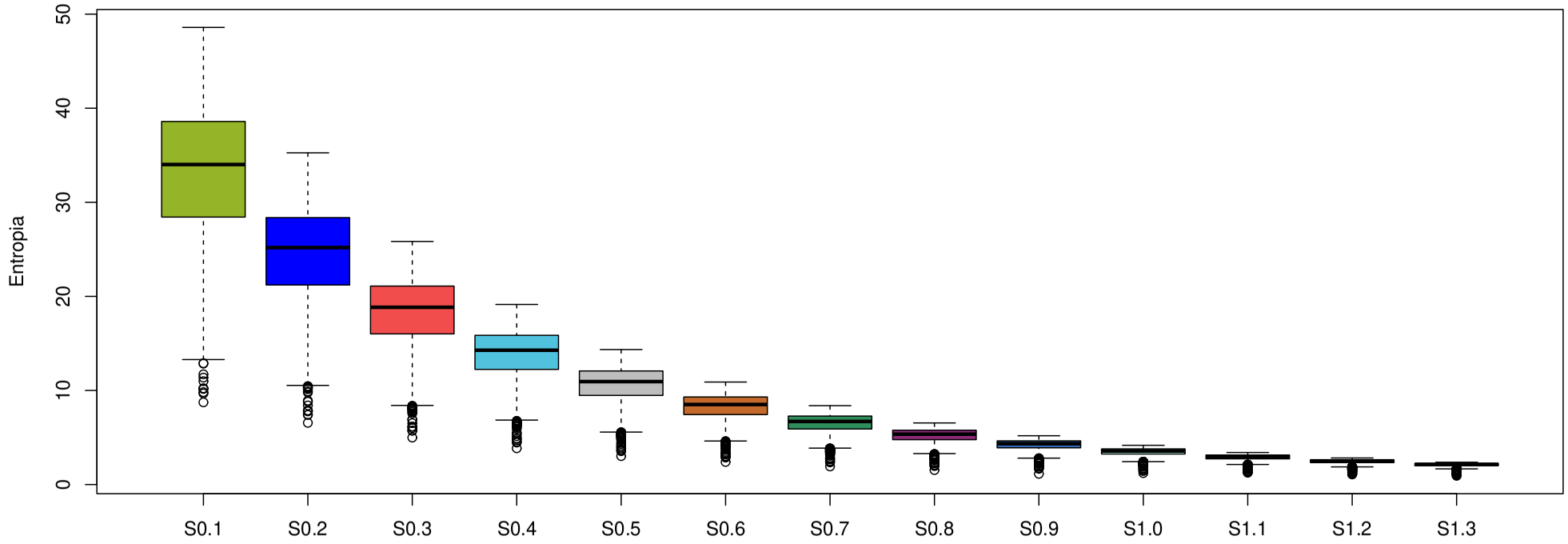

Figura 8.26: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas $s = 0.1$ ... $s = 1.3$, clãs=166, famílias=2557. Fonte: Boxplot construído no software R.

A Figura 8.26 contém o boxplot com todos os valores do parâmetro s de 0.1 a 1.3, com os valores de mínimo e máximo das médias de entropia Havrda-Charvat de probabilidades conjuntas para as 2557 famílias pertencentes a 166 clãs.

8.3 Análise dos histogramas dos clãs com 30 ou mais famílias

Foram gerados os histogramas de densidade das médias de entropia Havrda-Charvat (HC) com 18 clãs (CL0020, CL0023, CL0028, CL0029, CL0036, CL0058, CL0063, CL0110, CL0113, CL0123, CL0125, CL0126, CL0159, CL0172, CL0186, CL0193, CL0219, CL0236) que contenham 30 ou mais famílias, selecionando para o parâmetro s os valores: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2 e 1.3.

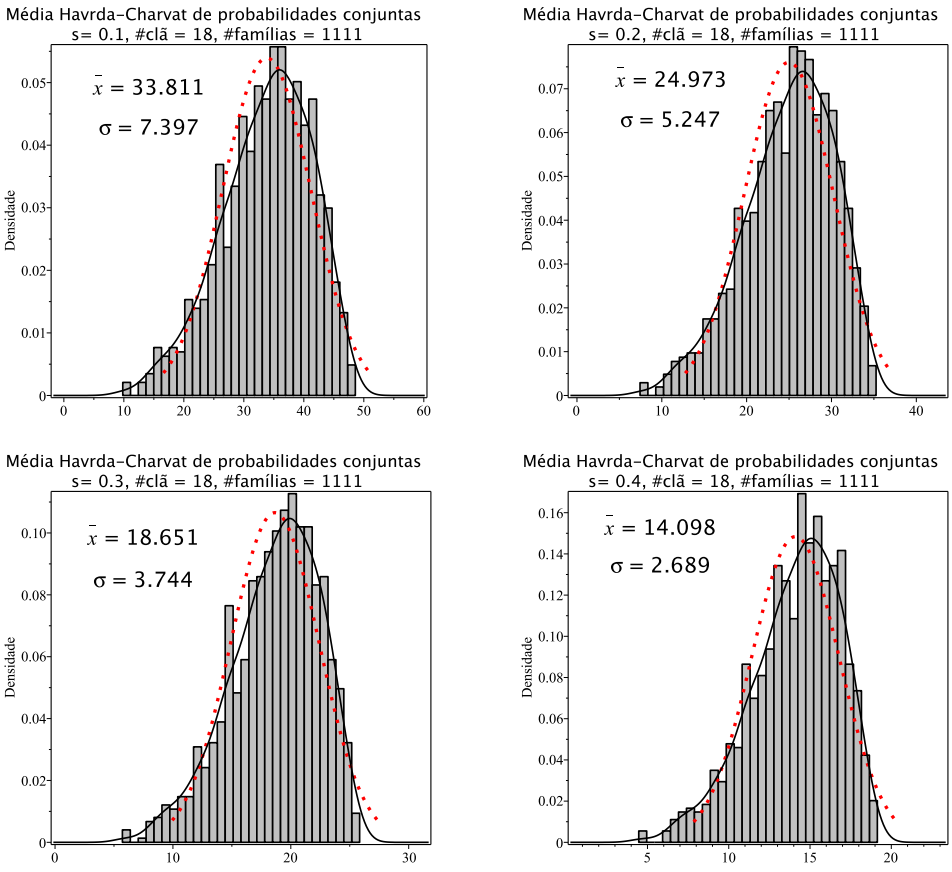

Figura 8.27: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

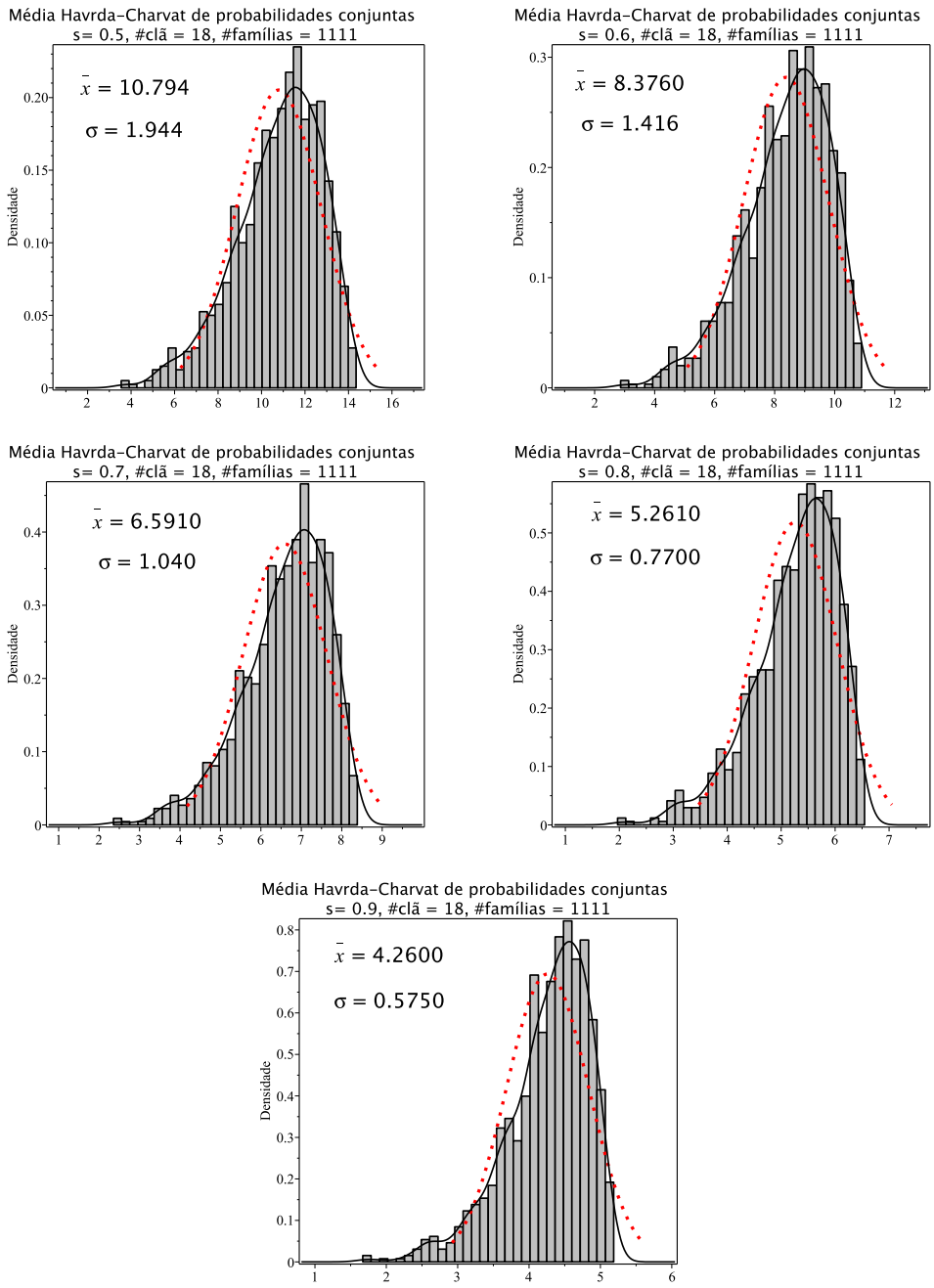

Figura 8.28: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

A Tabela 8.2 apresenta os resultados dos cálculos de assimetria e de curtose, assim como o nível de posição central, simetria da distribuição e amplitude de variação para as distribuições de médias de entropia Havrda-Charvat (HC) com clãs de 30 ou mais famílias.

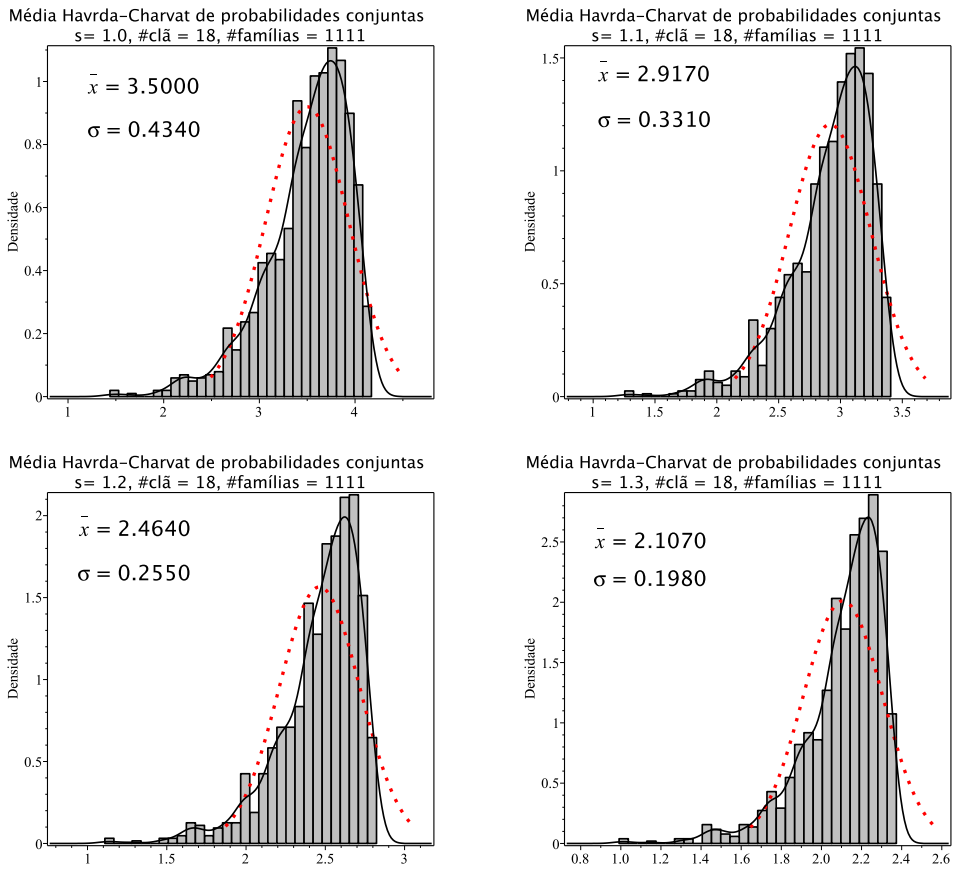

Figura 8.29: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

As Figuras 8.30 a 8.42 abaixo representam os boxplot das médias da entropia de Havrda-Charvat de probabilidade conjunta para as 1111 famílias de 18 clãs para os diferentes valores do parâmetro s (0.1 a 1.3), obtidos nas janelas (80×80). O Boxplot foi utilizado neste trabalho para explorar e comparar as características das distribuições de entropia das famílias em clãs como o nível de posição central, simetria da distribuição e amplitude de variação para as distribuições de médias de entropia Havrda-Charvat (HC) com clãs de 30 ou mais famílias.

Tabela 8.2: Estatística descritiva dos clãs com 30 ou mais famílias. Fonte: Geração da tabela no software Bioest 5.0.

	s=0.1	s=0.2	s=0.3	s=0.4	s=0.5	s=0.6	s=0.7	s=0.8	s=0.9	s=1.0	s=1.1	s=1.2	s=1.3
Tamanho da amostra	1111	1111	1111	1111	1111	1111	1111	1111	1111	1111	1111	1111	1111
Mínimo	9.0000	7.0000	5.0000	4.0000	3.0000	2.0000	2.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
Máximo	48.0000	35.0000	25.0000	19.0000	14.0000	10.0000	8.0000	6.0000	5.0000	4.0000	3.0000	2.0000	2.0000
Amplitude Total	39.0000	28.0000	20.0000	15.0000	11.0000	8.0000	6.0000	5.0000	4.0000	3.0000	2.0000	1.0000	2.0000
Mediana	34.0000	25.0000	19.0000	14.0000	11.0000	8.0000	6.0000	5.0000	4.0000	3.0000	2.0000	2.0000	2.0000
Primeiro Quartil(25%)	29.0000	21.0000	16.0000	11.5000	9.0000	7.0000	6.0000	4.0000	3.0000	3.0000	2.0000	2.0000	2.0000
Terceiro Quartil (75%)	39.0000	28.0000	21.0000	15.0000	12.0000	9.0000	7.0000	5.0000	4.0000	3.0000	3.0000	2.0000	2.0000
Desvio Interquartílico	10.0000	7.0000	21.0000	5.0000	3.5000	3.0000	2.0000	1.0000	1.0000	0.0000	1.0000	0.0000	0.0000
Média Aritmética	33.3087	24.4707	18.1413	13.2520	10.2862	7.8614	6.0891	4.7723	3.7435	2.9433	2.4626	1.9388	1.7759
Desvio Padrão	7.3890	5.2673	3.7533	2.6942	1.9620	1.4540	1.0748	0.8336	0.6032	0.6032	0.5421	0.2398	0.4193
Erro Padrão	0.2217	0.1580	0.1126	0.0808	0.0589	0.0322	0.0250	0.0181	0.0138	0.0163	0.0106	0.0072	0.0126
Assimetria	−0.4930	−0.5379	−0.5130	−0.5916	−0.7011	−0.7074	−0.7797	−0.6651	−1.0009	−0.4874	−0.2763	−3.6660	−1.3616
Curtose	−0.1709	−0.0777	−0.0080	0.0283	0.2843	0.2955	0.6279	0.7151	1.8557	2.6797	−1.0739	11.4603	−0.0311

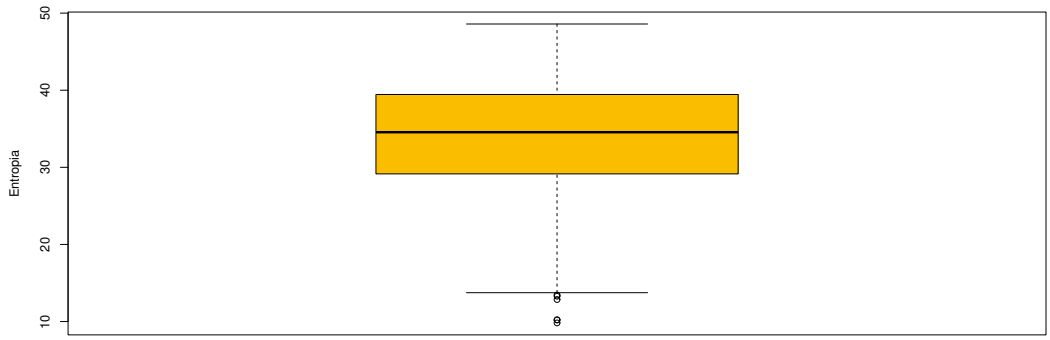

Figura 8.30: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.1, clãs=18, famílias=1111. Fonte: Boxplot construído no software R.

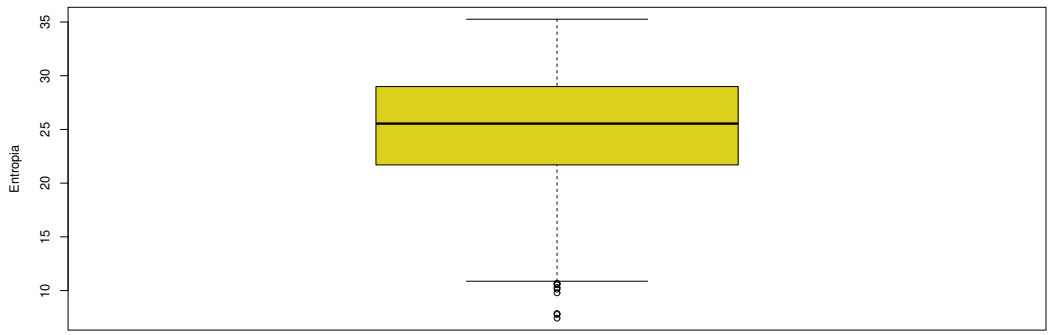

Figura 8.31: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.2, clãs=18, famílias=1111. Fonte: Boxplot construído no software R.

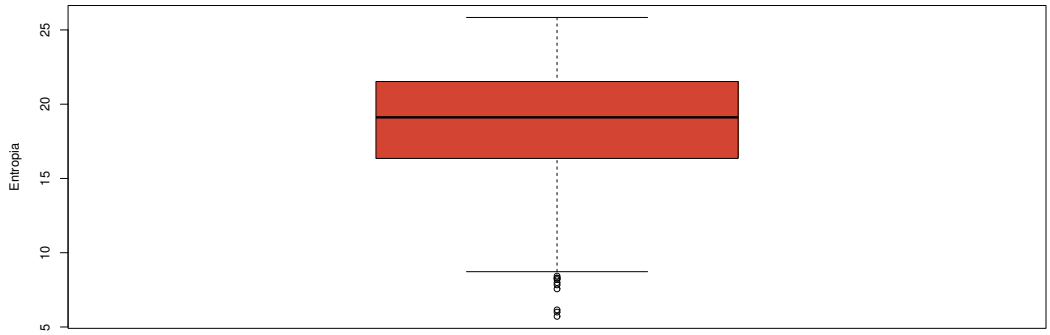

Figura 8.32: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.3, clãs=18, famílias=1111. Fonte: Boxplot construído no software R.

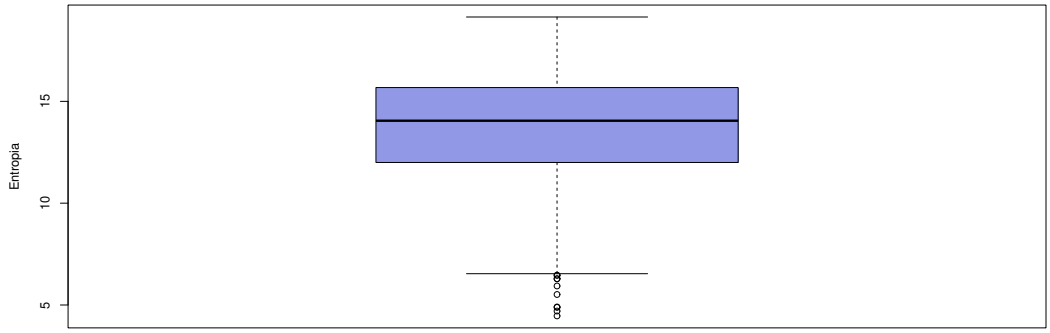

Figura 8.33: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.4, clãs=18, famílias=1111. Fonte: Boxplot construído no software R.

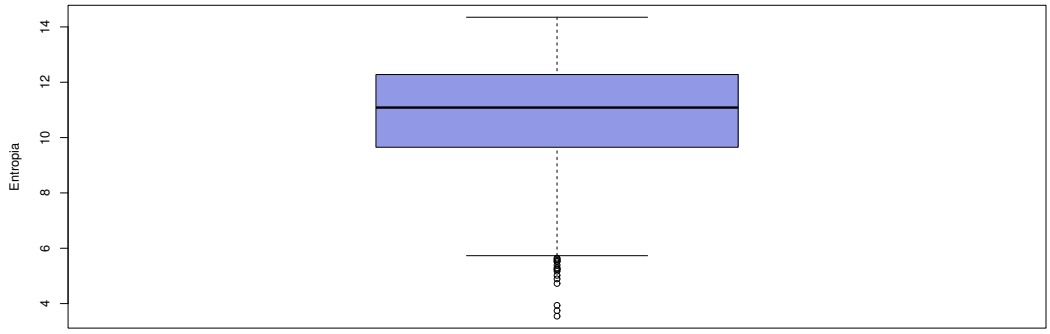

Figura 8.34: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.5, clãs=18, famílias=1111. Fonte: Boxplot construído no software R.

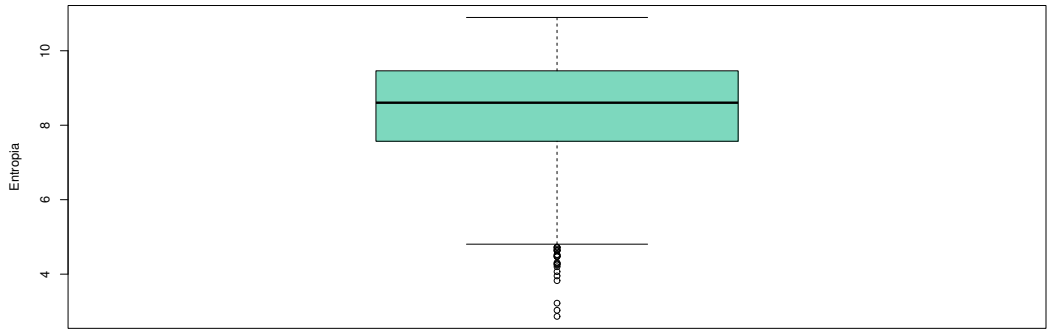

Figura 8.35: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.6, clãs=18, famílias=1111. Fonte: Boxplot construído no software R.

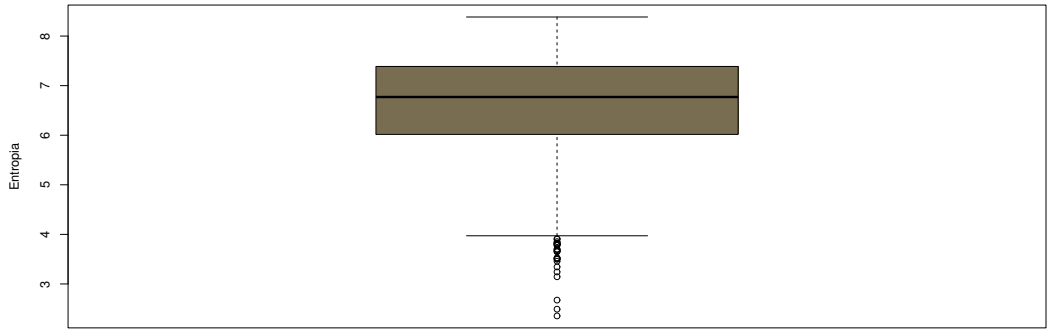

Figura 8.36: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.7, clãs=18, famílias=1111. Fonte: Boxplot construído no software R.

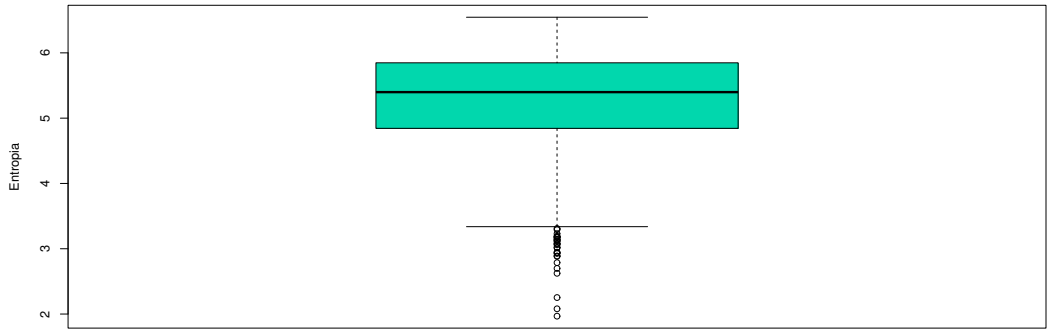

Figura 8.37: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.8, clãs=18, famílias=1111. Fonte: Boxplot construído no software R.

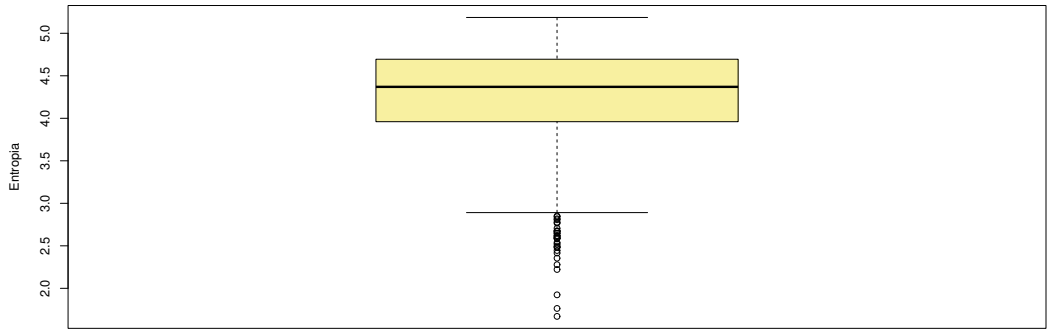

Figura 8.38: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.9, clãs=18, famílias=1111. Fonte: Boxplot construído no software R.

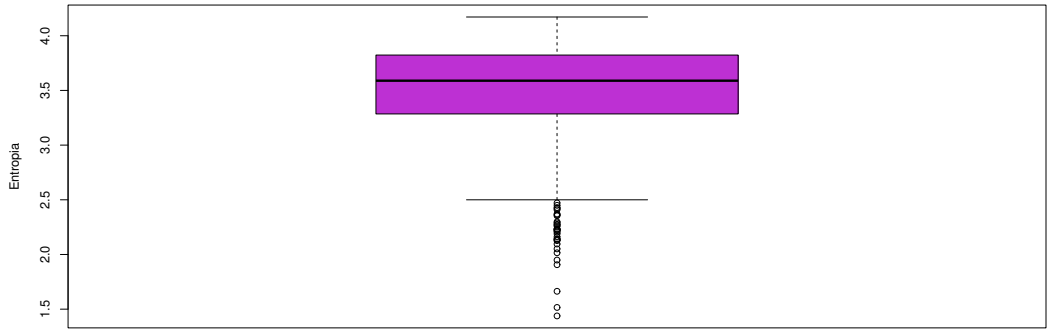

Figura 8.39: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=1.0, clãs=18, famílias=1111. Fonte: Boxplot construído no software R.

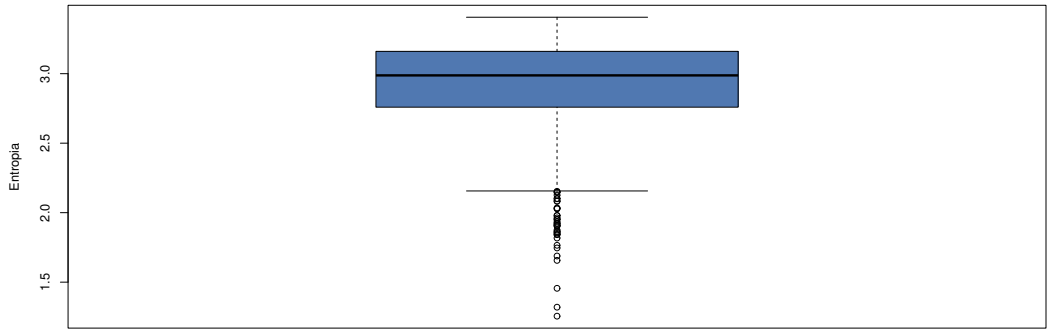

Figura 8.40: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=1.1, clãs=18, famílias=1111. Fonte: Boxplot construído no software R.

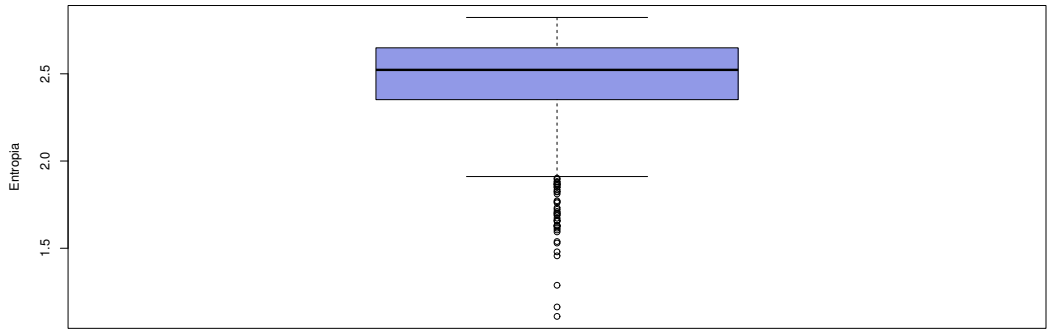

Figura 8.41: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=1.2, clãs=18, famílias=1111. Fonte: Boxplot construído no software R.

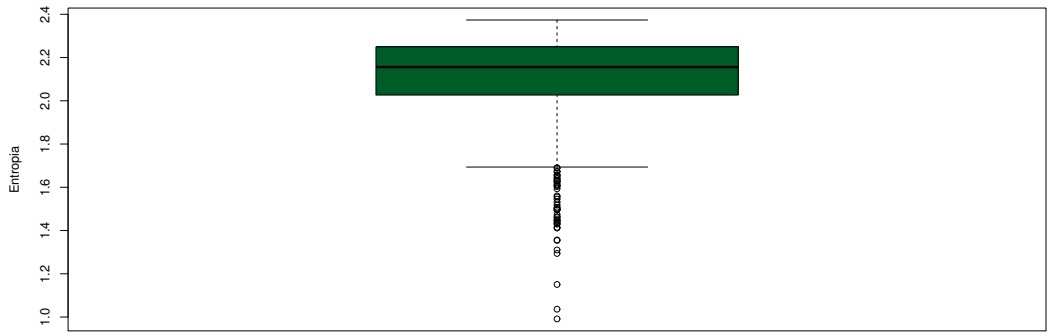

Figura 8.42: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=1.3, clãs=18, famílias=1111. Fonte: Boxplot construído no software R.

Em relação à simetria e o grau de achatamento das curvas (curtose), pode ser visto que os parâmetros 0.1, 0.2, 0.3, 0.4 e 1.1 possuem em comum uma curva simétrica (mais próxima da curva gaussiana) e leptocúrtica (curva mais afinalada). Os parâmetros 0.5, 0.6, 0.7, 0.8 e 1.0, têm em comum uma curva simétrica e platicúrtica (mais achatada). Os parâmetros 0.9 e 1.2 possuem uma curva considerada assimétrica negativa e platicúrtica. Finalmente, temos o parâmetro 1.3 que se destaca como o único a possuir uma curva assimétrica negativa e leptocúrtica. Essas conclusões foram feitas fundamentadas nos cálculos de assimetria e curtose.

Observando valores das entropias um a um de cada parâmetro, podemos notar especificidades. Considere que todos os valores estão aproximados (valores disponíveis na Tabela 8.2).

Podemos observar para o parâmetro $s = 0.1$, que entre os valores de médias de entropia que vão aproximadamente de 9 a 48, os valores de 9.82 a 13.43 são mais discrepantes em relação aos dados e os valores de 34 a 36 aparecem em maioria. Além disso, os cálculos da assimetria e curtose indicam uma curva mais simétrica (possui proximidade com a curva gaussiana) e leptocúrtica (possui sua curva mais afinalada).

Para o parâmetro $s = 0.2$, os valores de médias de entropia vão aproximadamente de 7 a 35, sendo os valores de 7.43 a 10.87 mais discrepantes em relação aos dados e os valores entre 25 a 27 aparecem em maioria. Além disso, os cálculos da assimetria e curtose indicam uma curva mais simétrica (possui proximidade com a curva gaussiana) e leptocúrtica (possui sua curva mais afinalada).

Para o parâmetro $s = 0.3$, os valores de médias de entropia vão aproximadamente de 5 a 25, sendo os valores de 5.71 a 8.5 mais discrepantes em relação aos dados e os valores entre 18 a 21 aparecem em maioria. Além disso, os cálculos da assimetria e curtose indicam uma curva mais simétrica (possui proximidade com a curva gaussiana) e leptocúrtica (possui sua curva mais afinalada).

Para o parâmetro $s = 0.4$, os valores de médias de entropia vão aproximadamente de 4 a 19, sendo os valores de 4.46 a 6.48 mais discrepantes em relação aos dados e os valores entre 14 a 16 aparecem em maioria. Além disso, os cálculos da assimetria e curtose indicam uma curva mais simétrica (possui proximidade com a curva gaussiana) e leptocúrtica (possui sua curva mais afinalada).

Para o parâmetro $s = 0.5$, os valores de médias de entropia vão aproximadamente de 3 a 14, sendo os valores de 3.54 a 5.99 mais discrepantes em relação aos dados e os valores entre 10 a 12 aparecem em maioria. Além disso, os cálculos da assimetria e curtose indicam uma curva mais simétrica (possui proximidade com a curva gaussiana) e platicúrtica (possui sua curva mais achatada).

Para o parâmetro $s = 0.6$, os valores de médias de entropia vão aproximadamente de 2 a 10, sendo os valores de 2.86 a 4.73 mais discrepantes em relação aos dados e os valores entre 7 a 9 aparecem em maioria. Além disso, os cálculos da assimetria e curtose indicam uma curva mais simétrica (possui proximidade com a curva gaussiana) e platicúrtica (possui sua curva mais achatada).

Para o parâmetro $s = 0.7$, os valores de médias de entropia vão aproximadamente de 2 a 8, sendo os valores de 2.35 a 3.95 mais discrepantes em relação aos dados e os valores entre 6 a 7 aparecem em maioria. Além disso, os cálculos da assimetria e curtose indicam uma curva mais simétrica (possui proximidade com a curva gaussiana) e platicúrtica (possui sua curva mais achatada).

Para o parâmetro $s = 0.8$, os valores de médias de entropia vão aproximadamente de 1 a 6, sendo os valores de 1.96 a 2.94 mais discrepantes em relação aos dados e os valores entre 3 a 4 aparecem em maioria. Além disso, os cálculos da assimetria e curtose indicam uma curva mais simétrica (possui proximidade com a curva gaussiana) e platicúrtica (possui sua curva mais achatada).

Para o parâmetro $s = 0.9$, os valores de médias de entropia vão aproximadamente de 1 a 5, sendo os valores de 1.66 a 2.85 mais discrepantes em relação aos dados e os valores entre 3 a 4 aparecem em maioria. Além disso, os cálculos da assimetria

e curtose indicam uma curva mais assimétrica negativa e platicúrtica (possui sua curva mais achatada).

Para o parâmetro $s = 1.0$, os valores de médias de entropia vão aproximadamente de 1 a 4, sendo os valores de 1.48 a 2.47 mais discrepantes em relação aos dados e os valores de 3 a 4 aparecem em maioria. Além disso, os cálculos da assimetria e curtose indicam uma curva mais simétrica (possui proximidade com a curva gaussiana) e platicúrtica (possui sua curva mais achatada).

Para o parâmetro $s = 1.1$, os valores de médias de entropia vão aproximadamente de 1 a 3, sendo os valores de 1.25 a 2.15 mais discrepantes em relação aos dados e os valores de 2.16 a 3 aparecem em maioria. Além disso, os cálculos da assimetria e curtose indicam uma curva mais simétrica (possui proximidade com a curva gaussiana) e leptocúrtica (possui sua curva mais afinalada).

Para o parâmetro $s = 1.2$, os valores de médias de entropia vão aproximadamente de 1 a 3,sendo os valores de 1.10 a 1.90 mais discrepantes em relação aos dados e os valores de 1.91 a 3 aparecem em maioria. Além disso, os cálculos da assimetria e curtose indicam uma curva mais assimétrica negativa e platicúrtica (possui sua curva mais achatada).

Para o parâmetro $s = 1.3$, os valores de médias de entropia vão aproximadamente de 0.9 a 2,sendo os valores de 0.99 a $1,72$ mais discrepantes em relação aos dados e os valores de 1.73 a 2 aparecem em maioria. Além disso, os cálculos da assimetria e curtose indicam uma curva mais assimétrica negativa e leptocúrtica (possui sua curva mais afinalada).

A Figura 8.43 abaixo contém o boxplot com todos os valores do parâmetro s de 0.1 a 1.3, com os valores de mínimo e máximo das médias de entropia Havrda-Charvat de probabilidades conjuntas para as 1111 famílias pertencentes a 18 clãs.

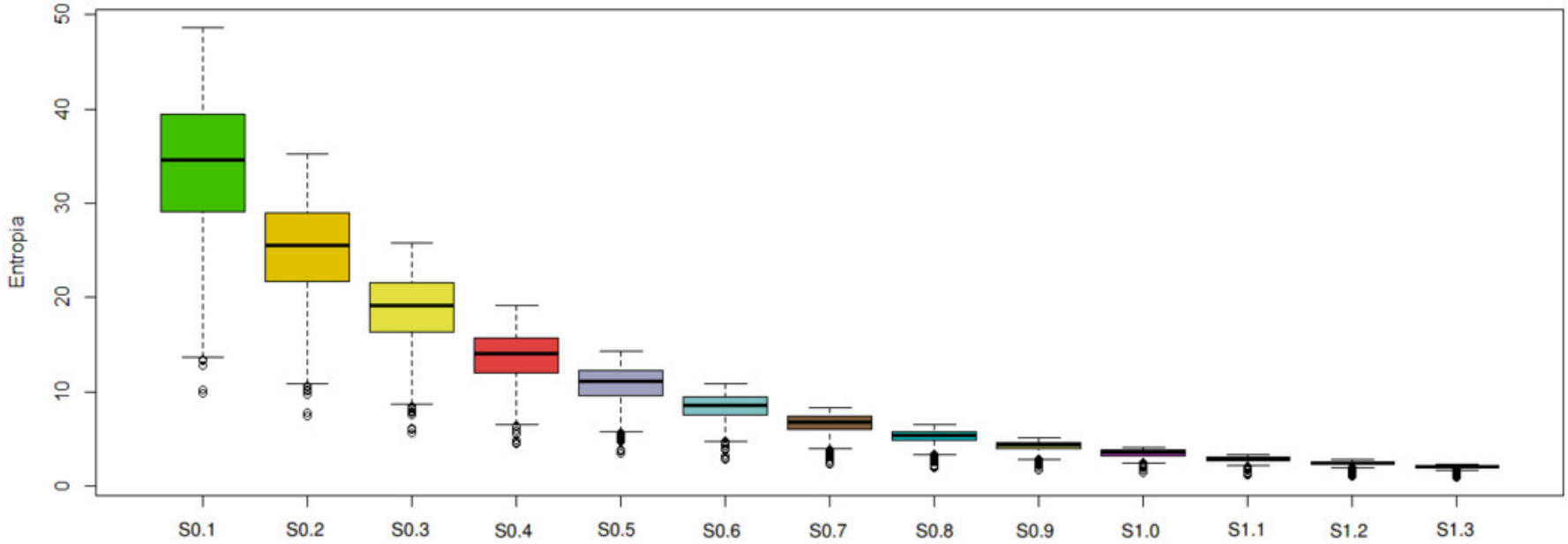

Figura 8.43: Box-Plot Entropia de Havrda-Charvat de probabilidades conjuntas s=0.1 ... s=1.3, clãs=18, famílias=1111. Fonte: Boxplot construído no software *R*.

Abaixo seguem os histogramas de densidade individualizados clã a clã, selecionando os valores do parâmetro s igual a 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2 e 1.3.

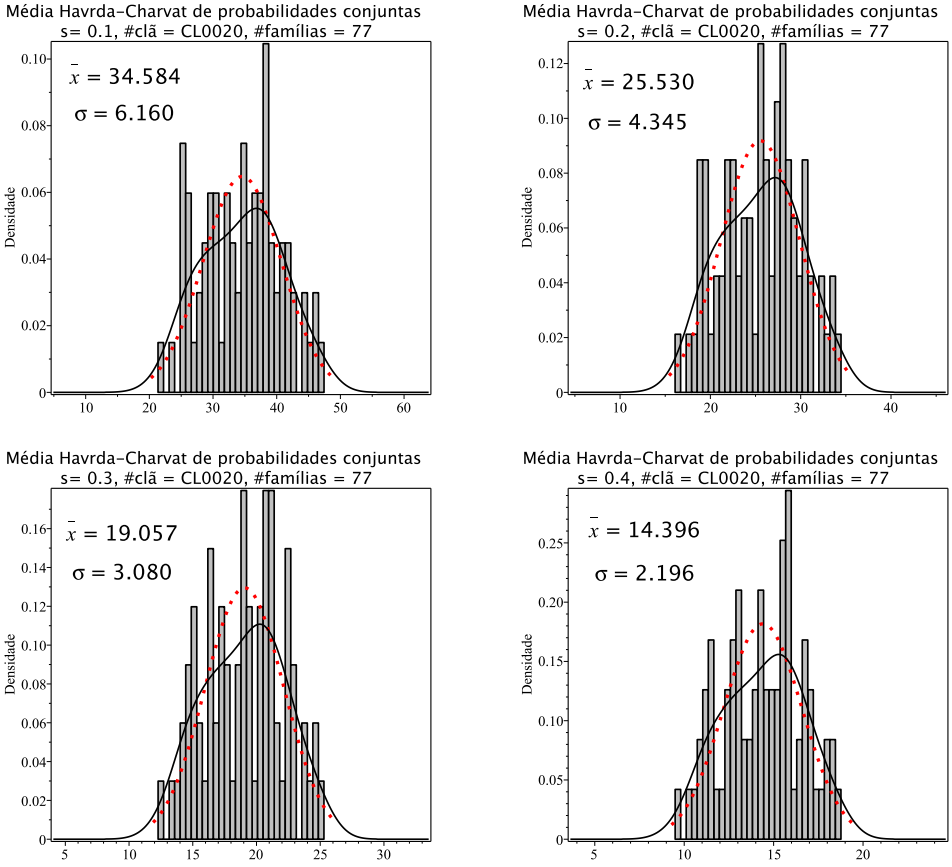

Figura 8.44: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0020. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

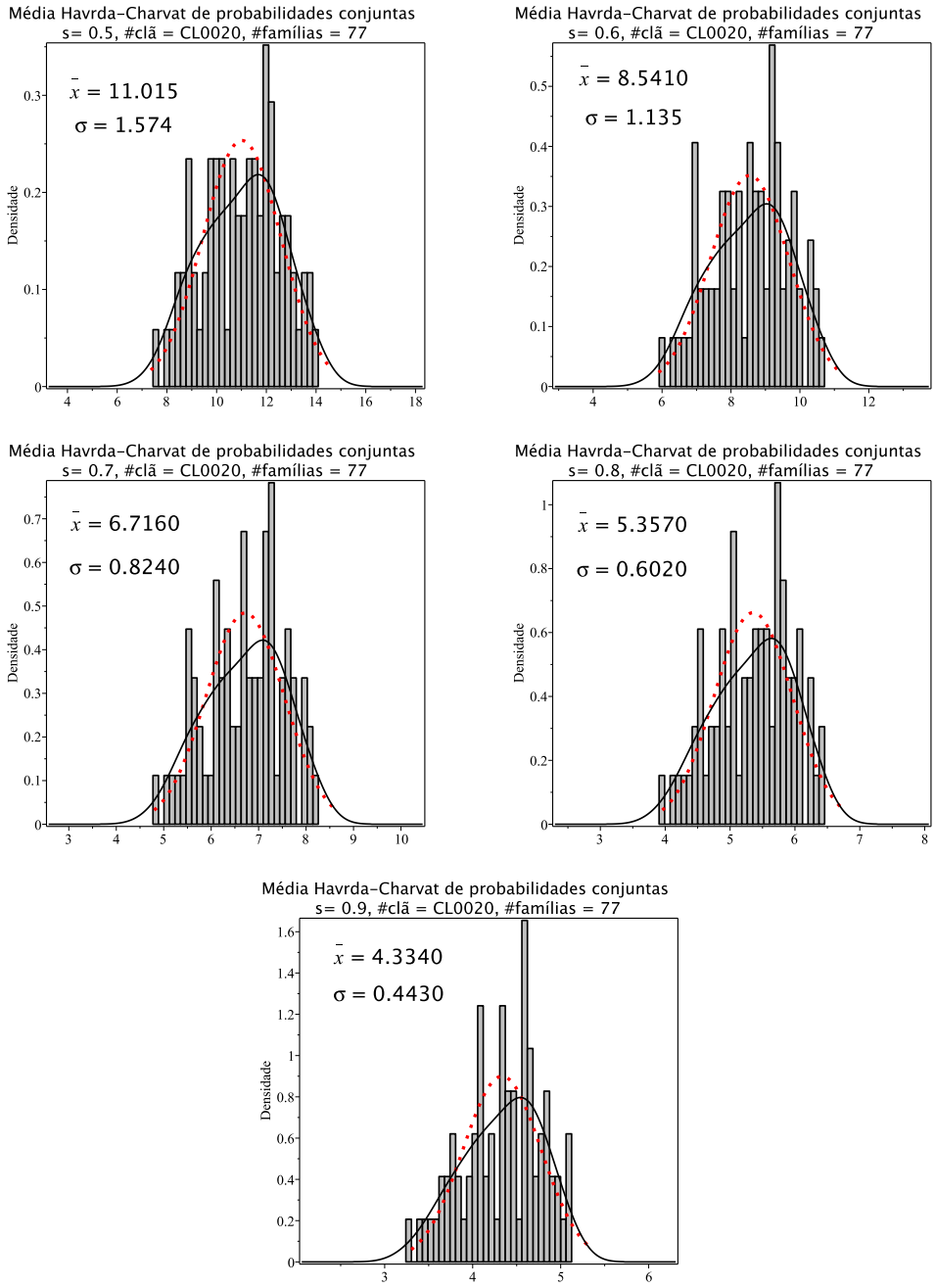

Figura 8.45: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0020. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

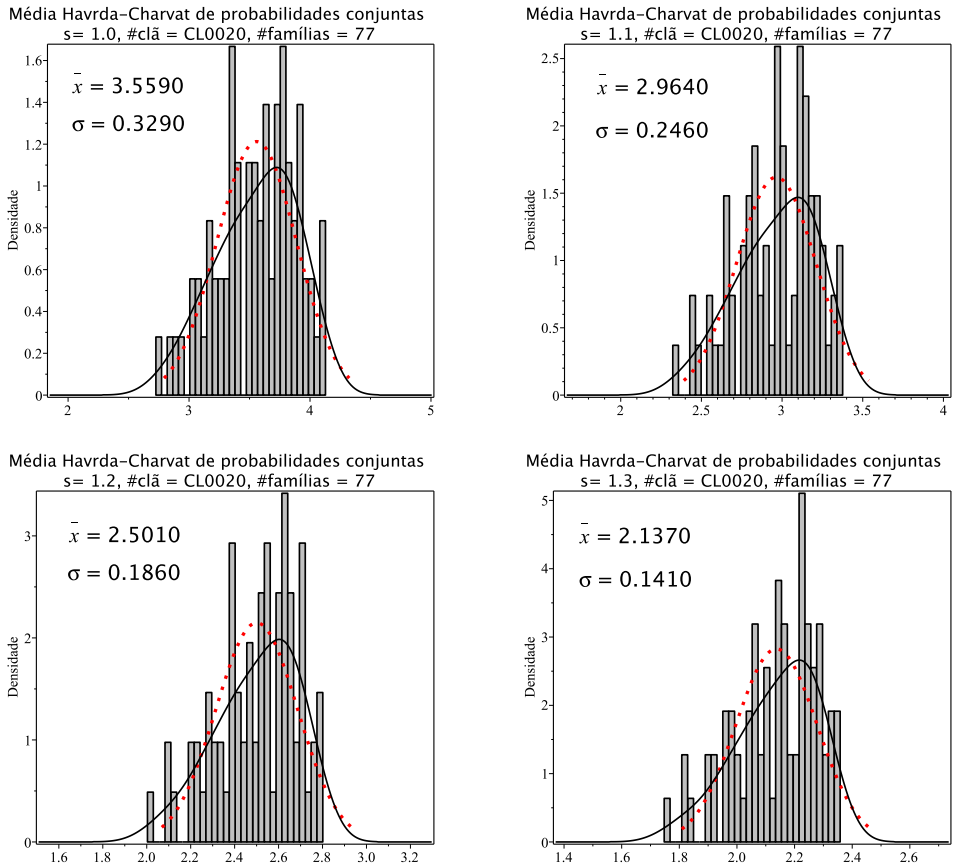

Figura 8.46: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0020. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

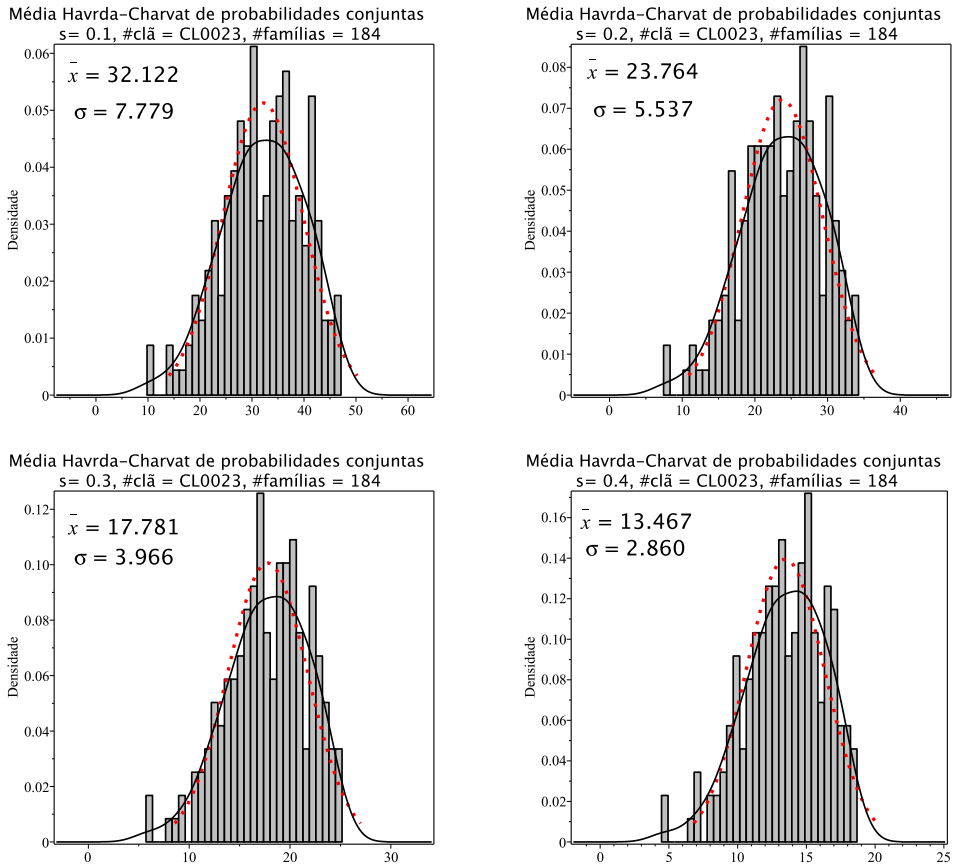

Figura 8.47: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0023. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

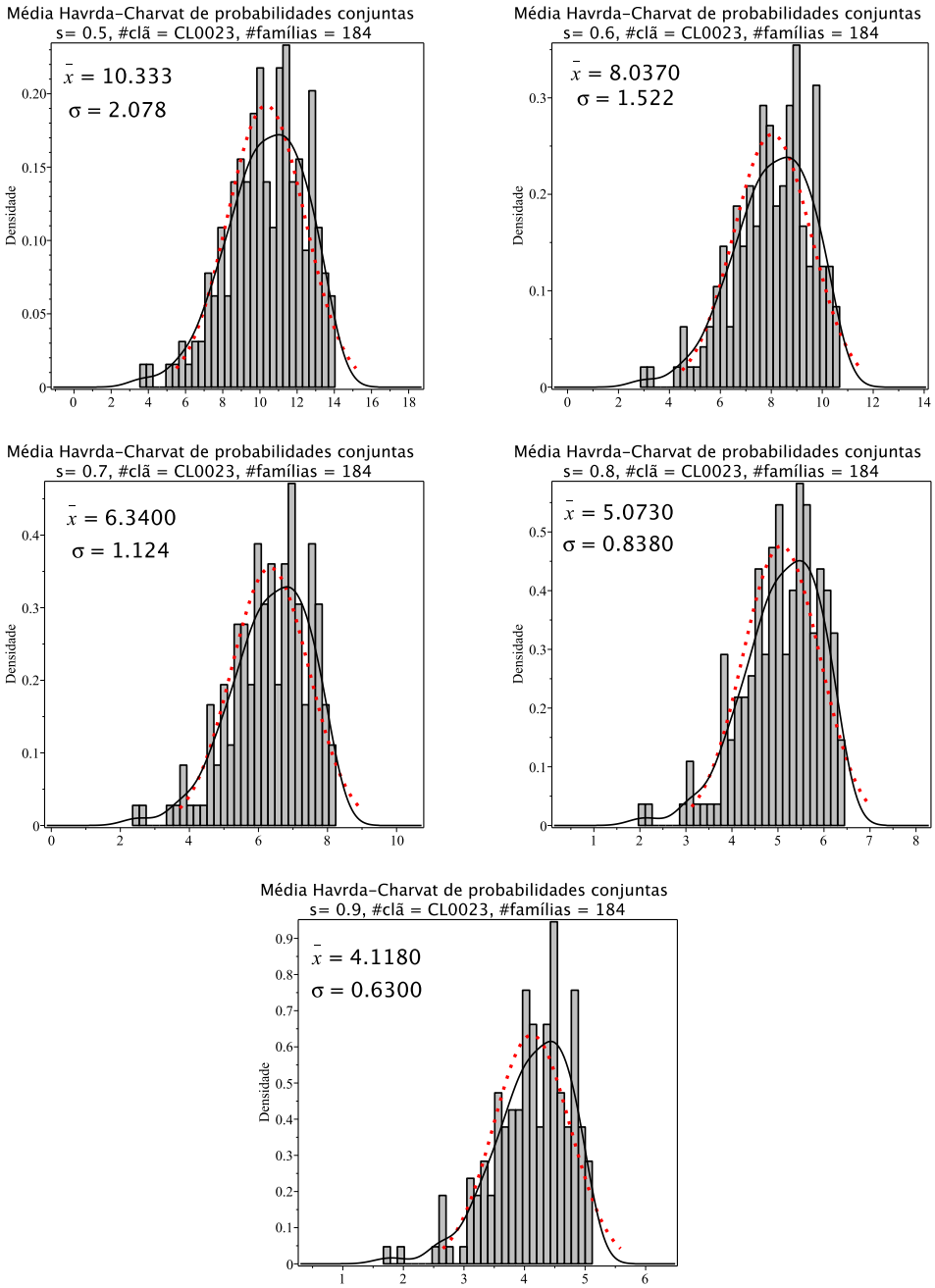

Figura 8.48: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0023. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

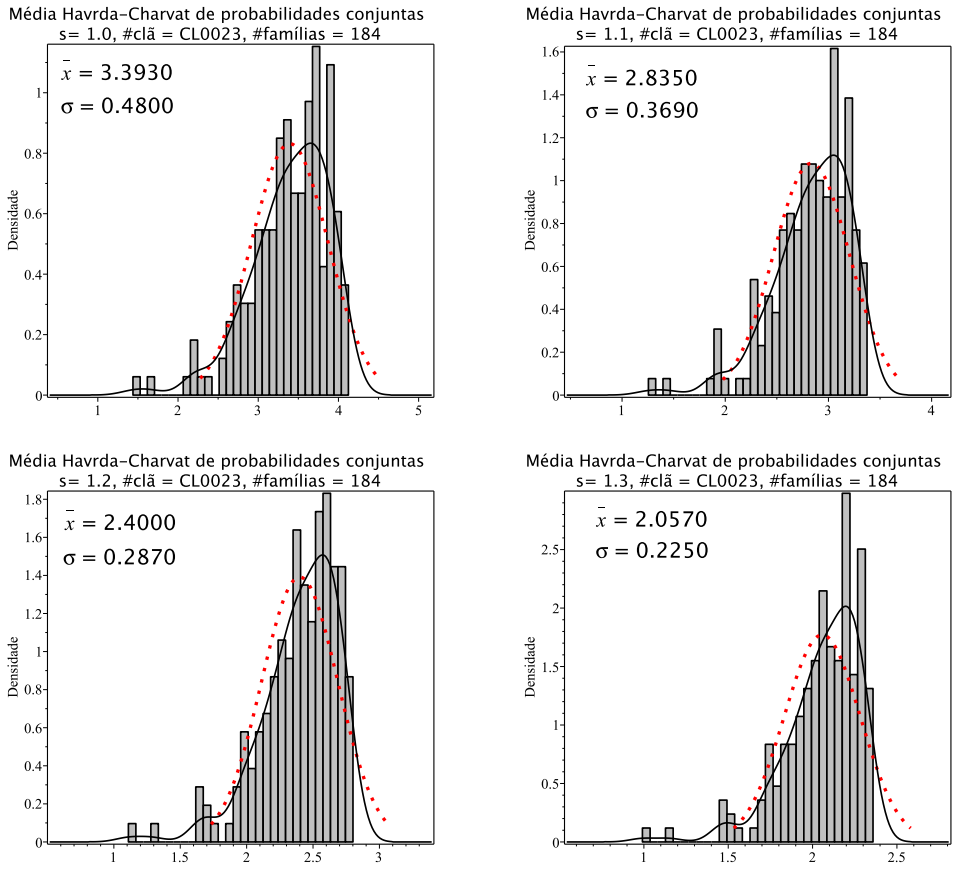

Figura 8.49: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0023. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

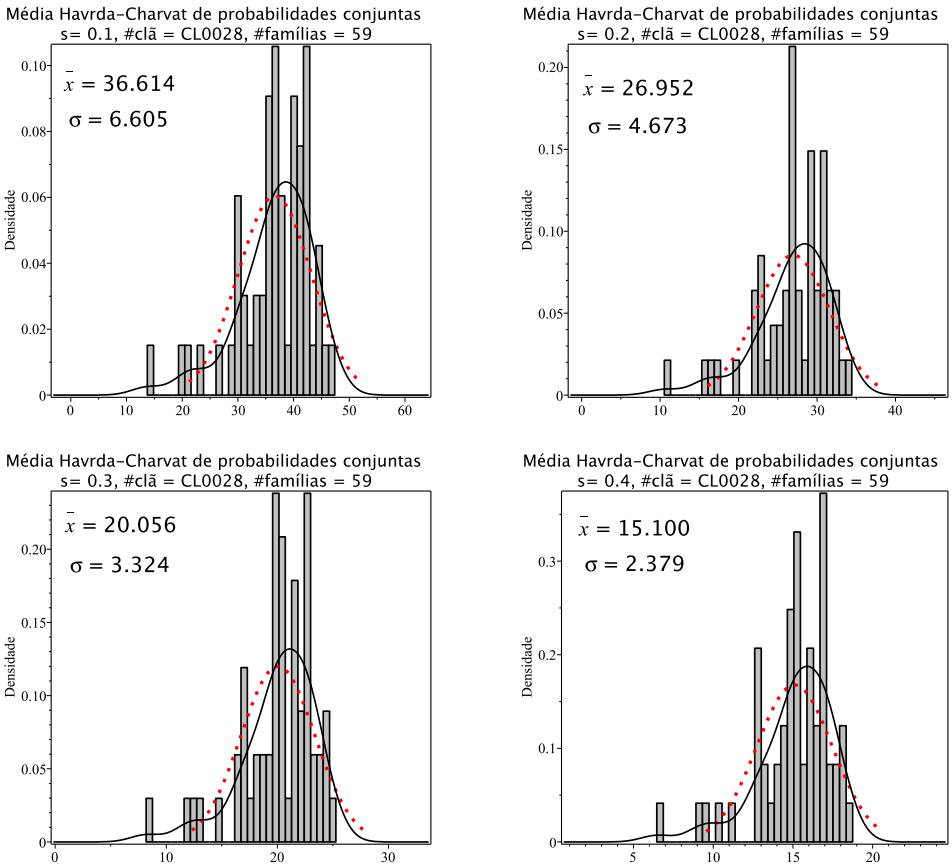

Figura 8.50: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0028. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

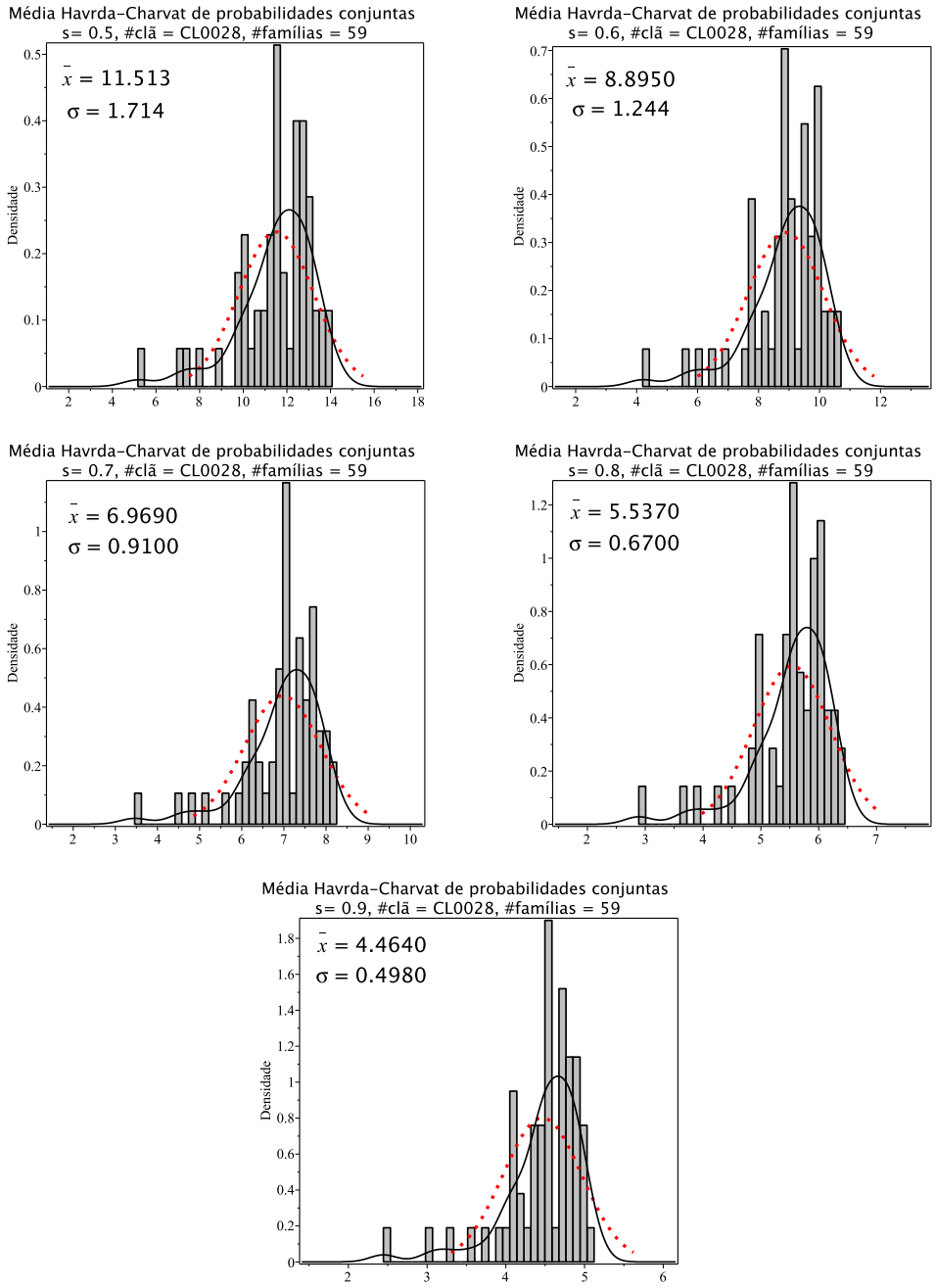

Figura 8.51: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0028. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

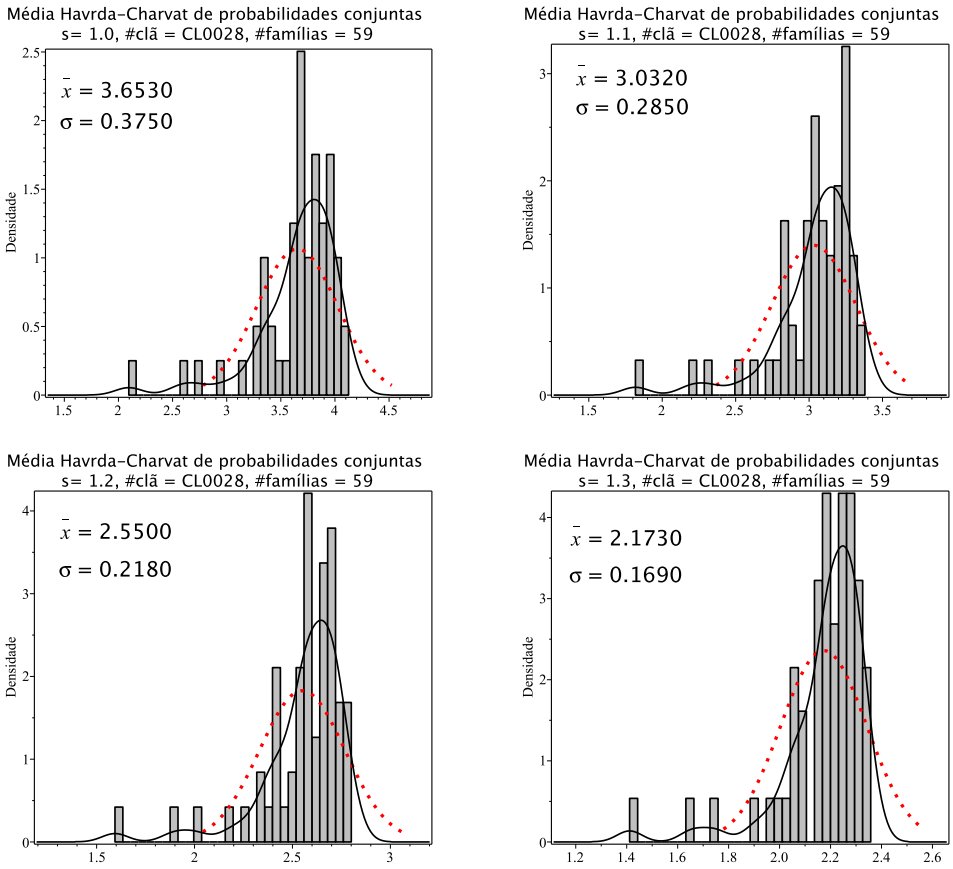

Figura 8.52: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0028. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

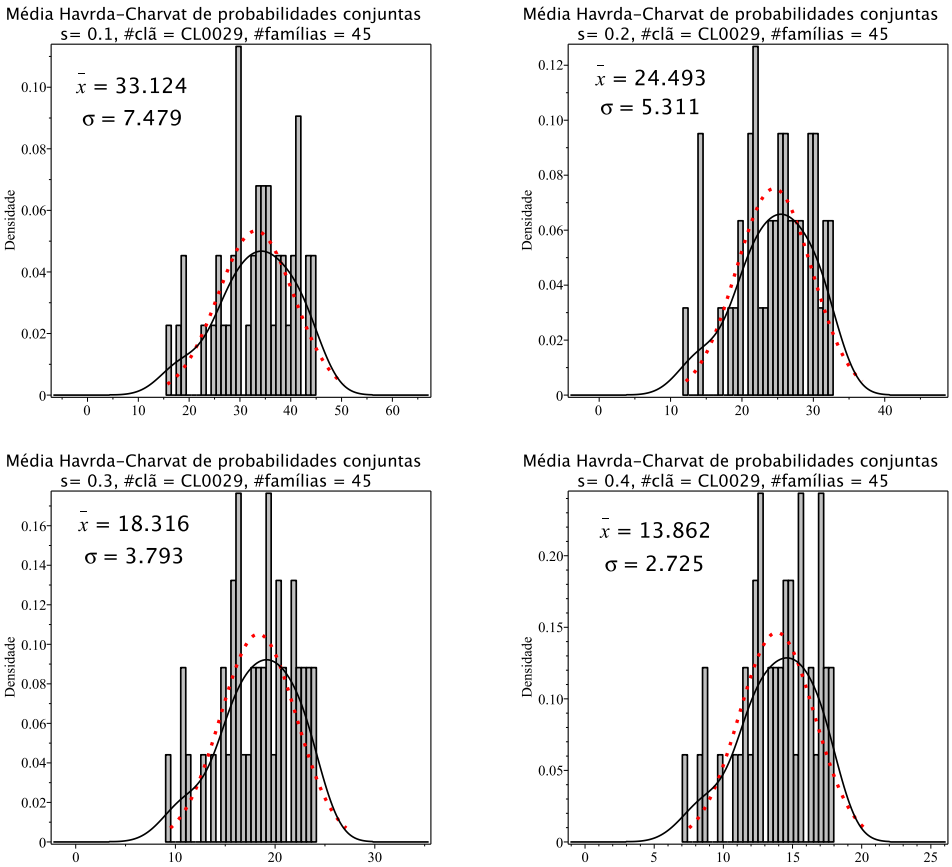

Figura 8.53: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0029. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

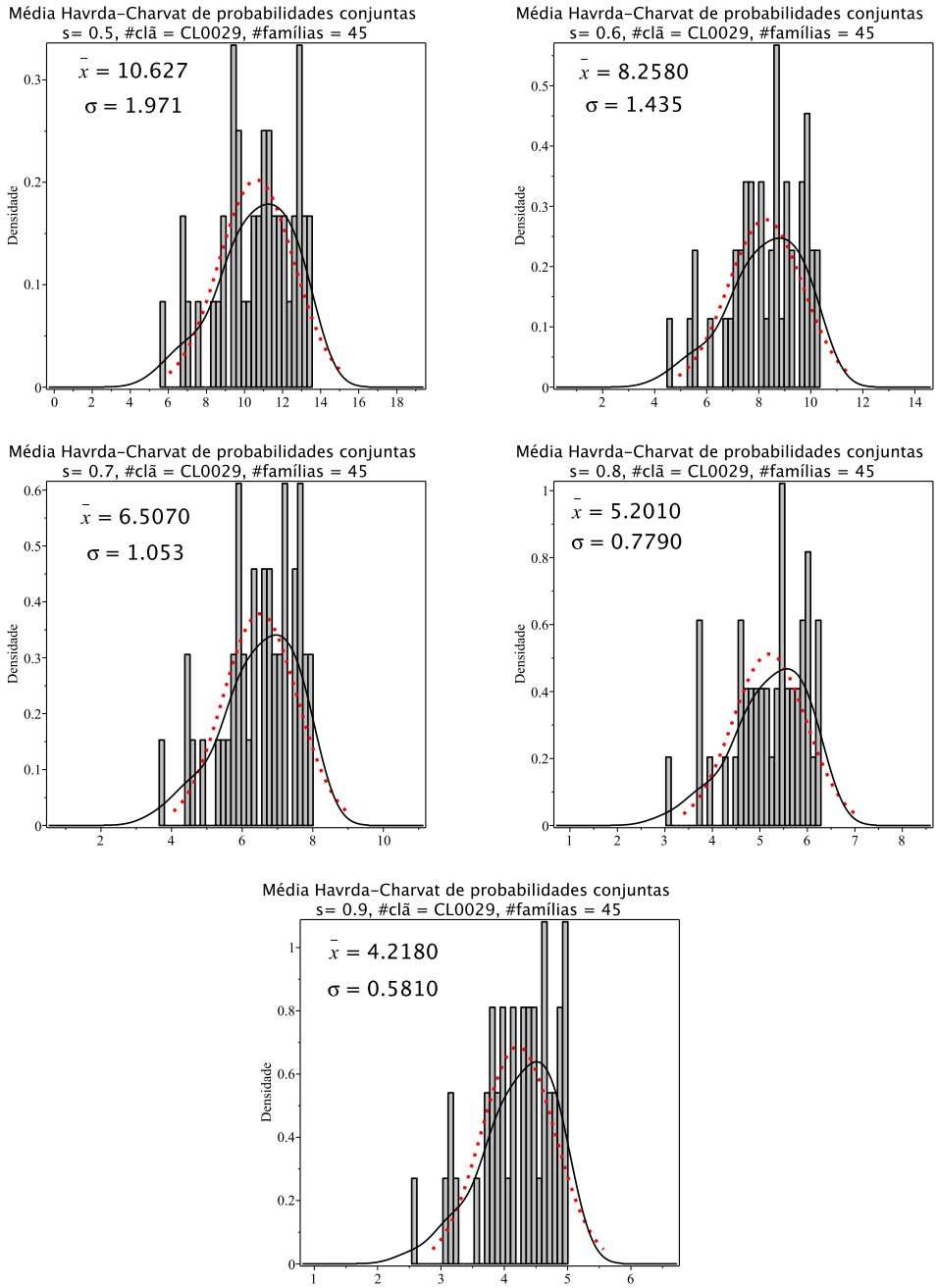

Figura 8.54: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0029. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

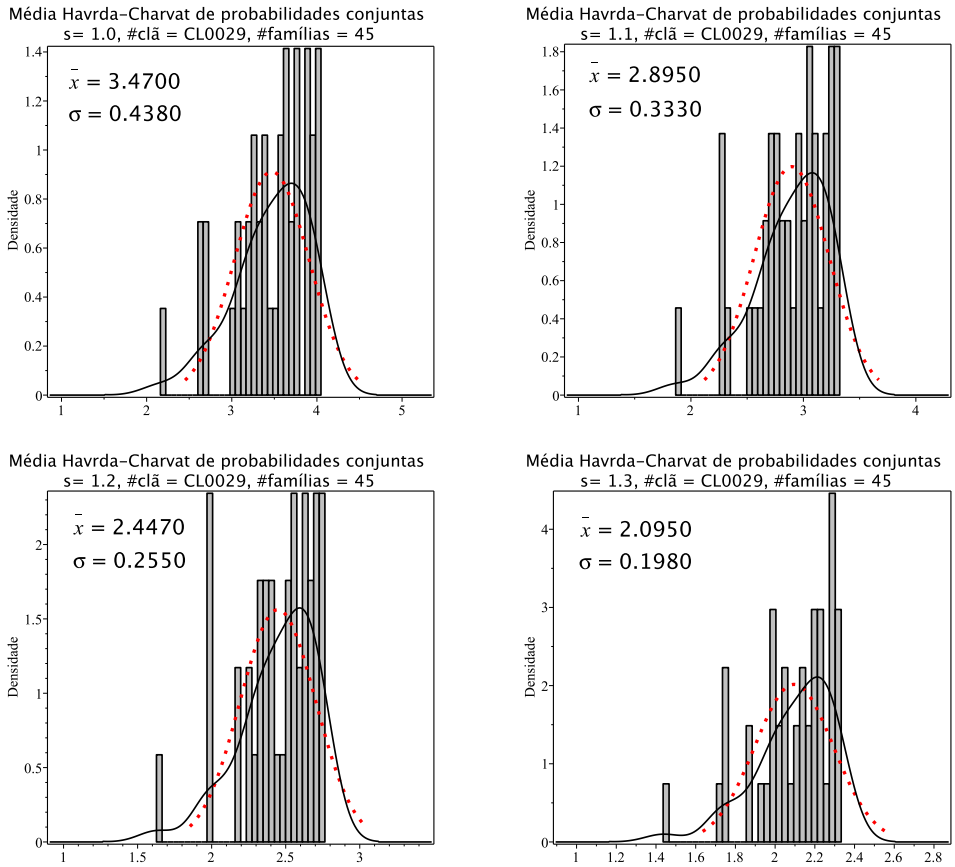

Figura 8.55: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0029. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

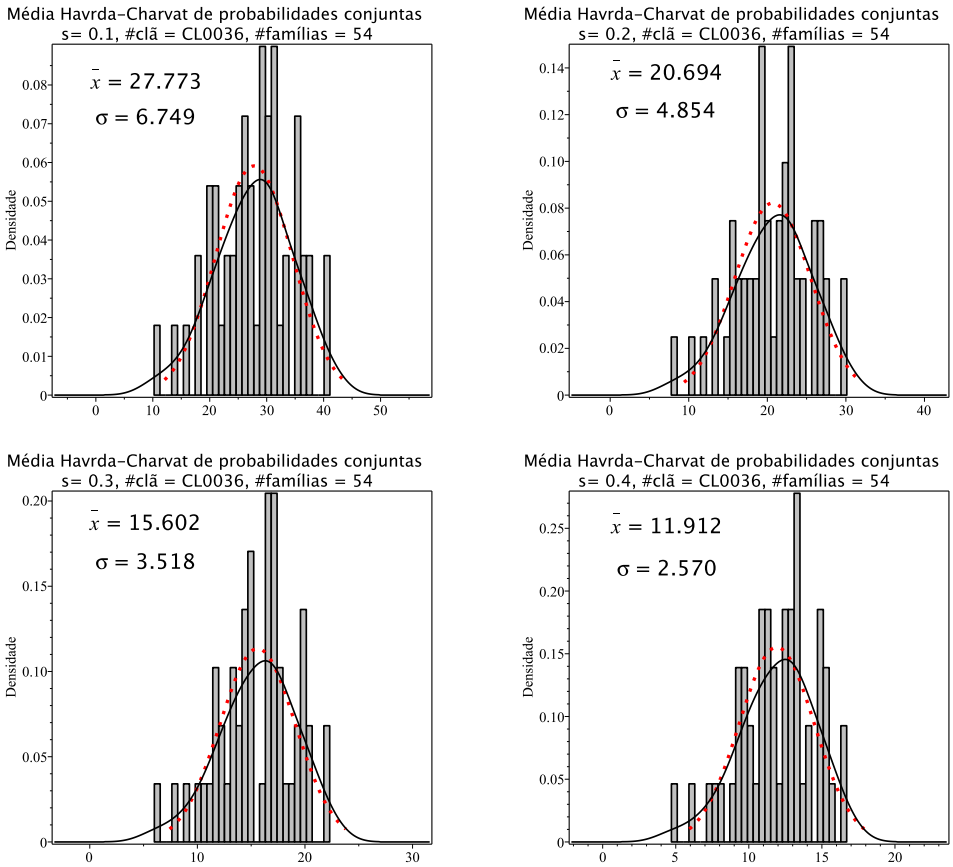

Figura 8.56: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0036. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

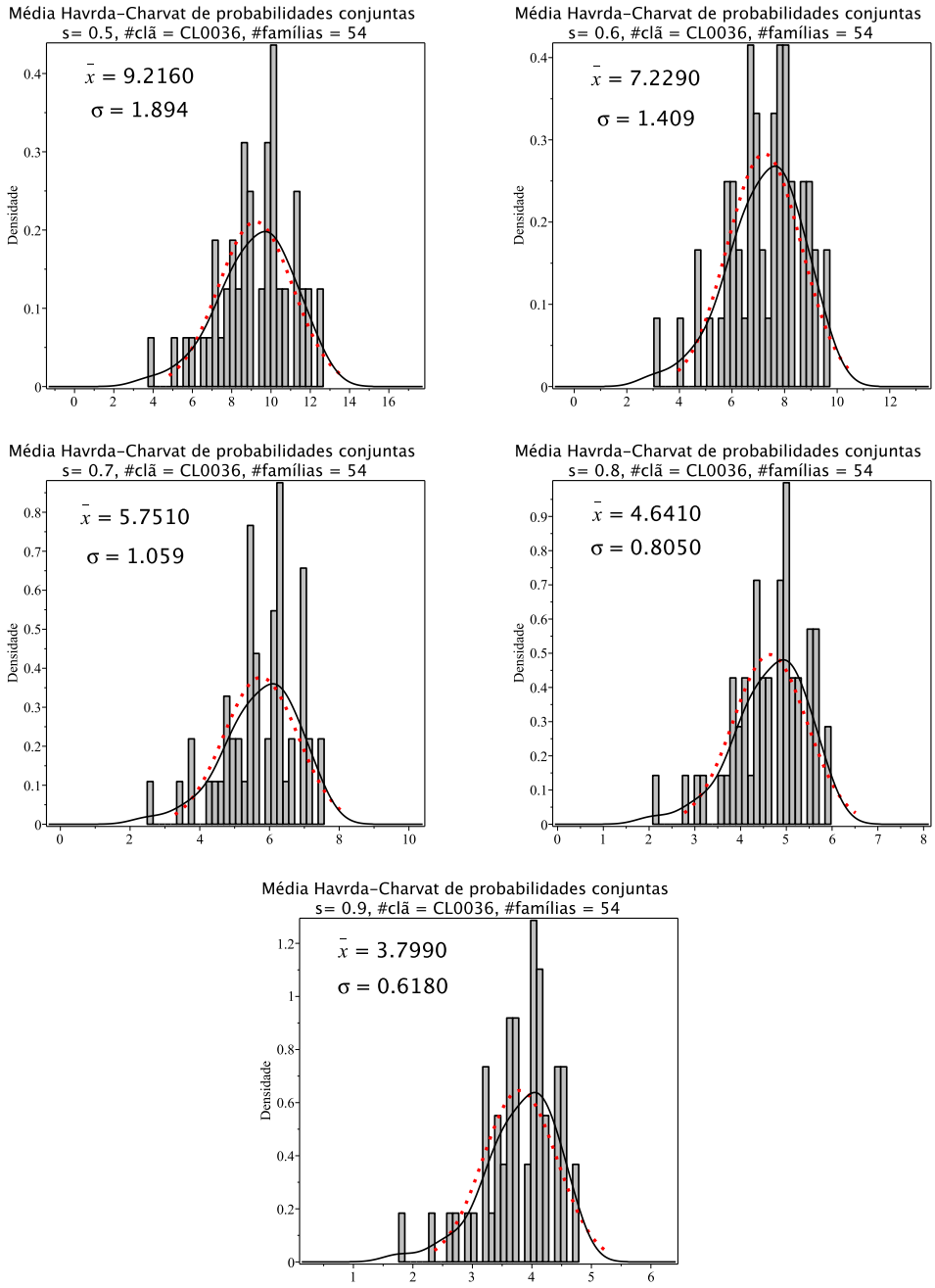

Figura 8.57: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0036. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

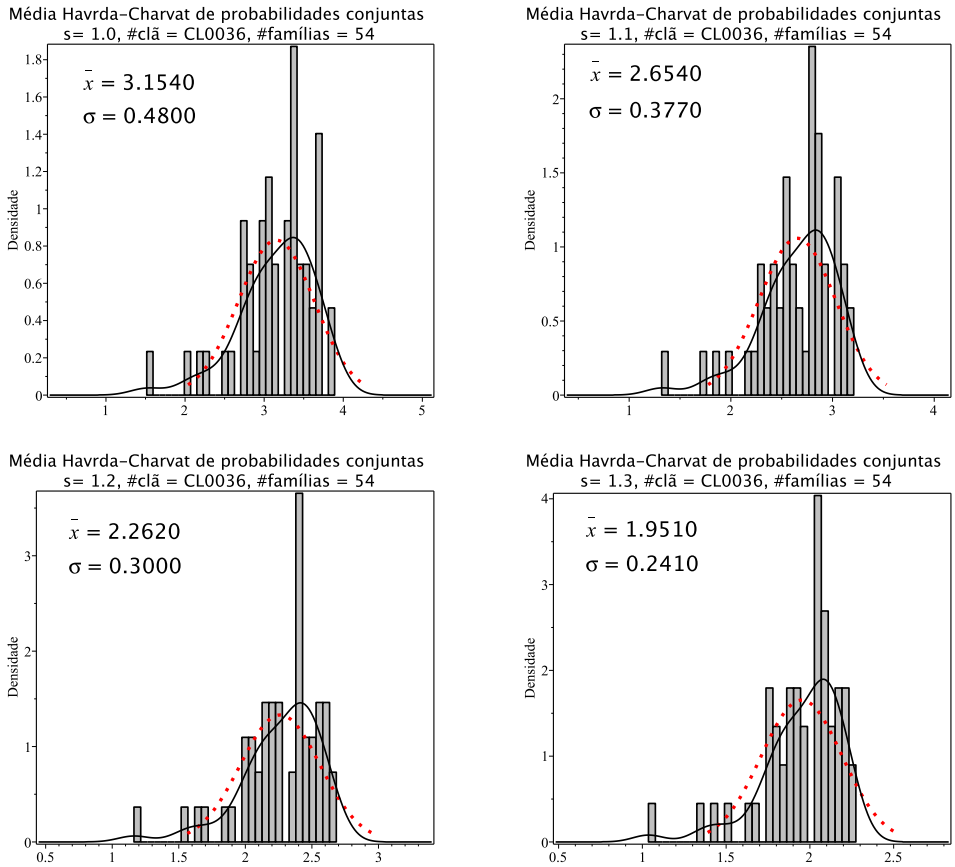

Figura 8.58: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0036. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

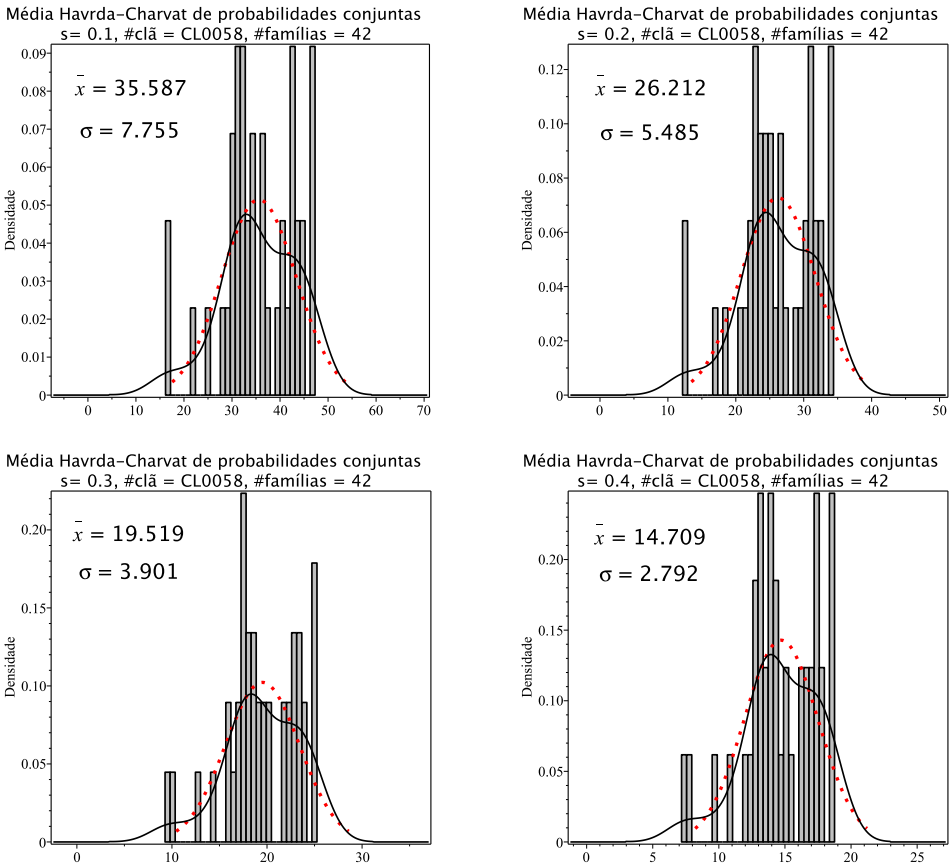

Figura 8.59: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0058. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

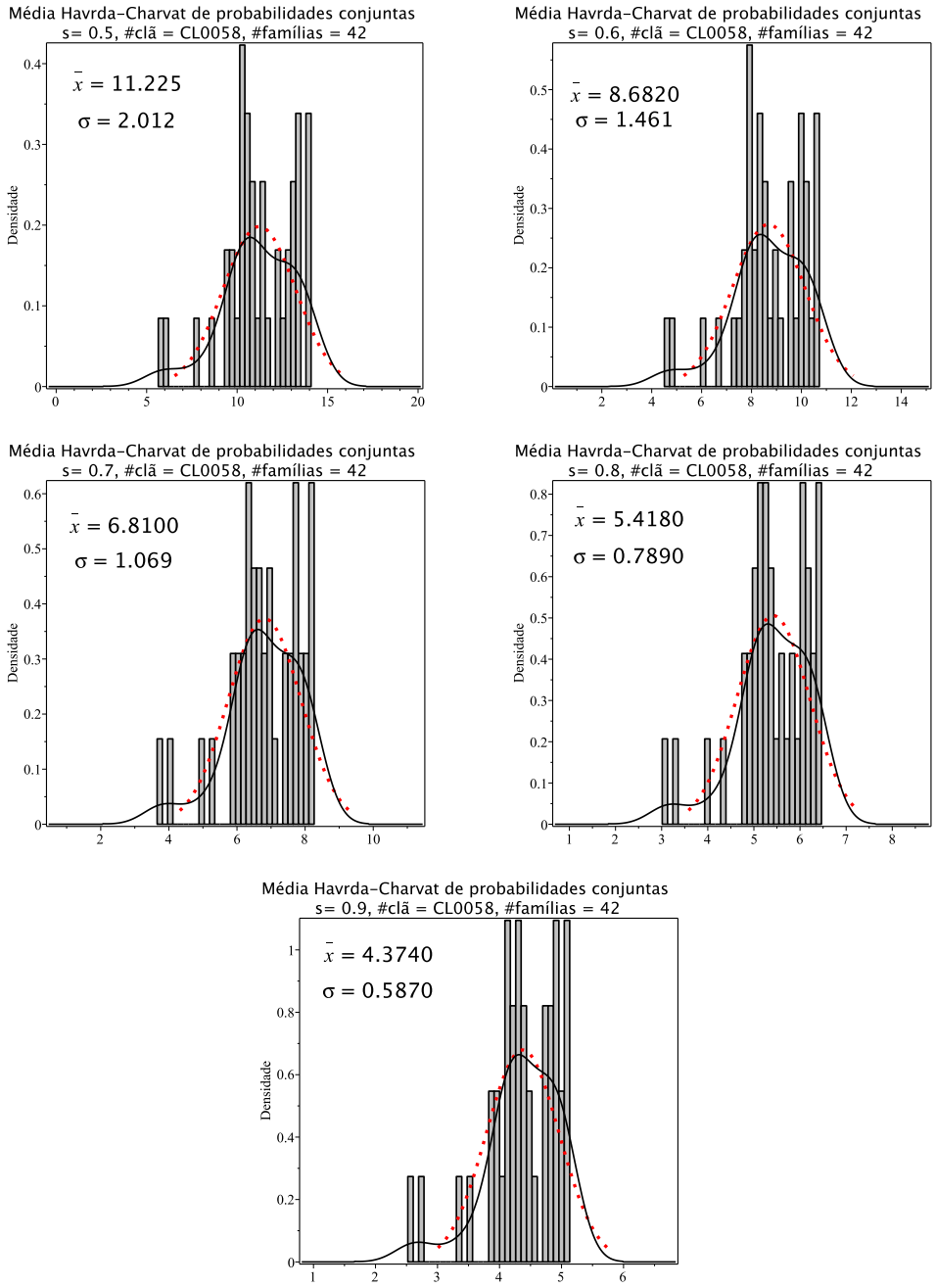

Figura 8.60: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0058. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

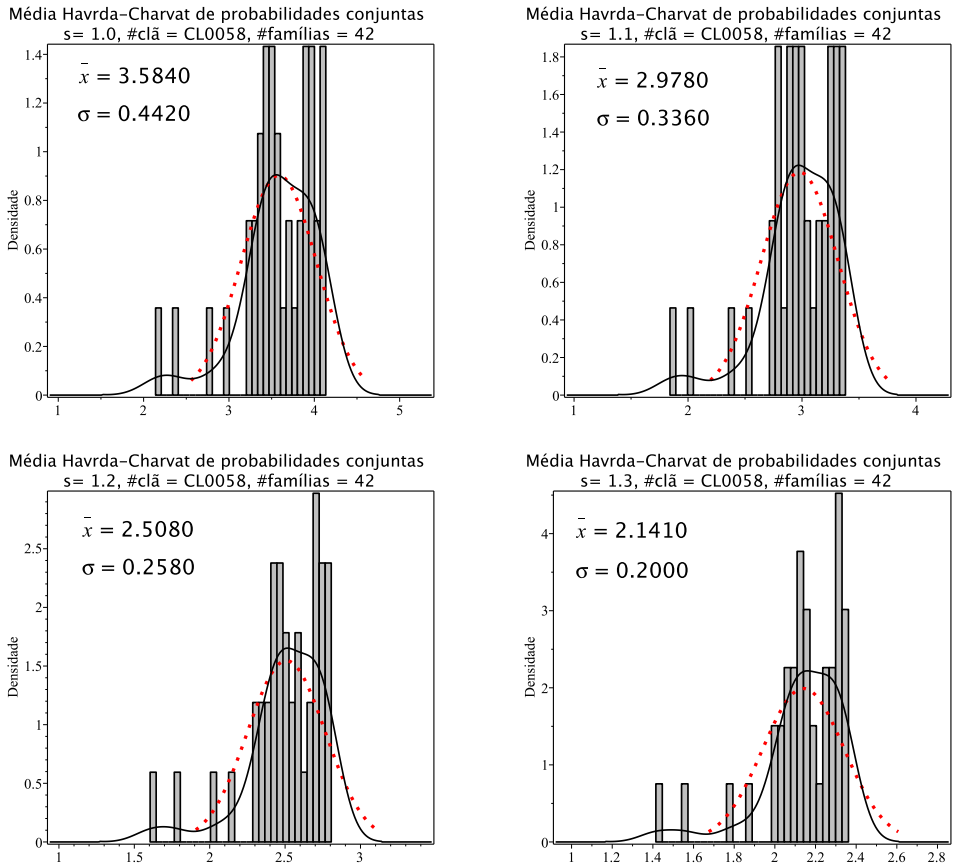

Figura 8.61: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0058. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

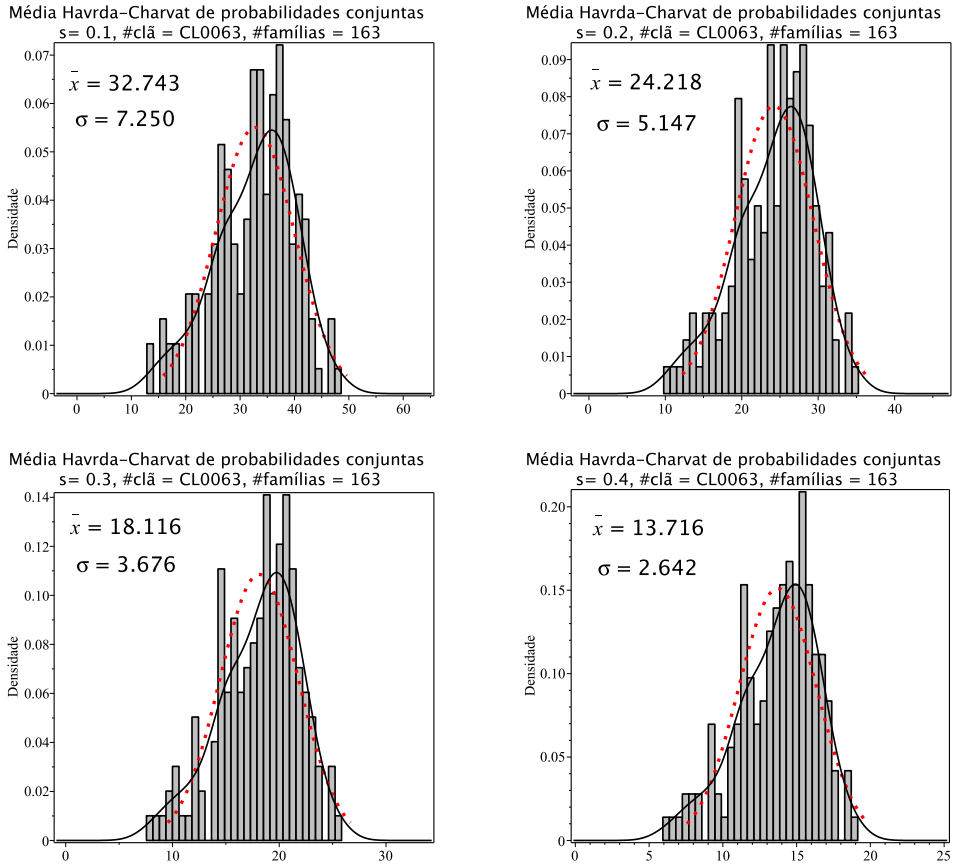

Figura 8.62: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0063. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

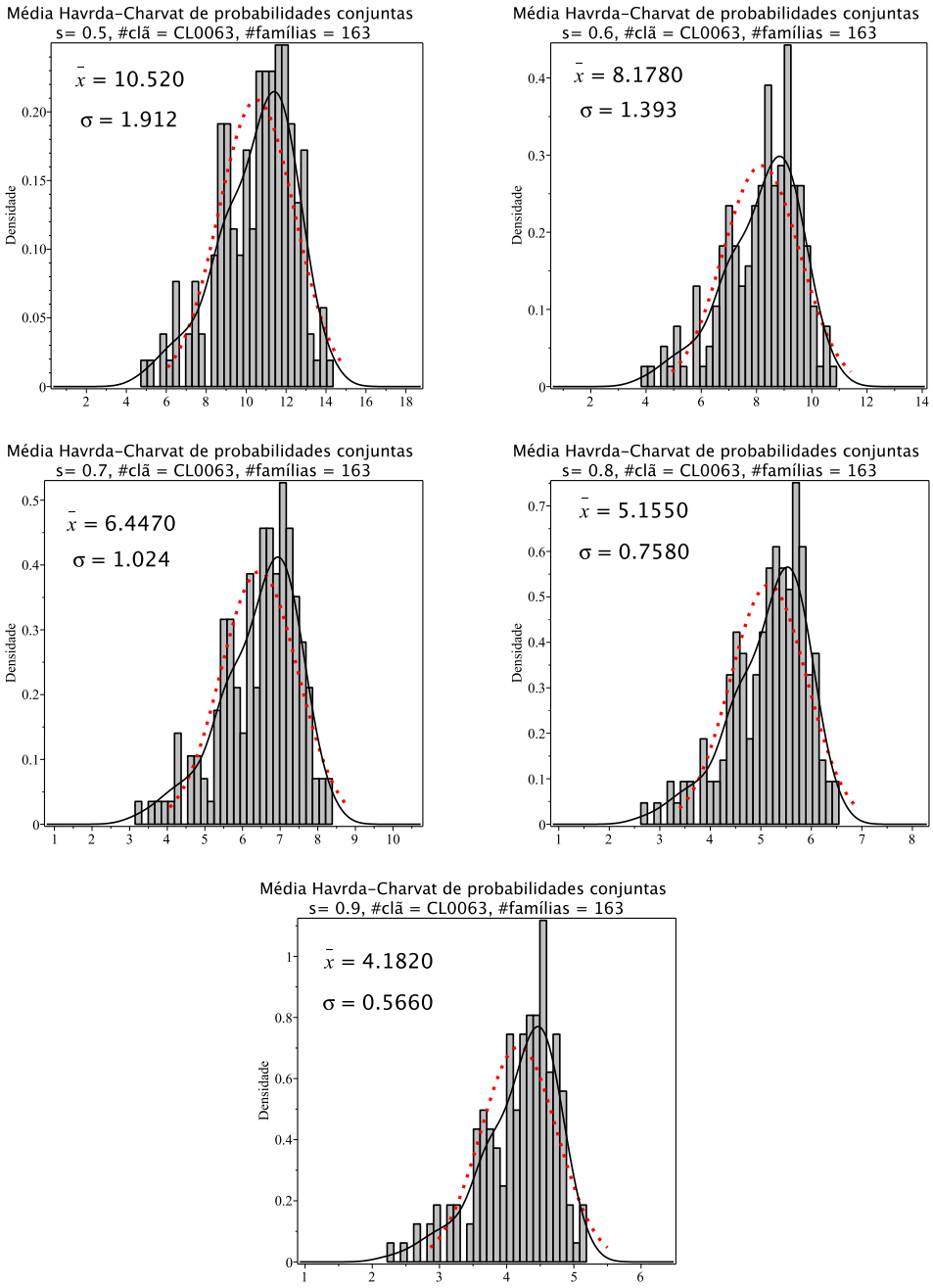

Figura 8.63: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0063. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

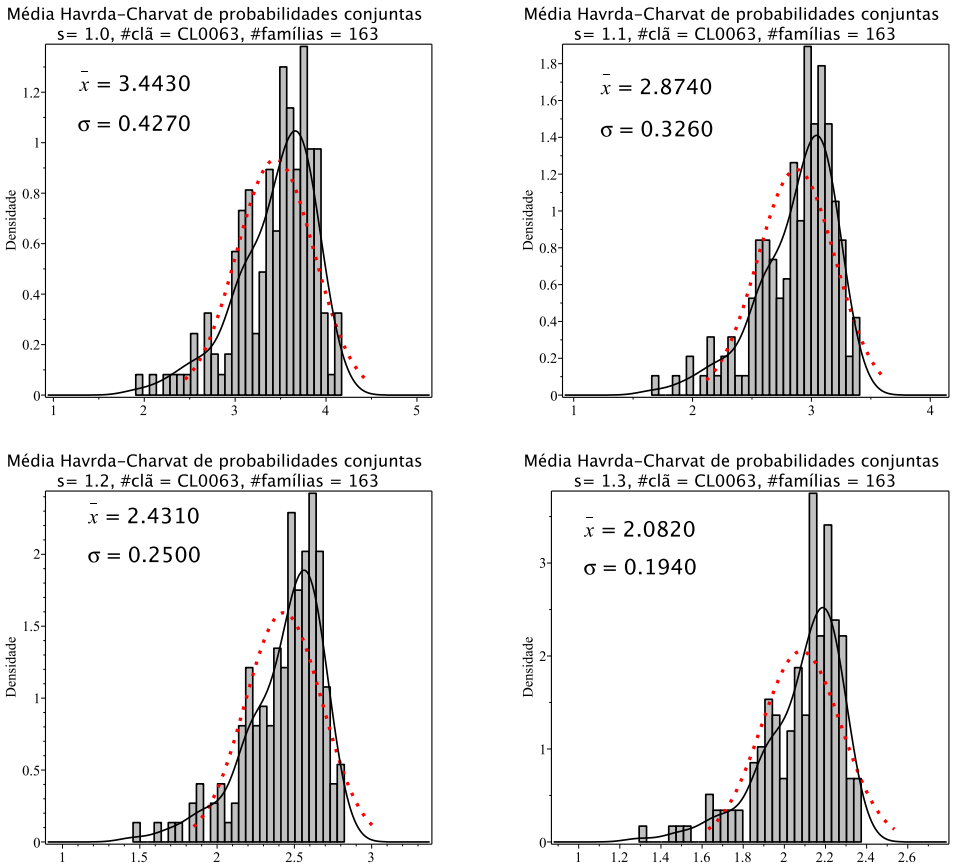

Figura 8.64: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0063. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

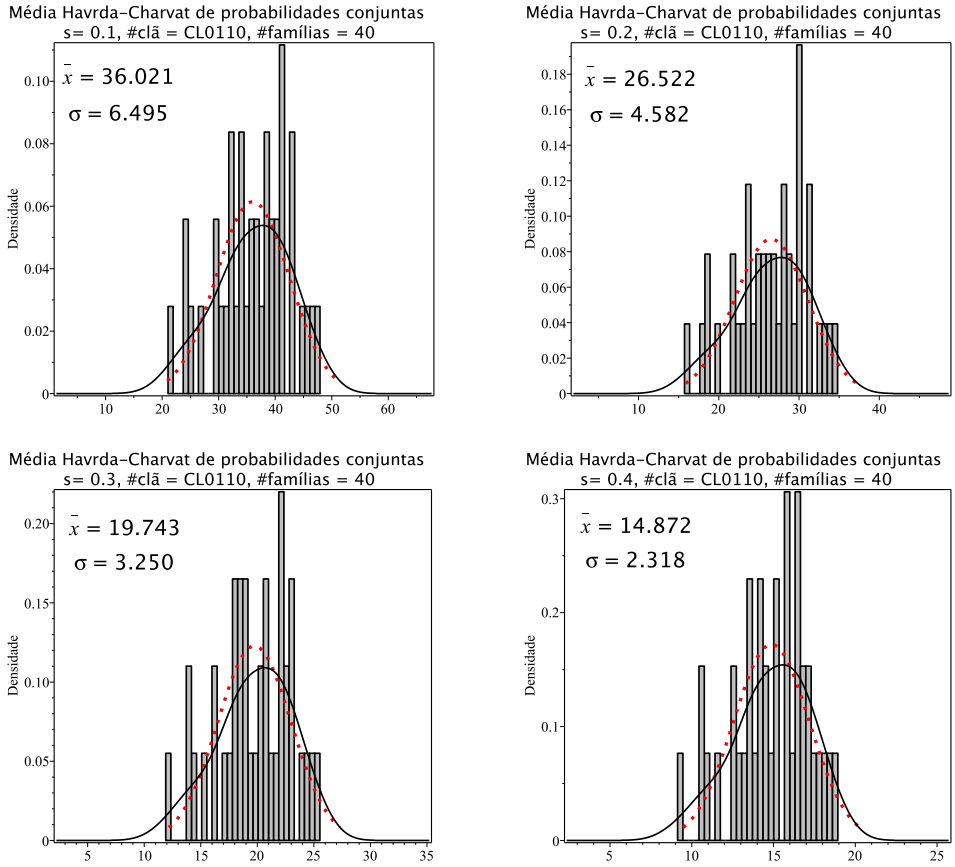

Figura 8.65: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0110. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

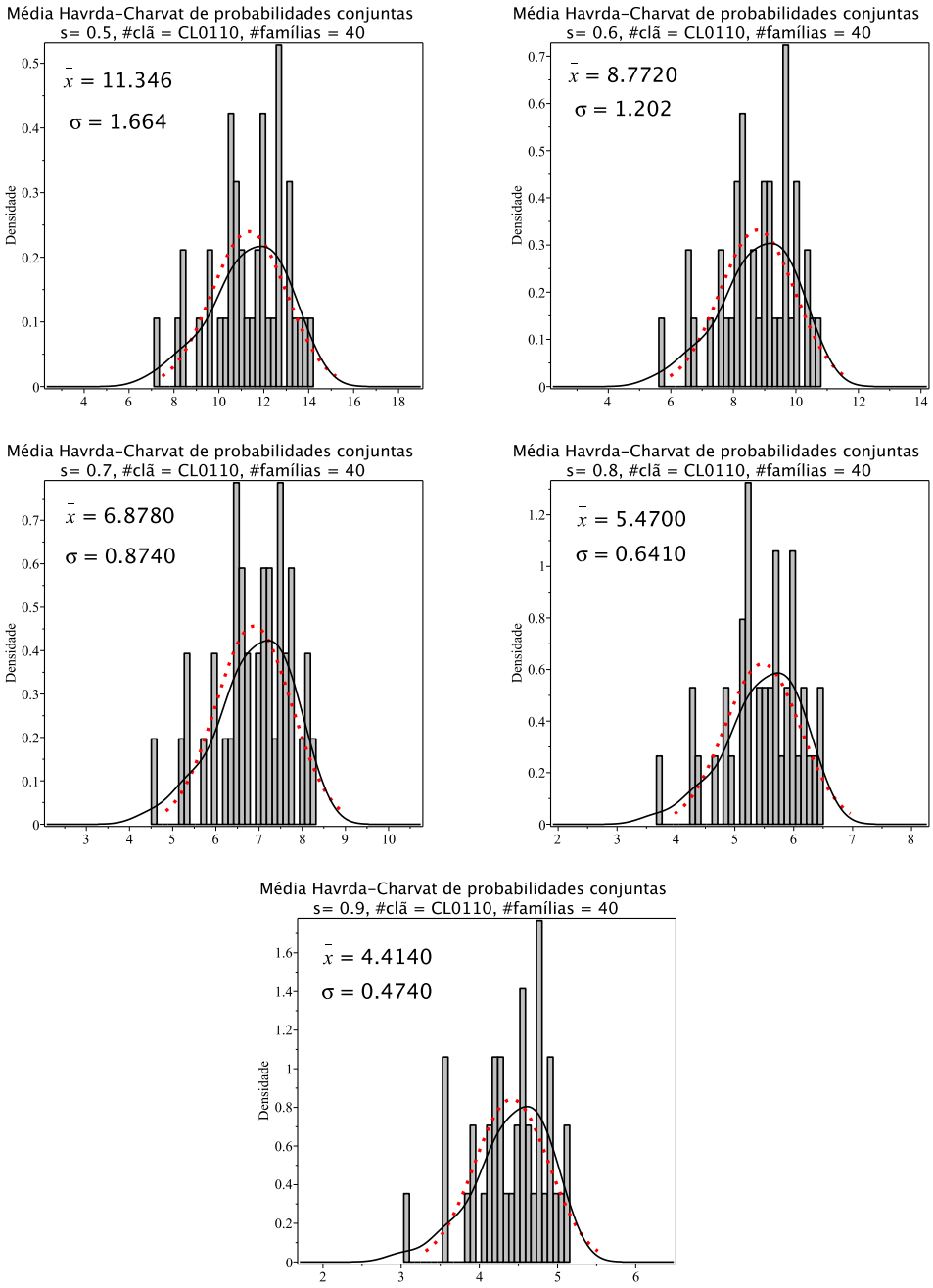

Figura 8.66: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0110. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

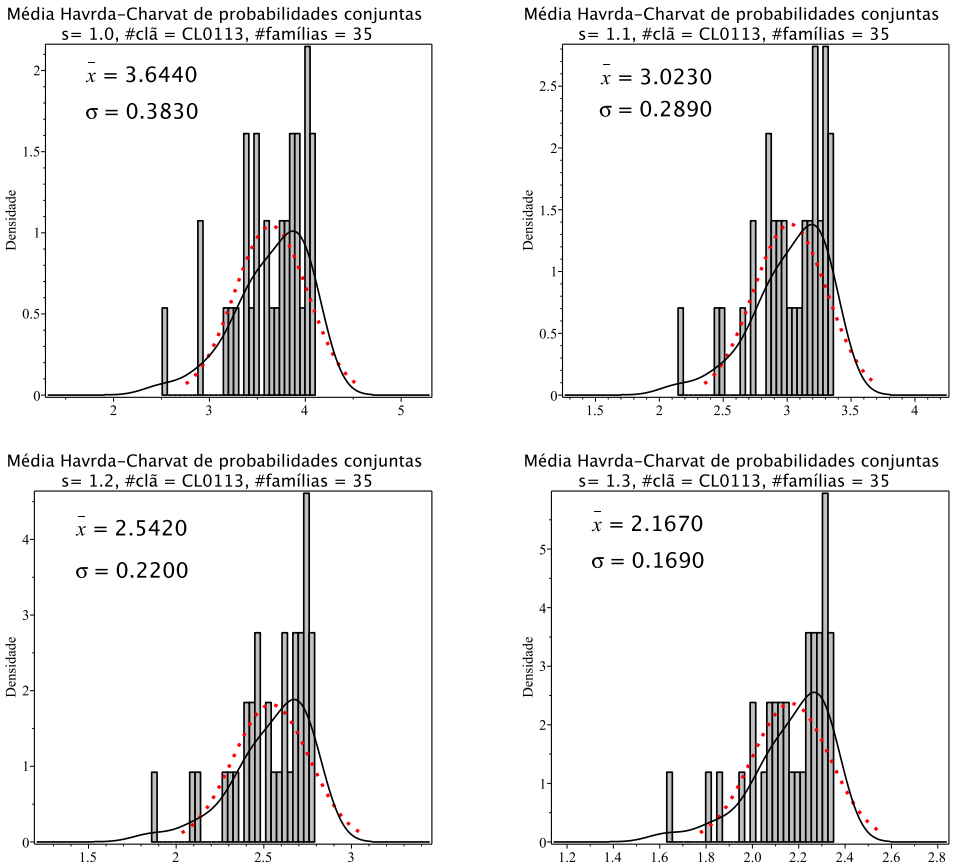

Figura 8.67: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0113. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

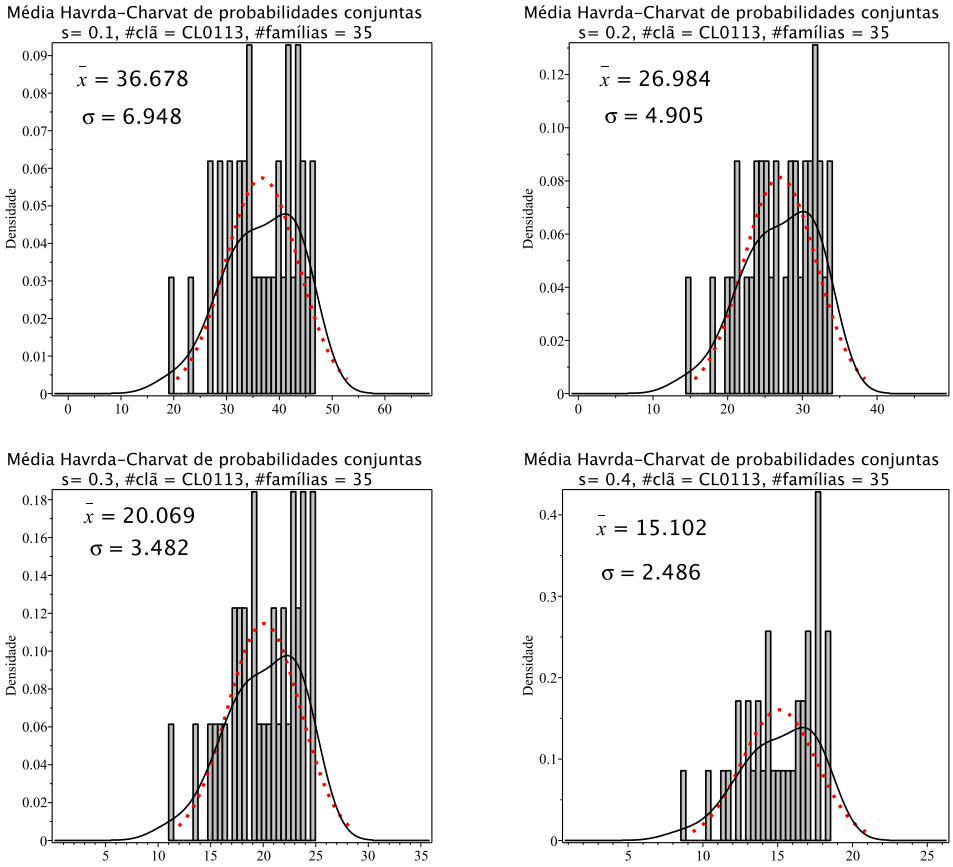

Figura 8.68: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0113. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

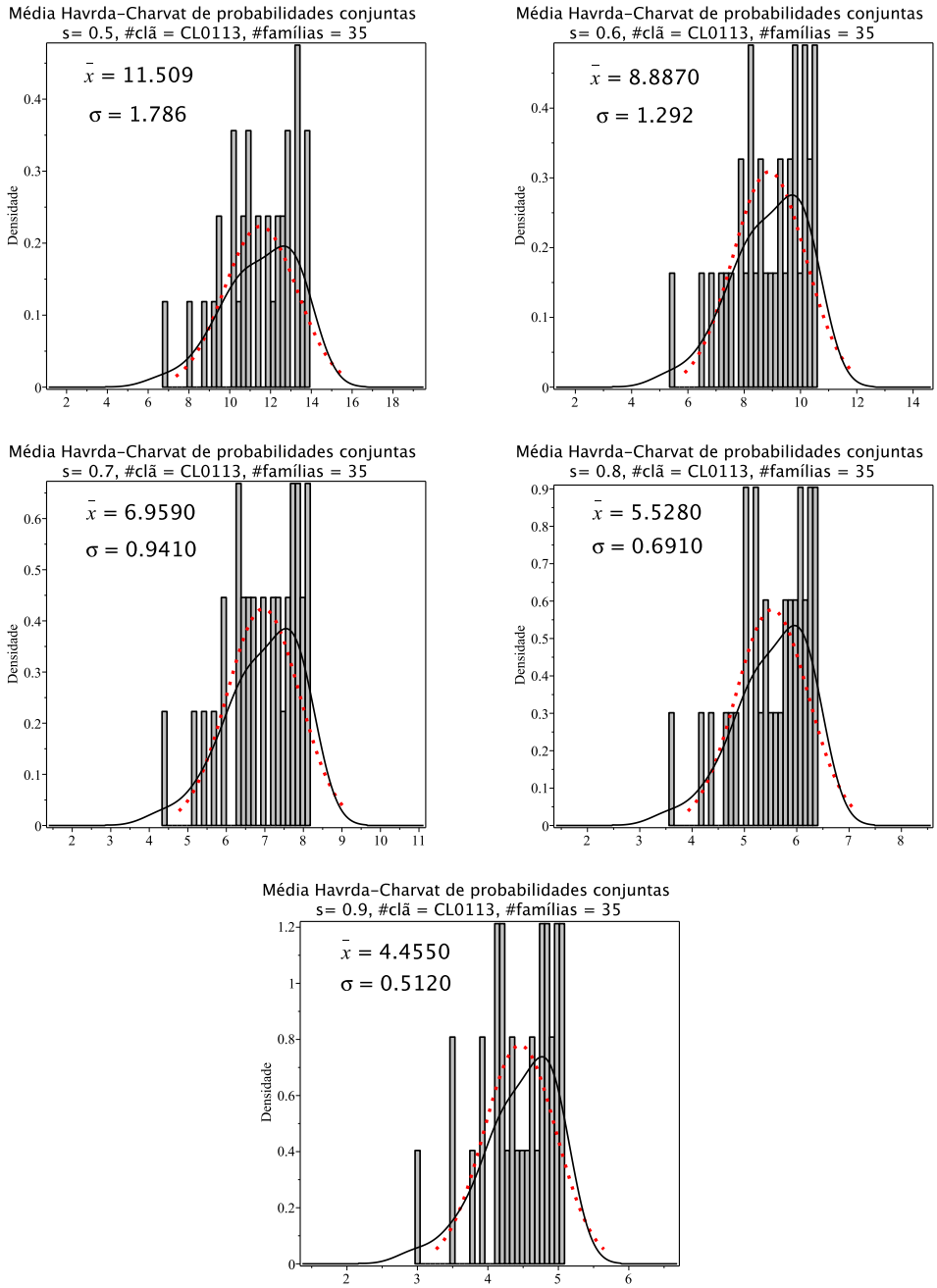

Figura 8.69: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0113. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

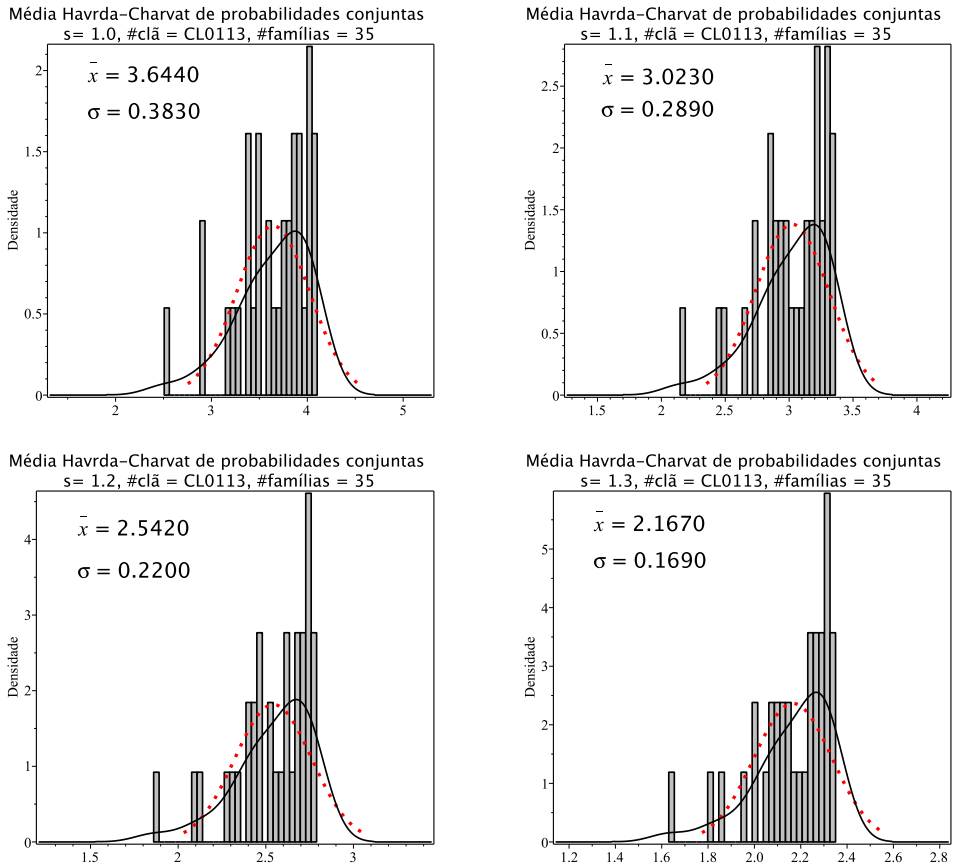

Figura 8.70: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0113. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

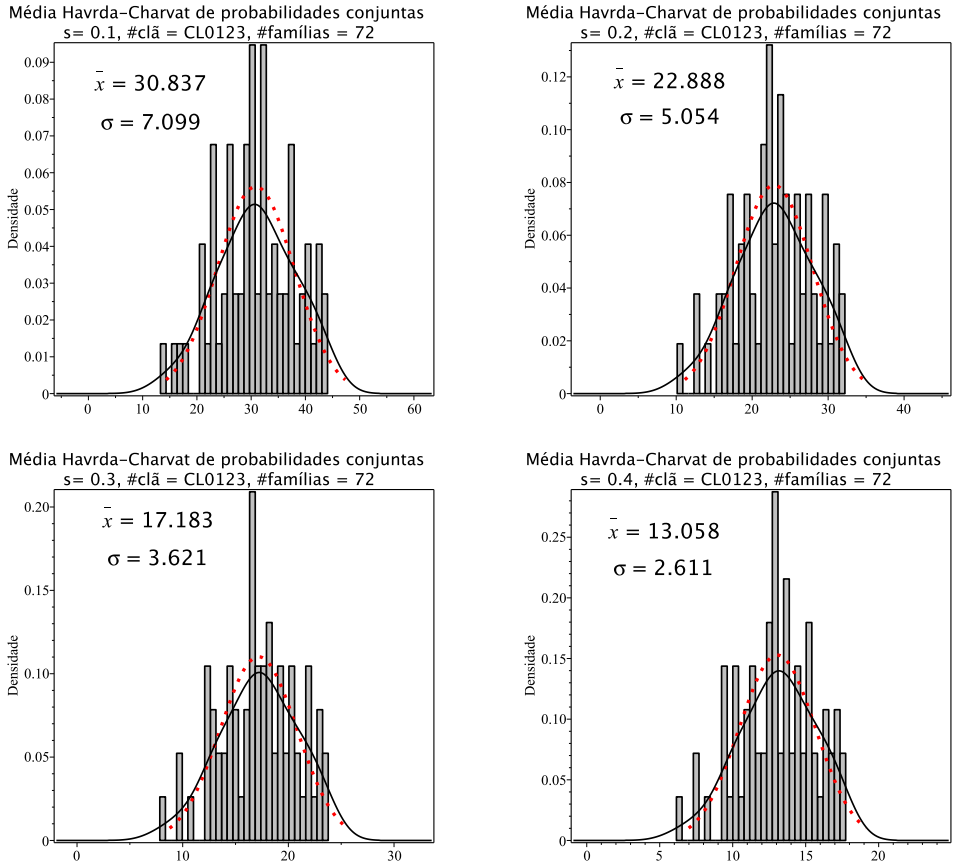

Figura 8.71: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0123. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

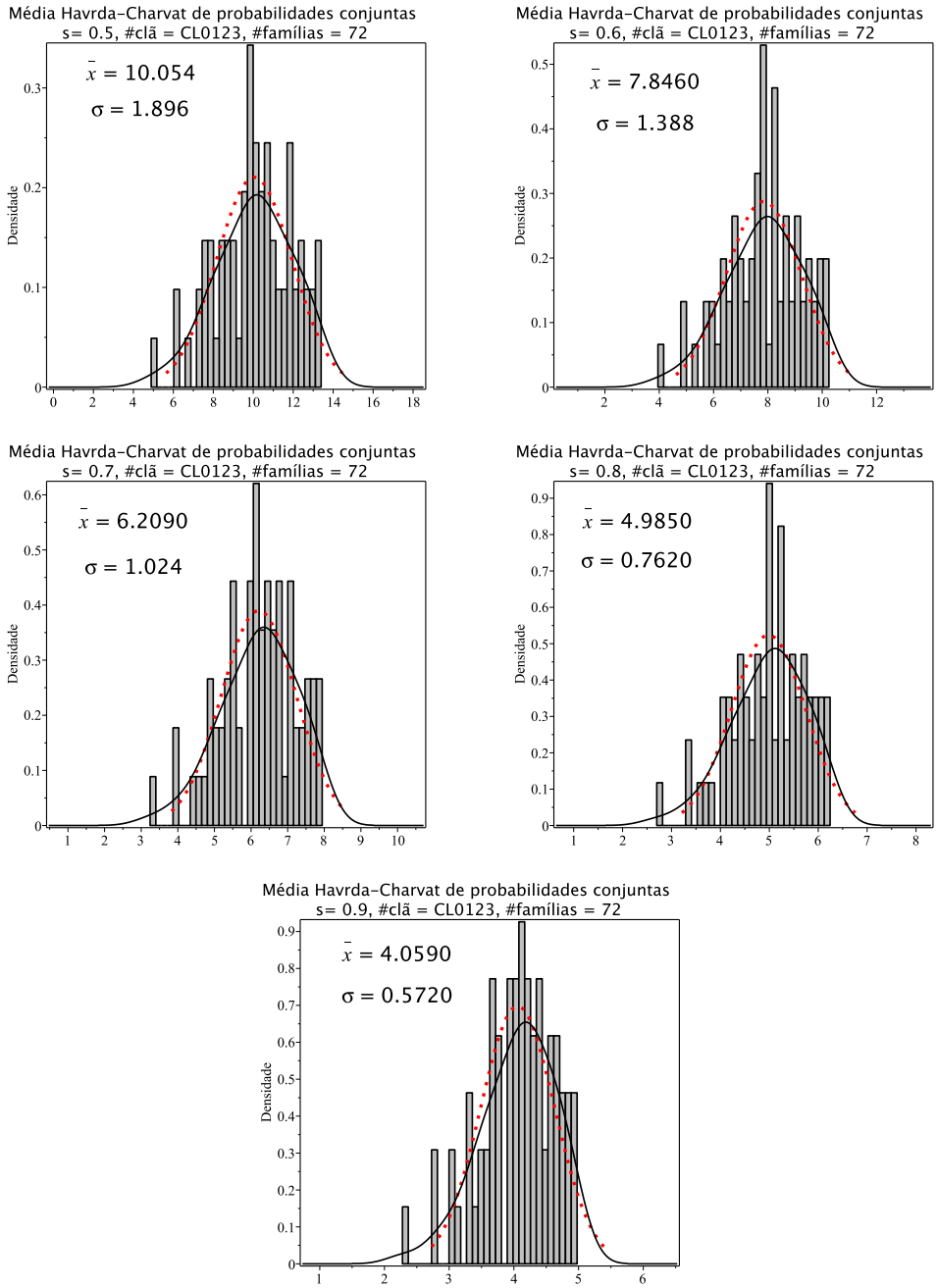

Figura 8.72: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0123. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

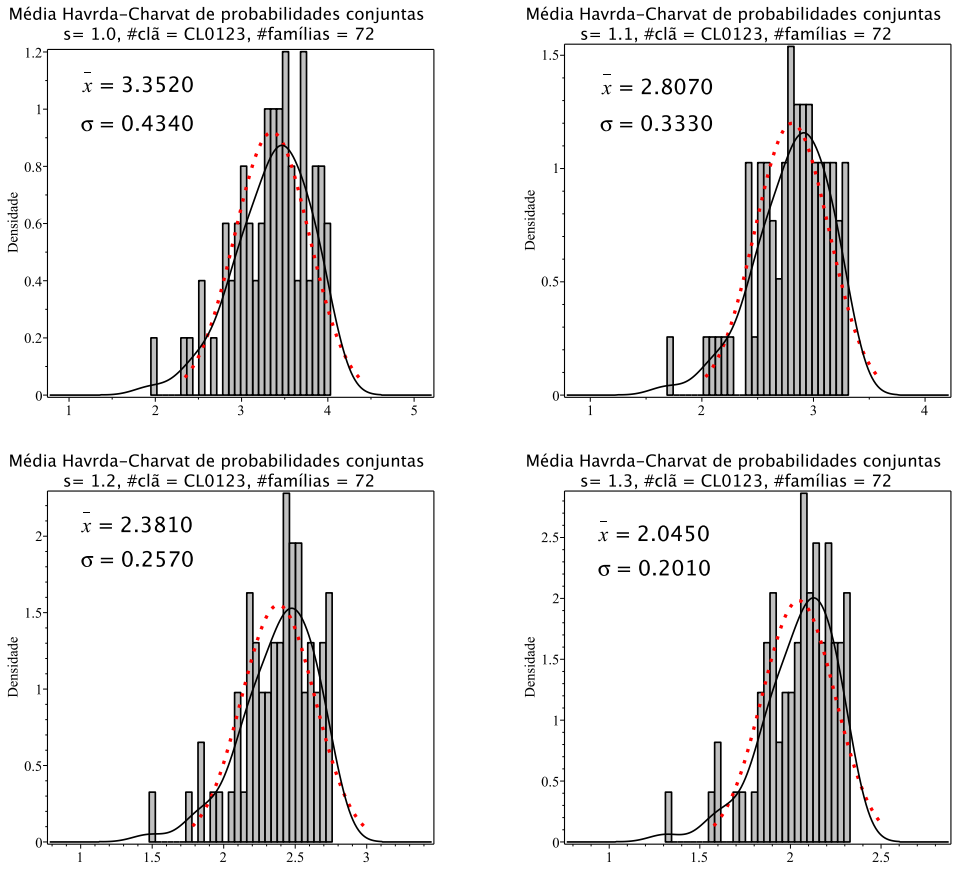

Figura 8.73: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0123. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

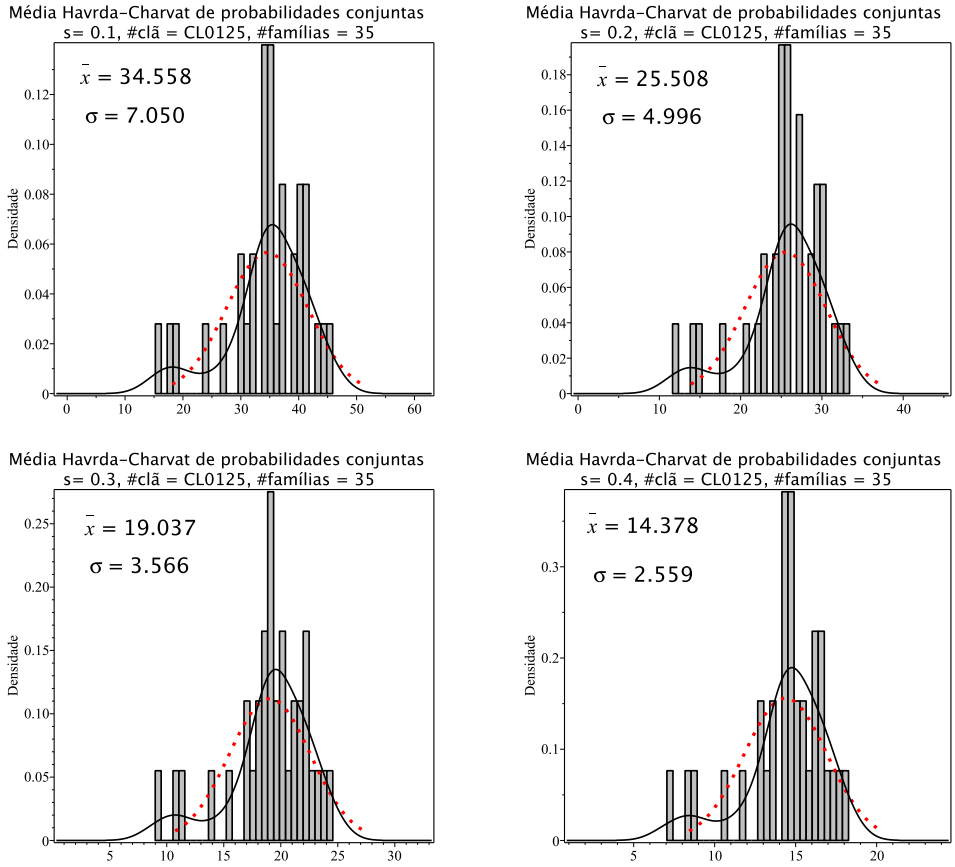

Figura 8.74: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0125. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

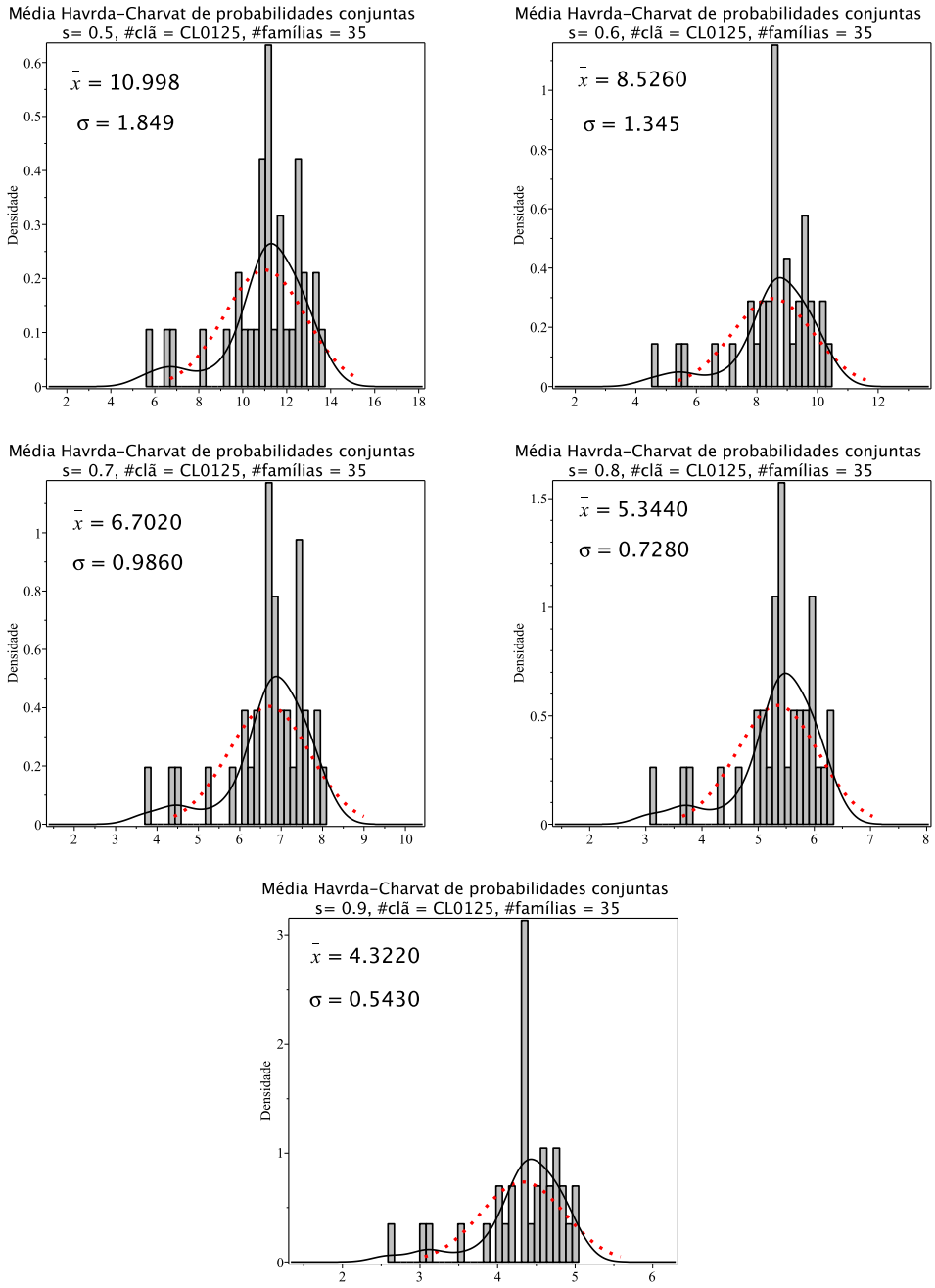

Figura 8.75: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0125. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

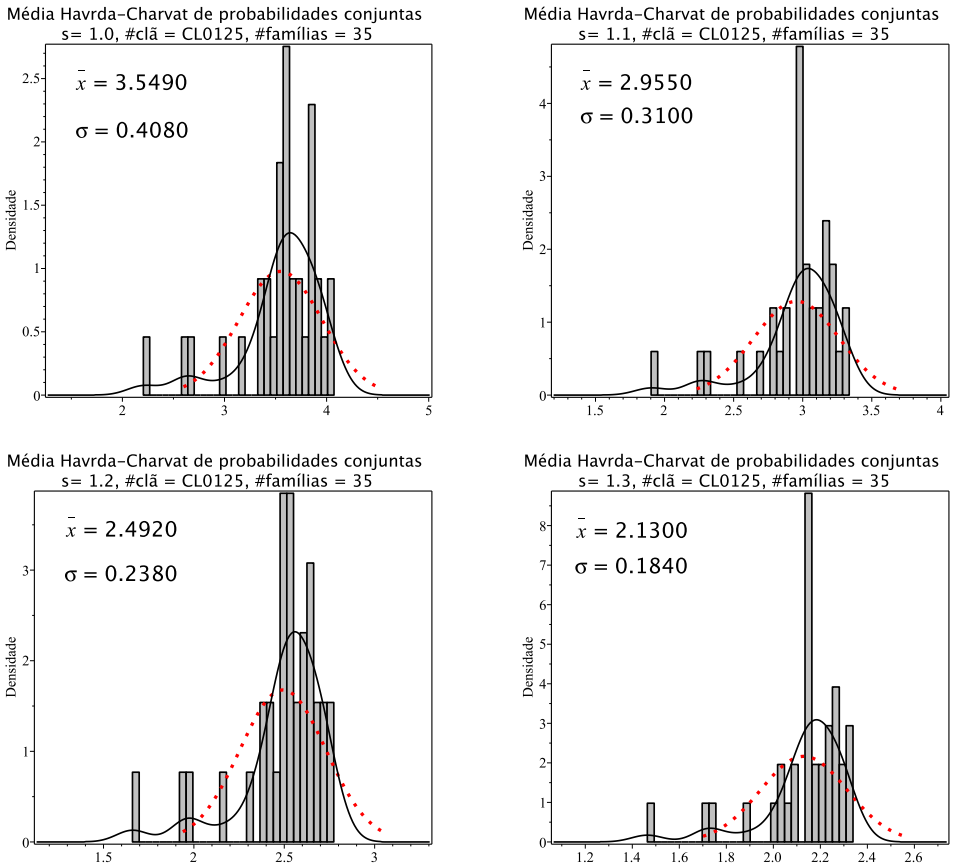

Figura 8.76: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0125. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

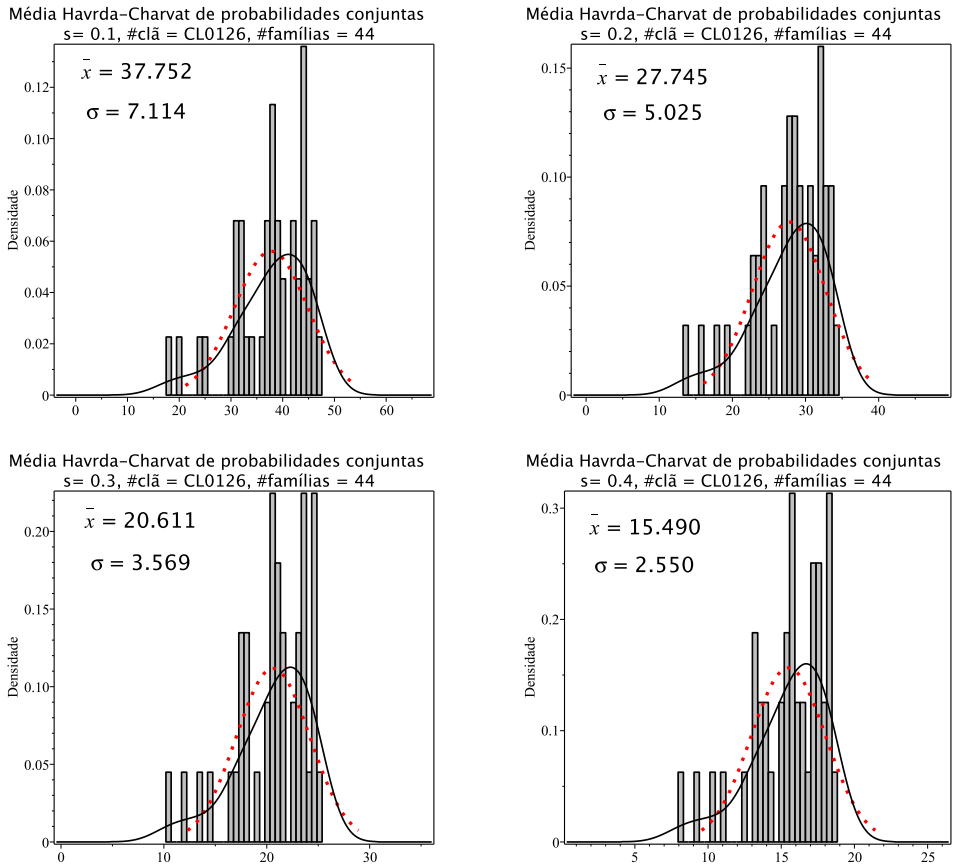

Figura 8.77: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0126. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

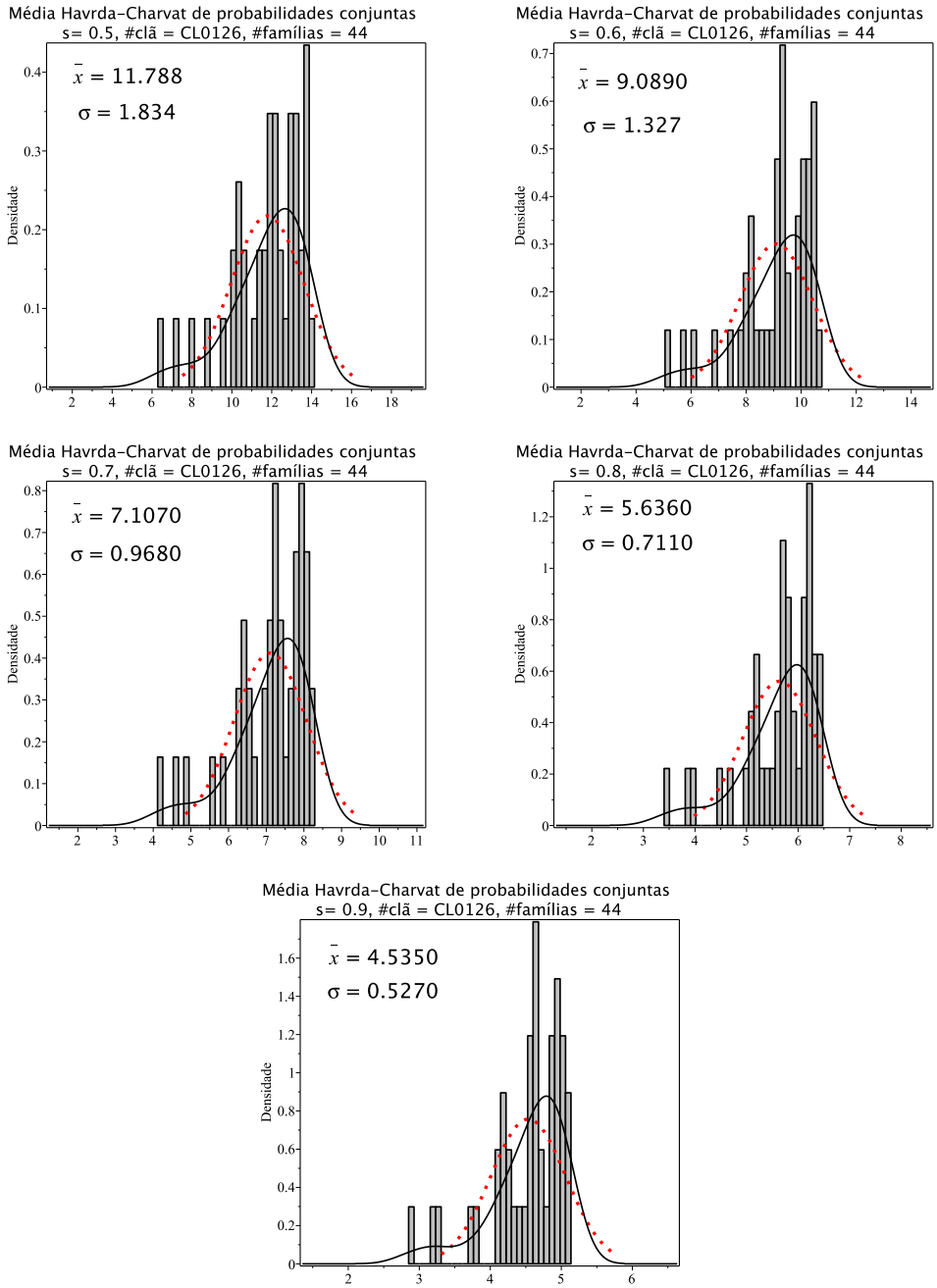

Figura 8.78: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0126. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

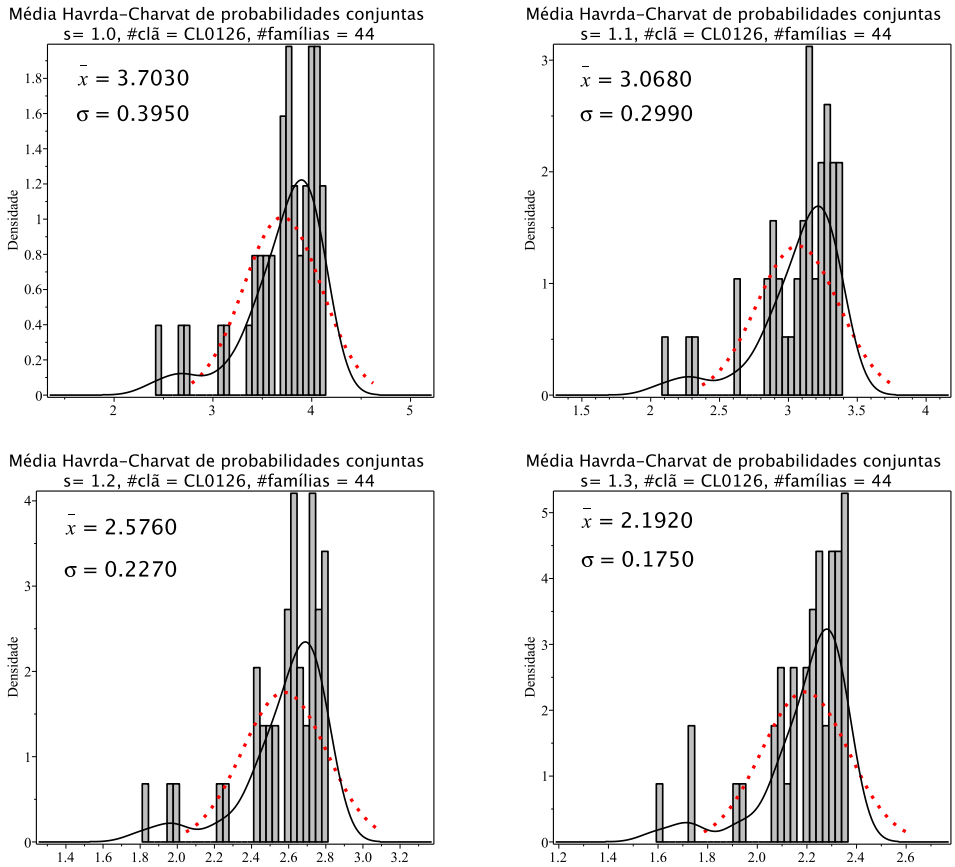

Figura 8.79: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0126. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

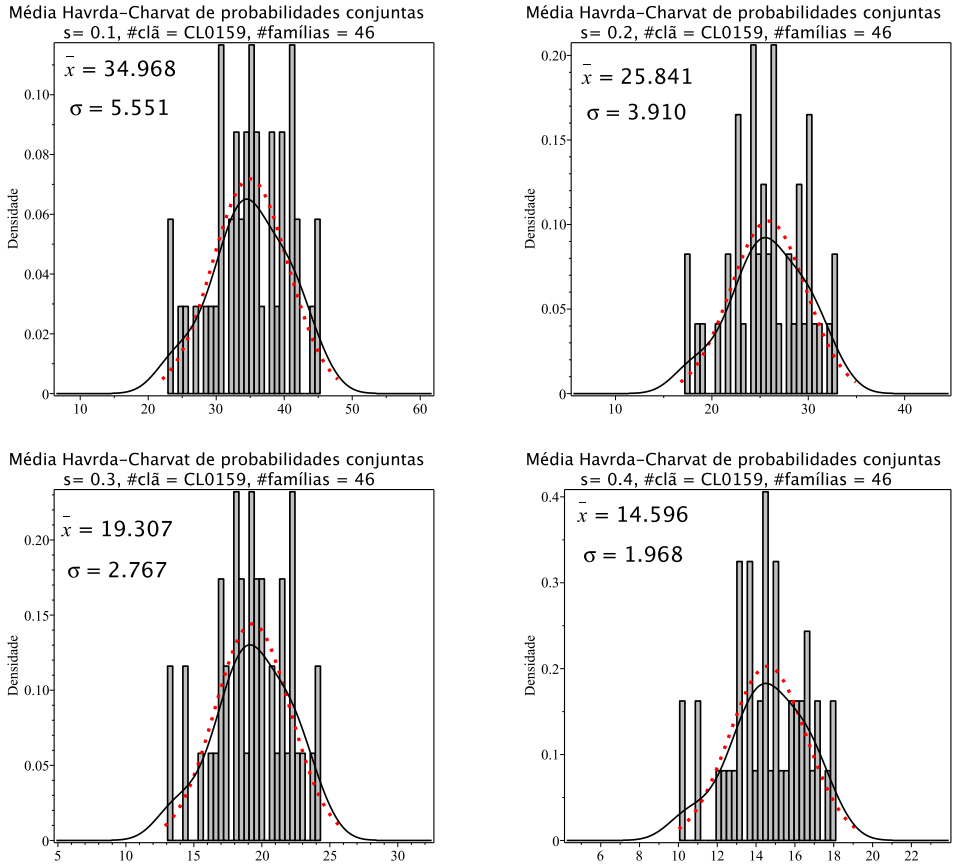

Figura 8.80: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0159. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

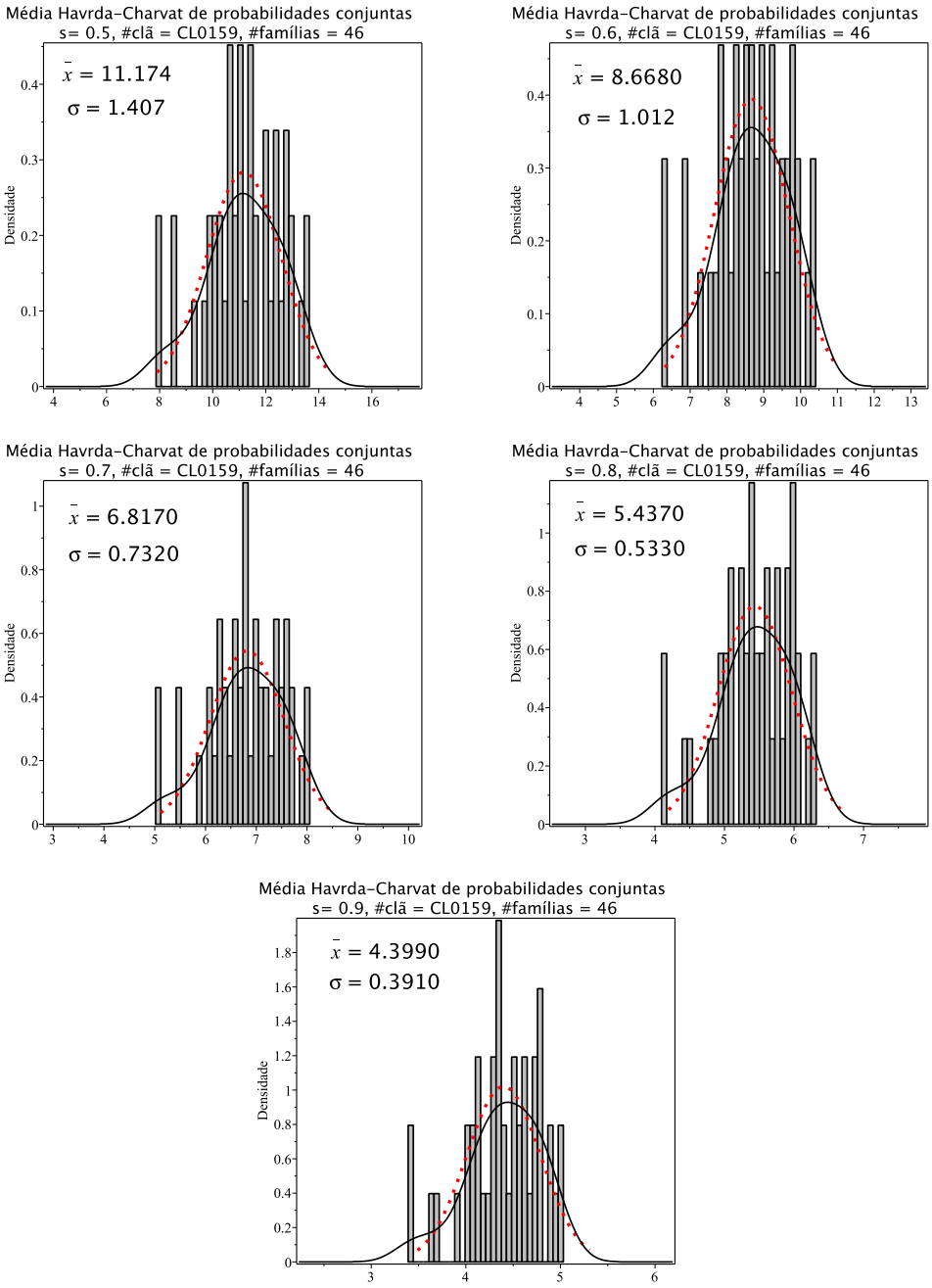

Figura 8.81: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0159. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

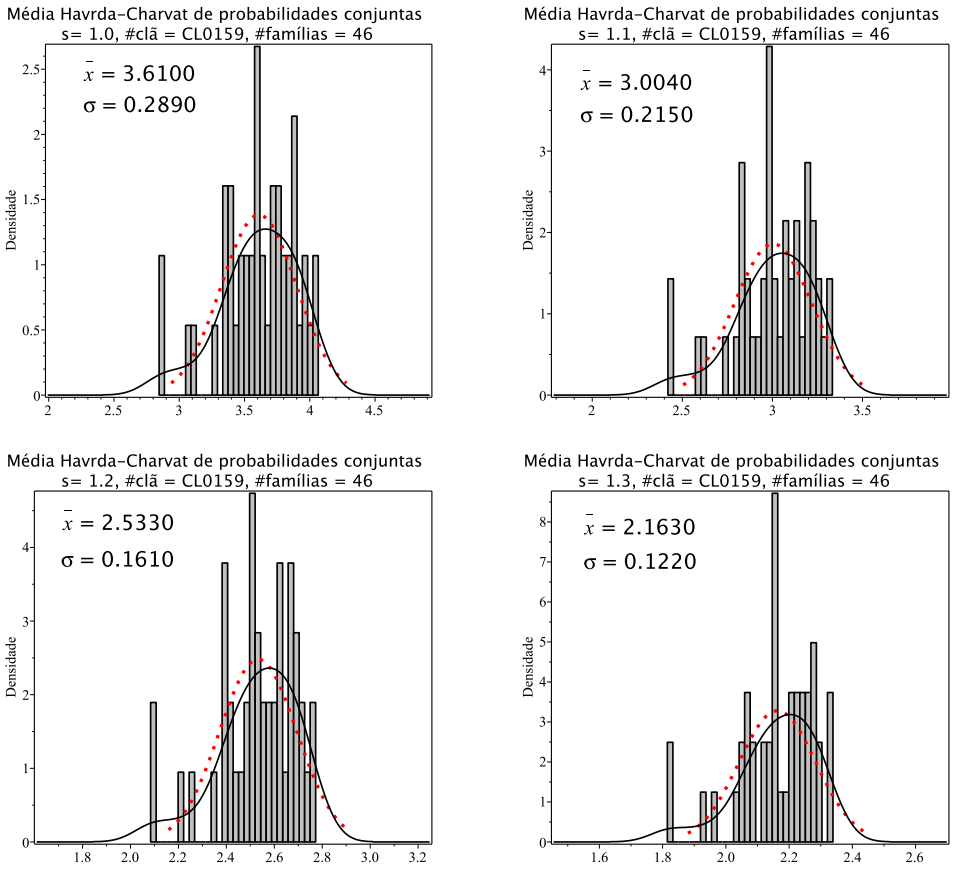

Figura 8.82: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0159. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

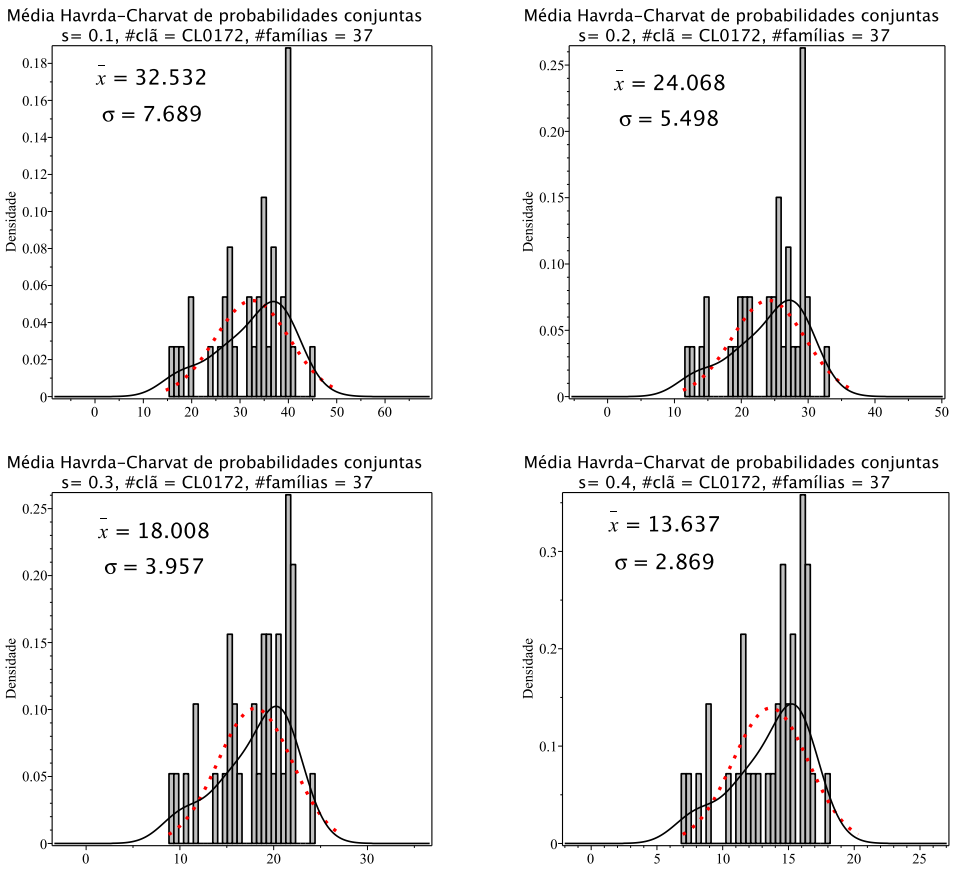

Figura 8.83: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0172. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

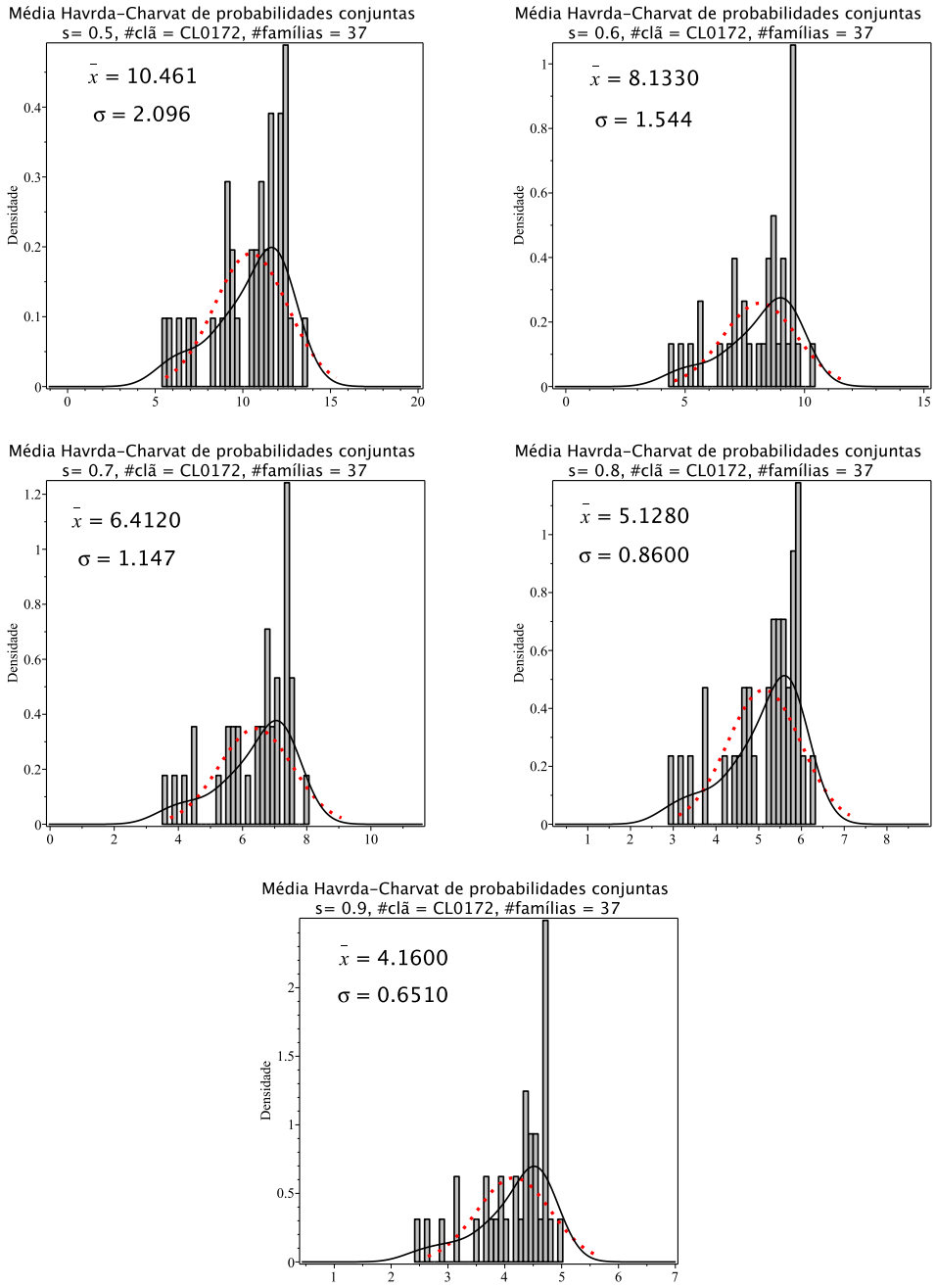

Figura 8.84: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0172. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

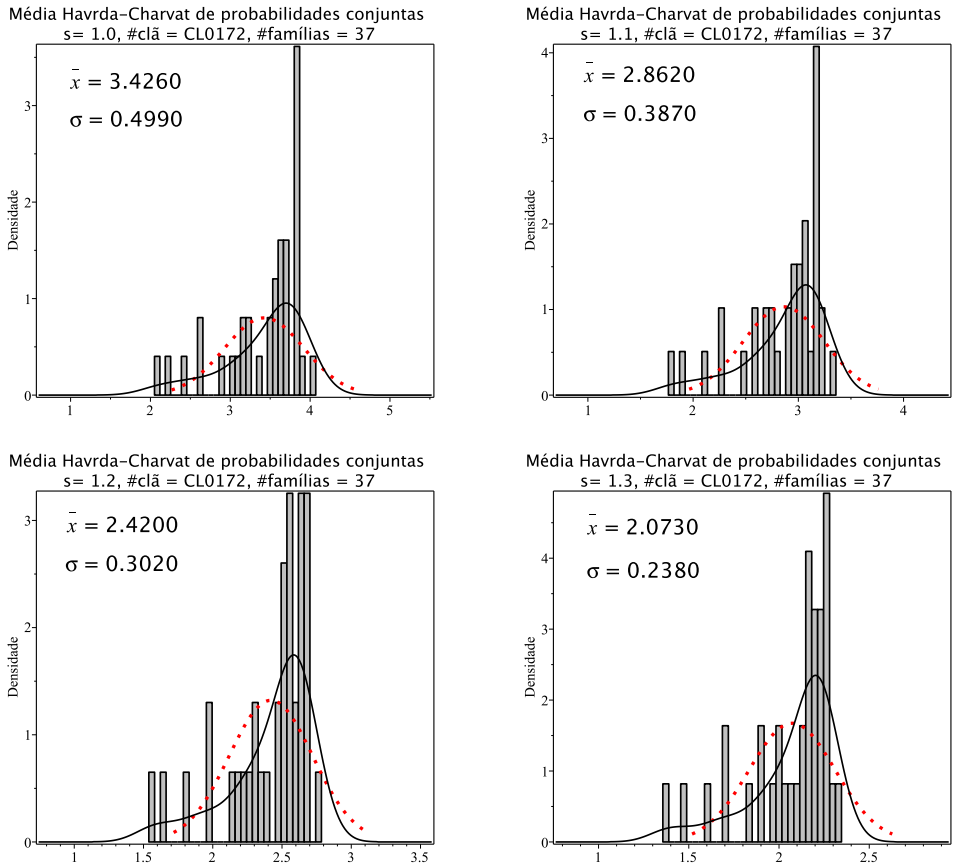

Figura 8.85: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0172. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

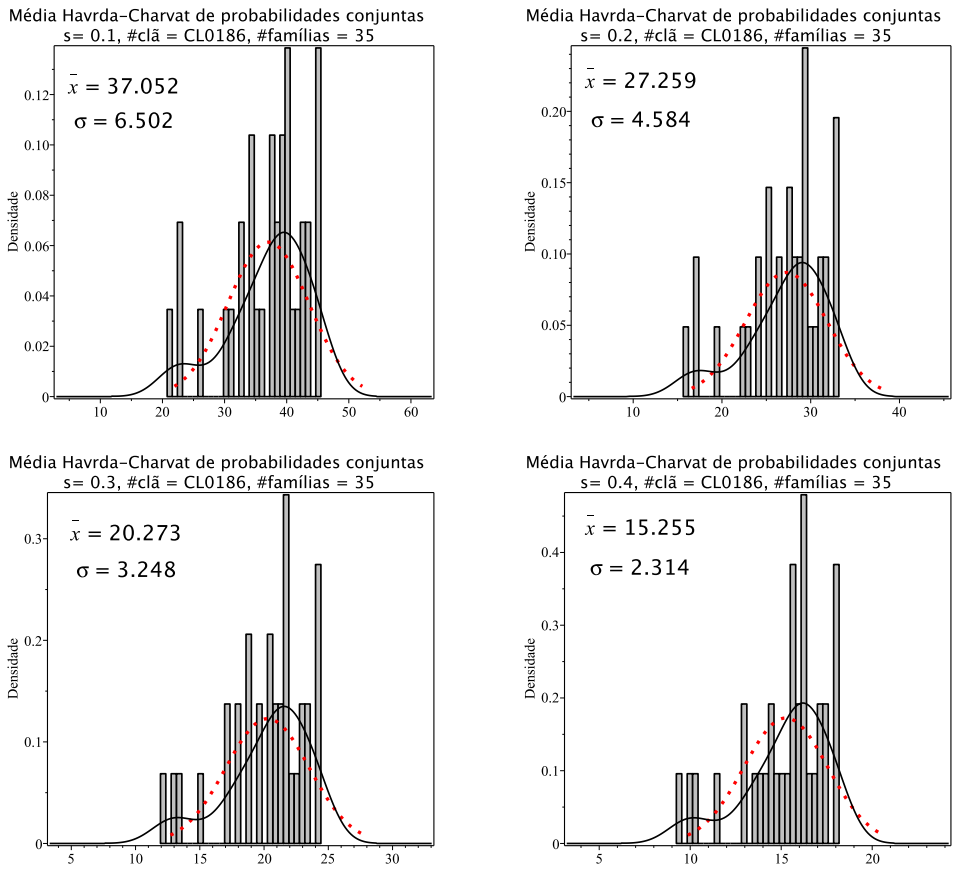

Figura 8.86: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0186. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

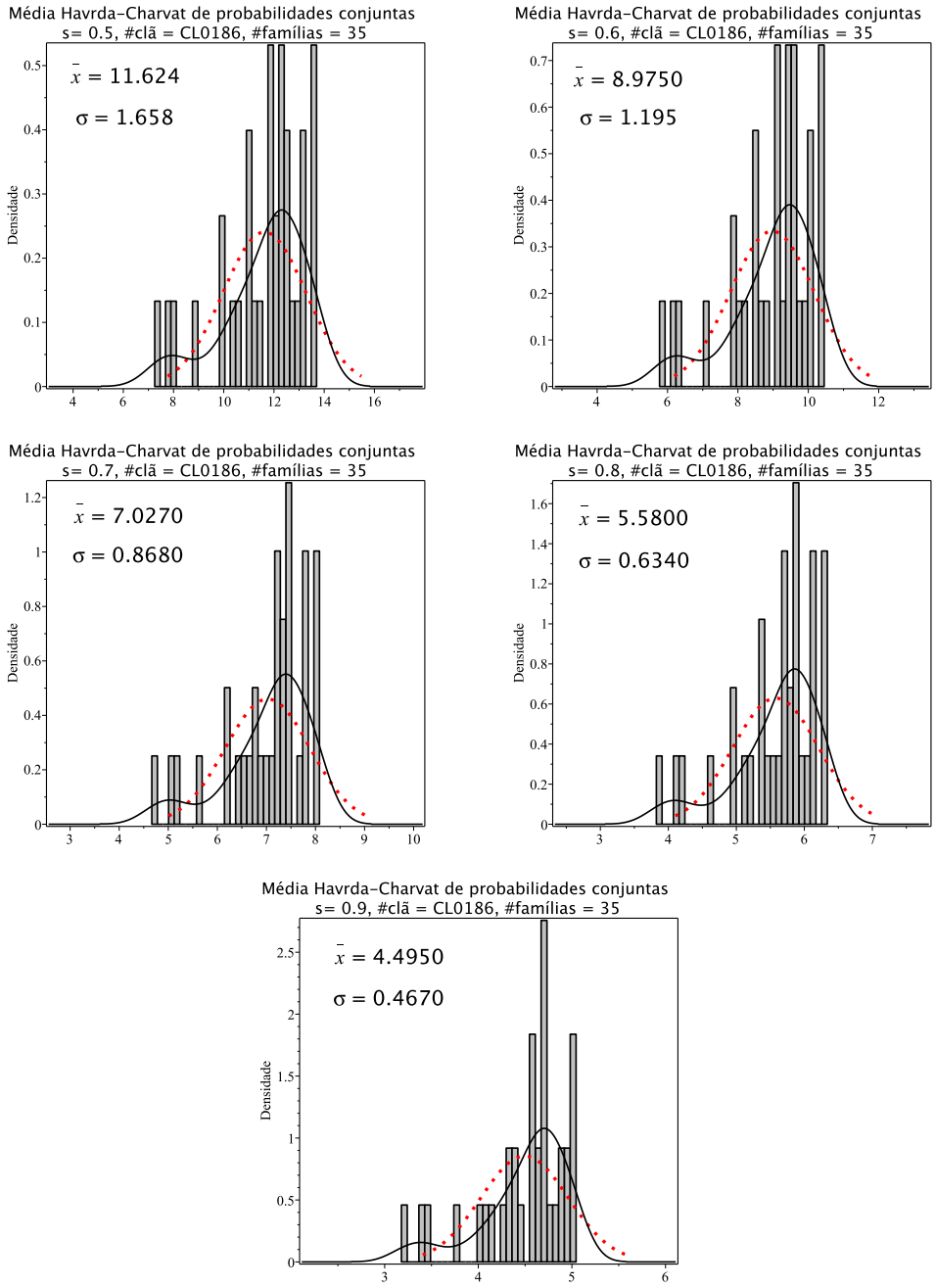

Figura 8.87: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0186. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

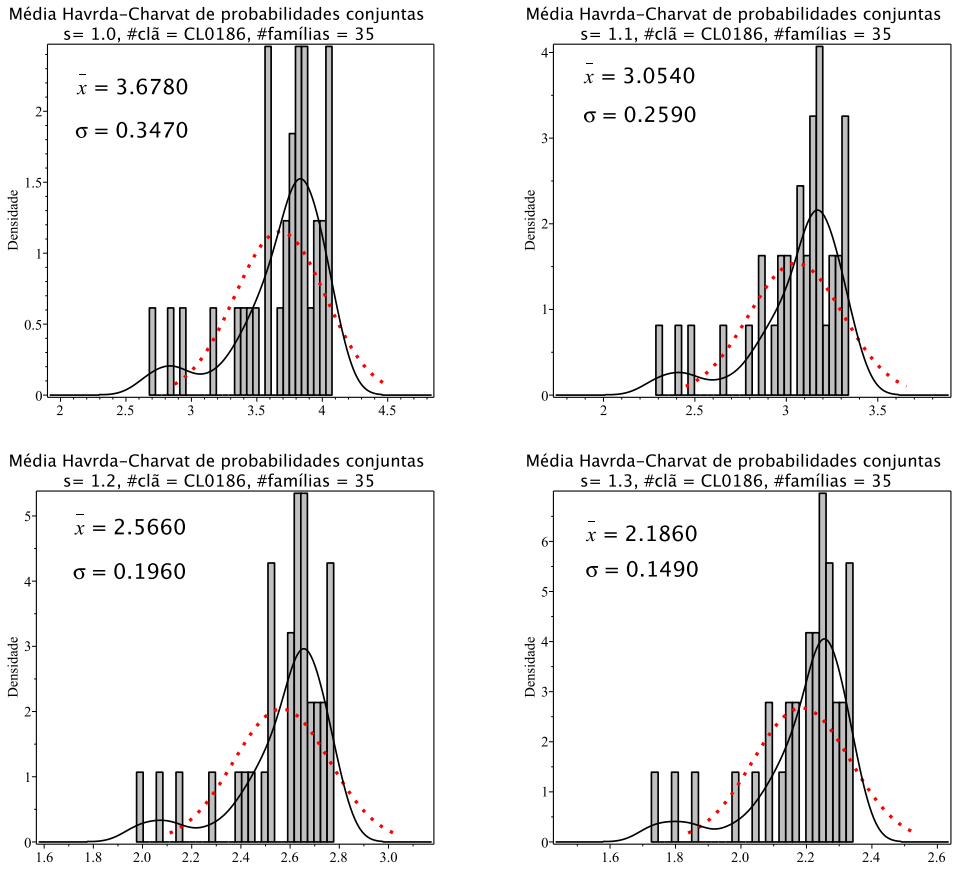

Figura 8.88: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0186. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

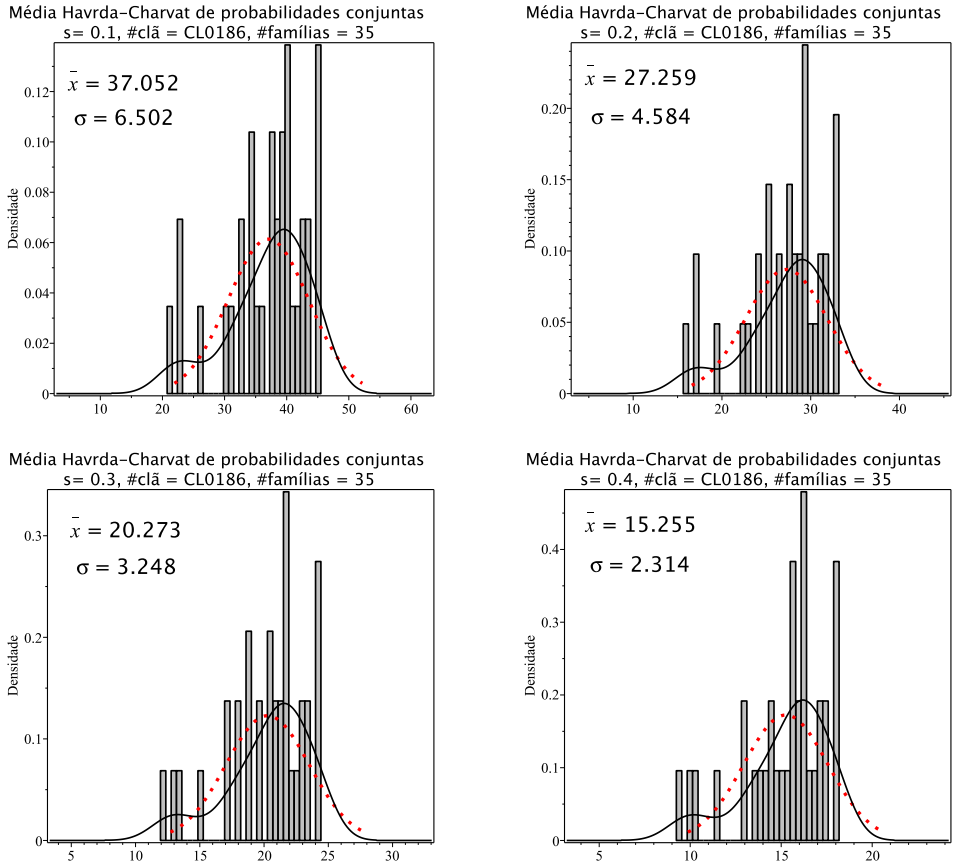

Figura 8.89: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0186. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

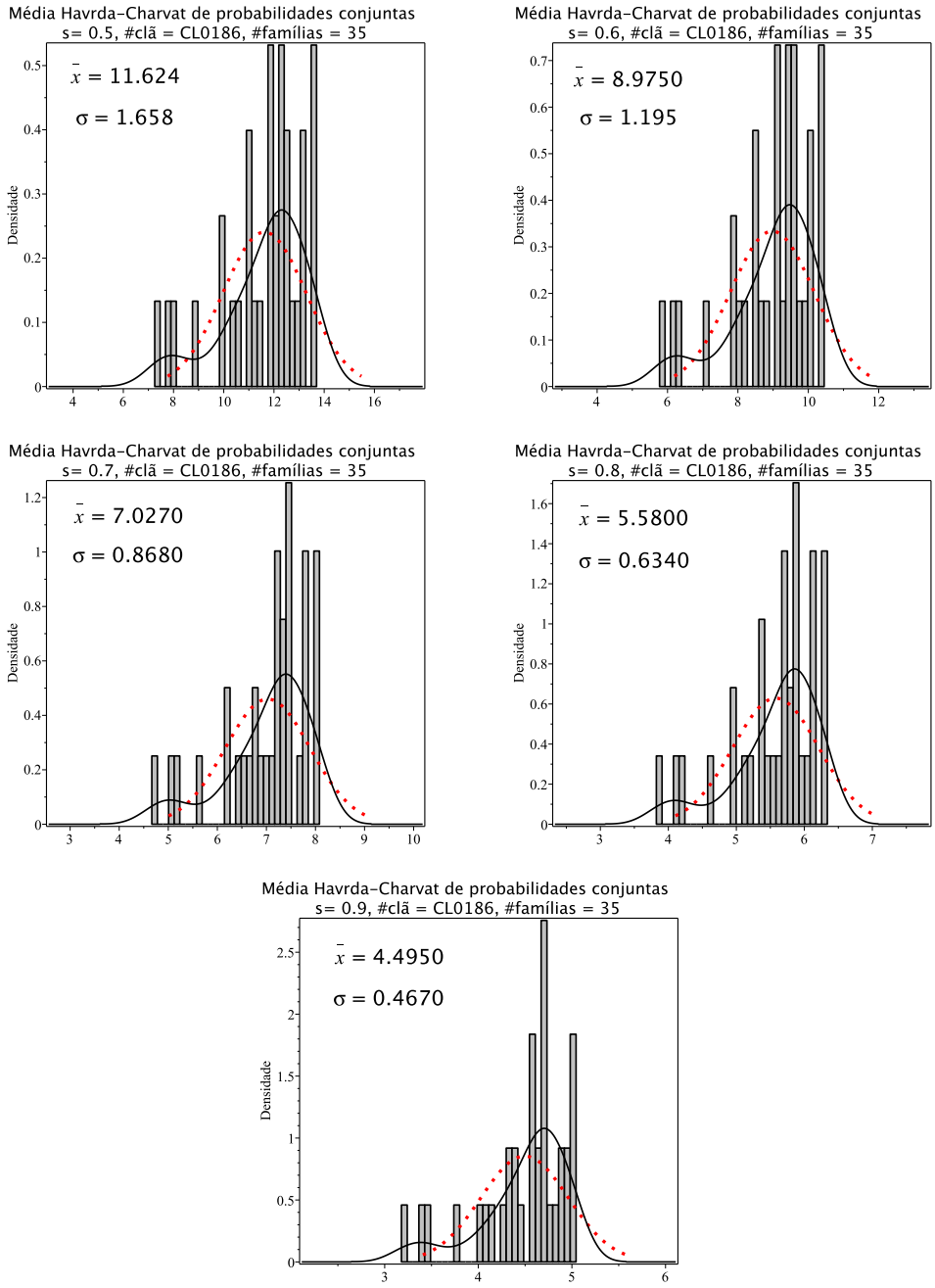

Figura 8.90: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0186. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

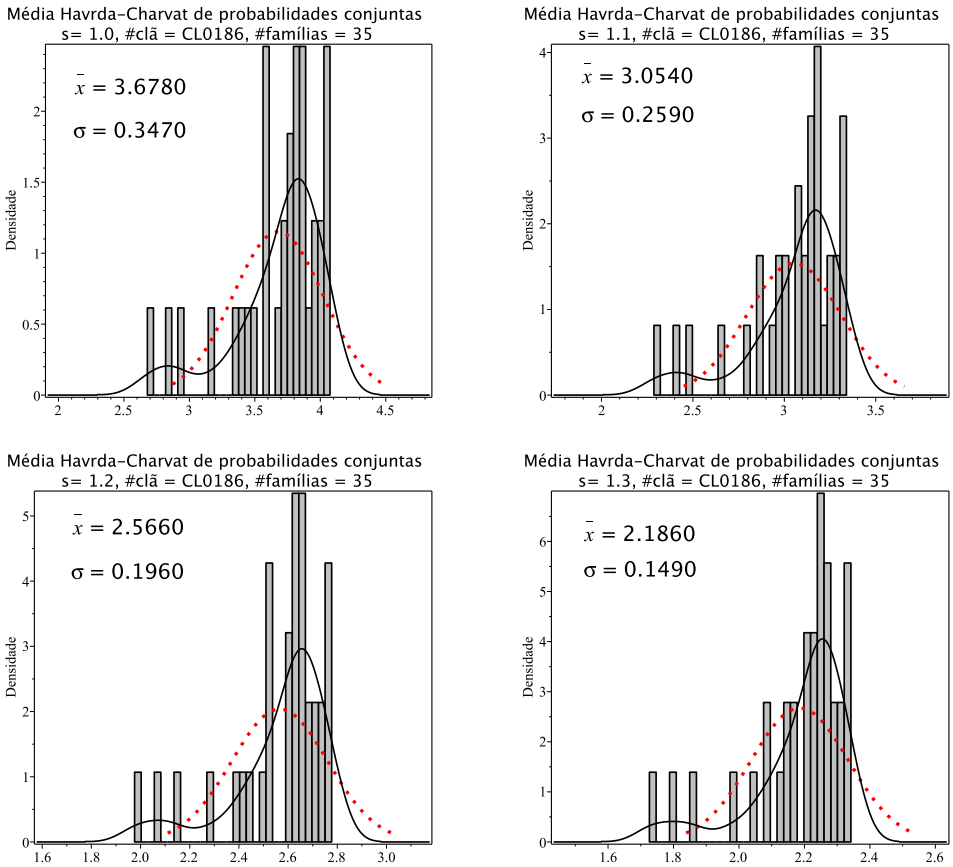

Figura 8.91: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0186. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

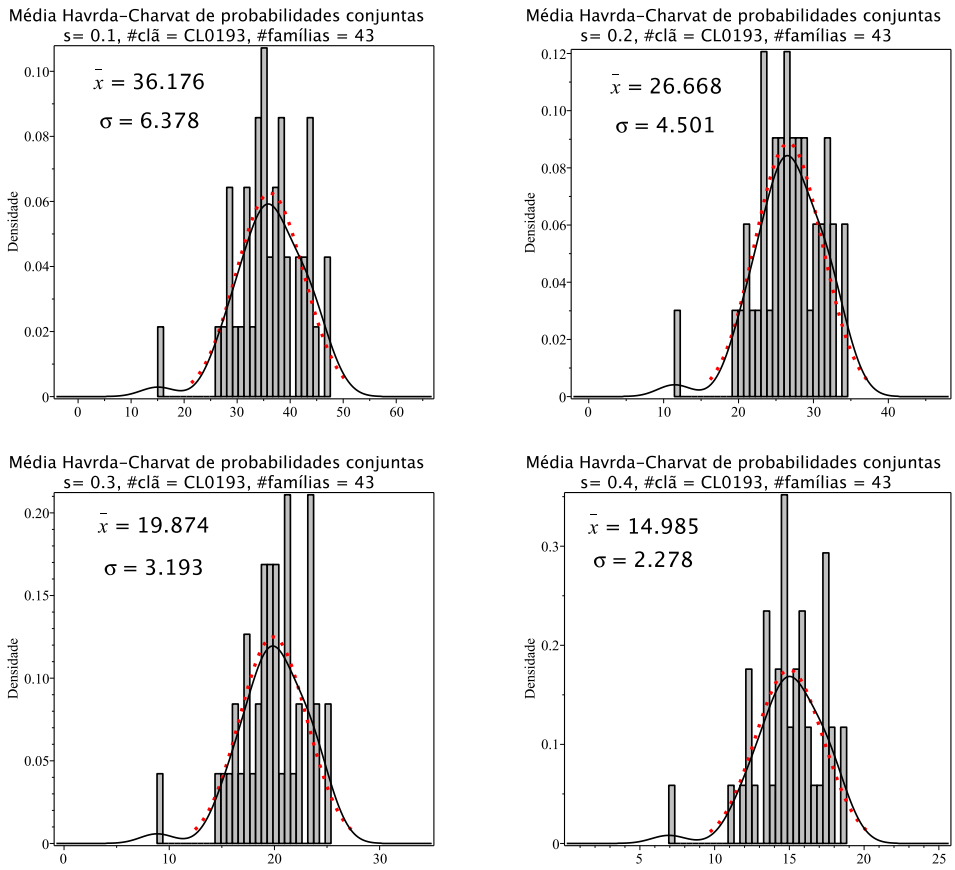

Figura 8.92: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0193. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

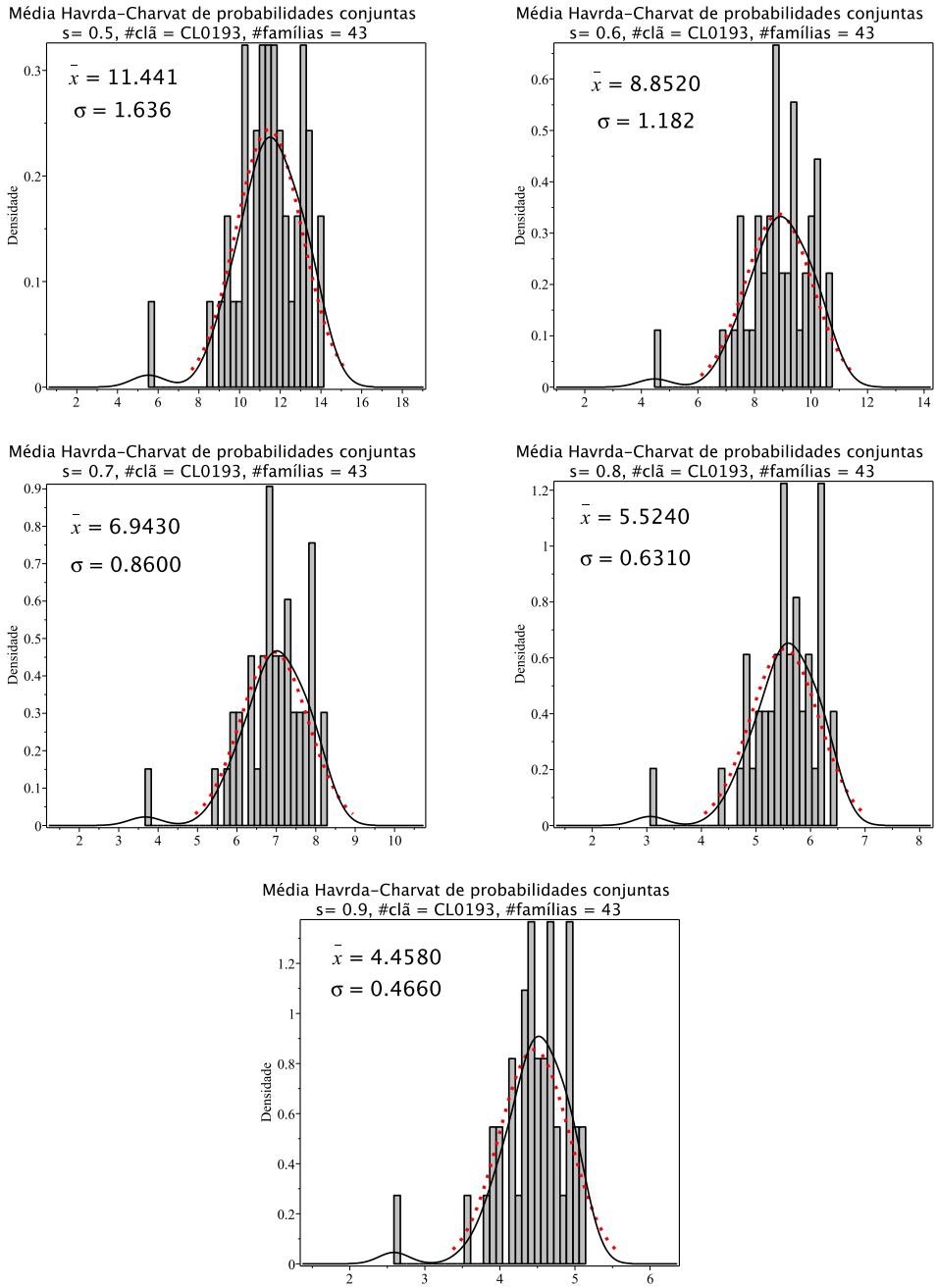

Figura 8.93: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0193. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

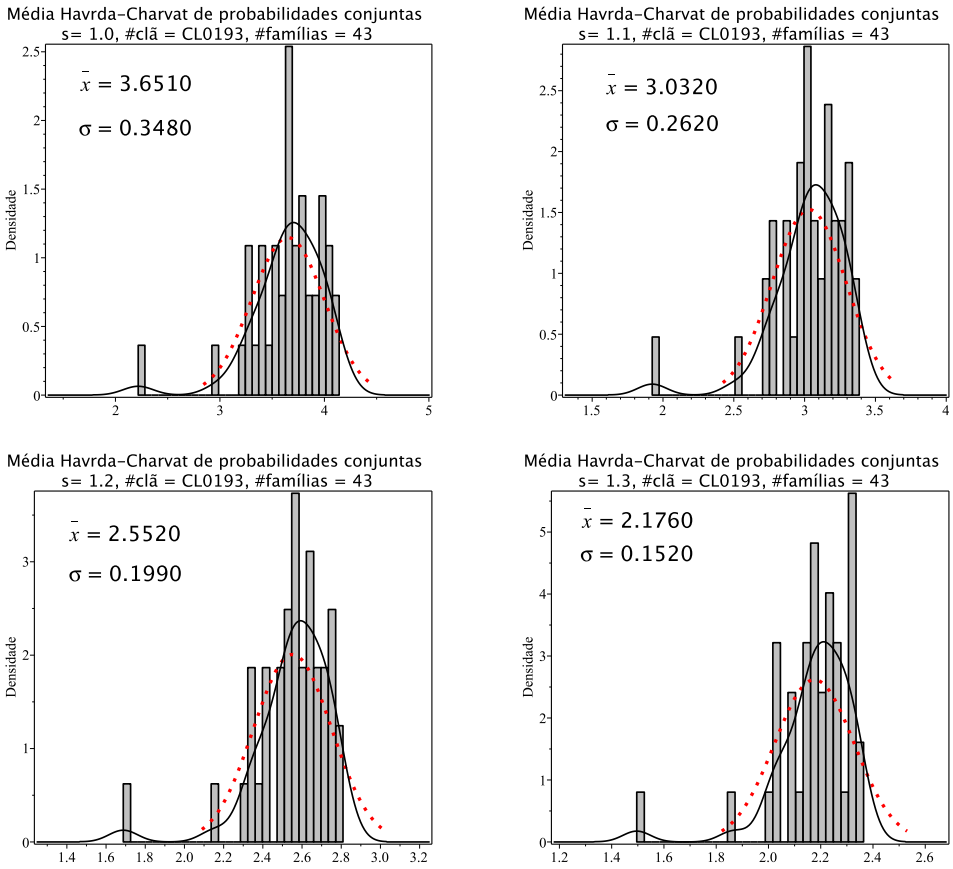

Figura 8.94: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0193. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

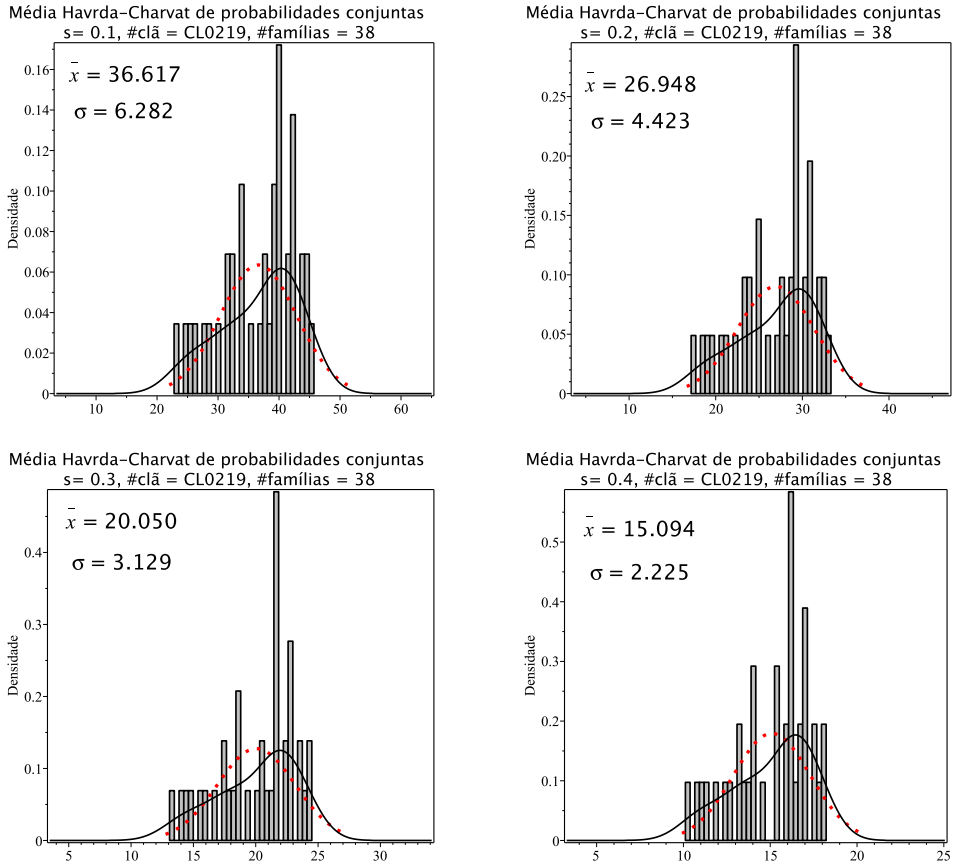

Figura 8.95: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0219. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

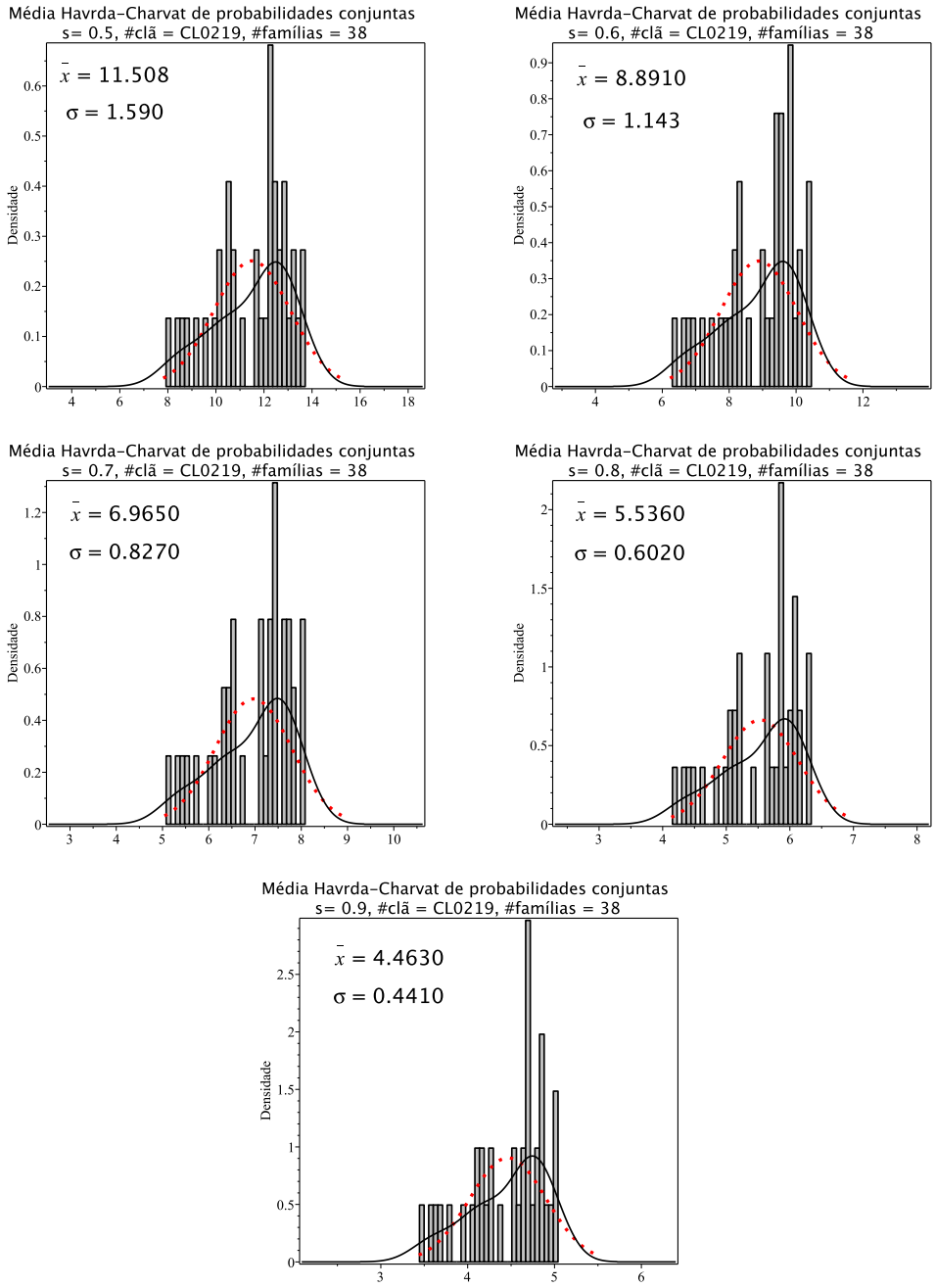

Figura 8.96: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0219. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

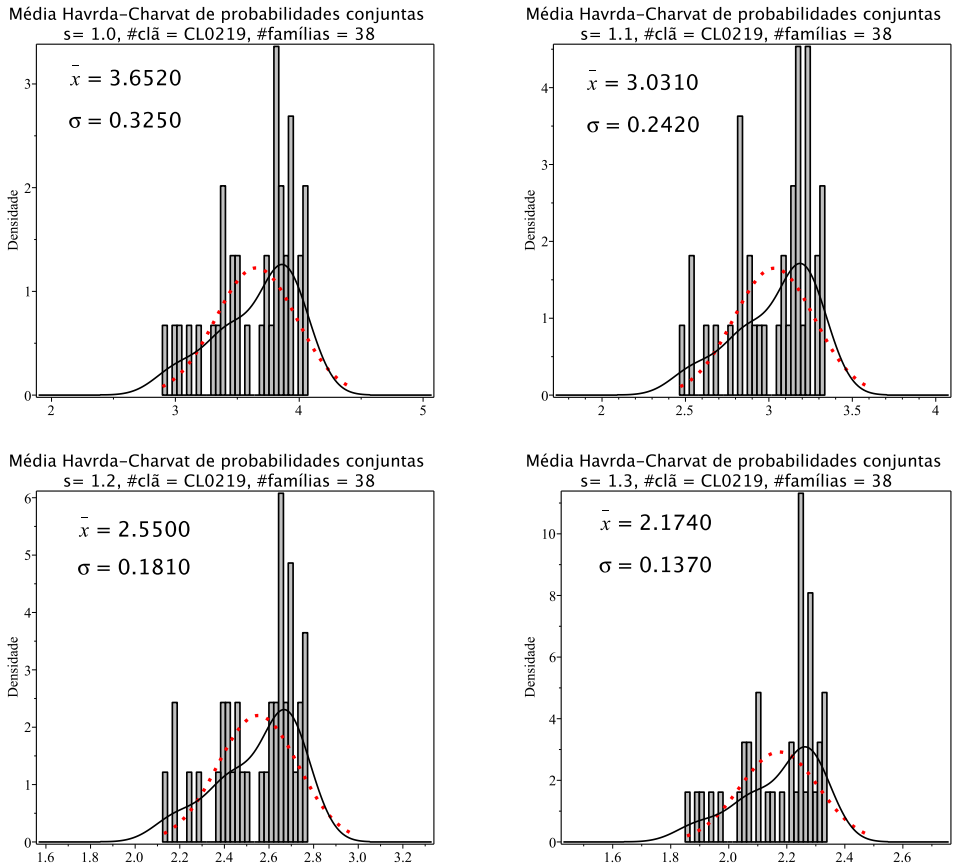

Figura 8.97: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0219. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

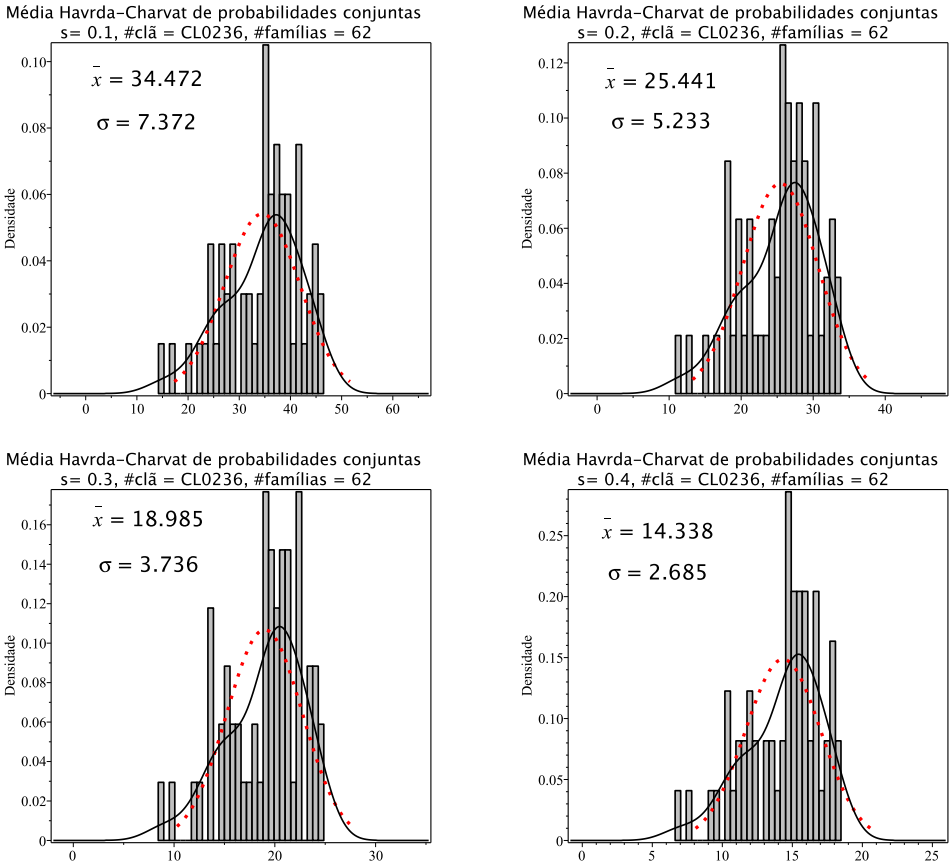

Figura 8.98: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4 do clã CL0236. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

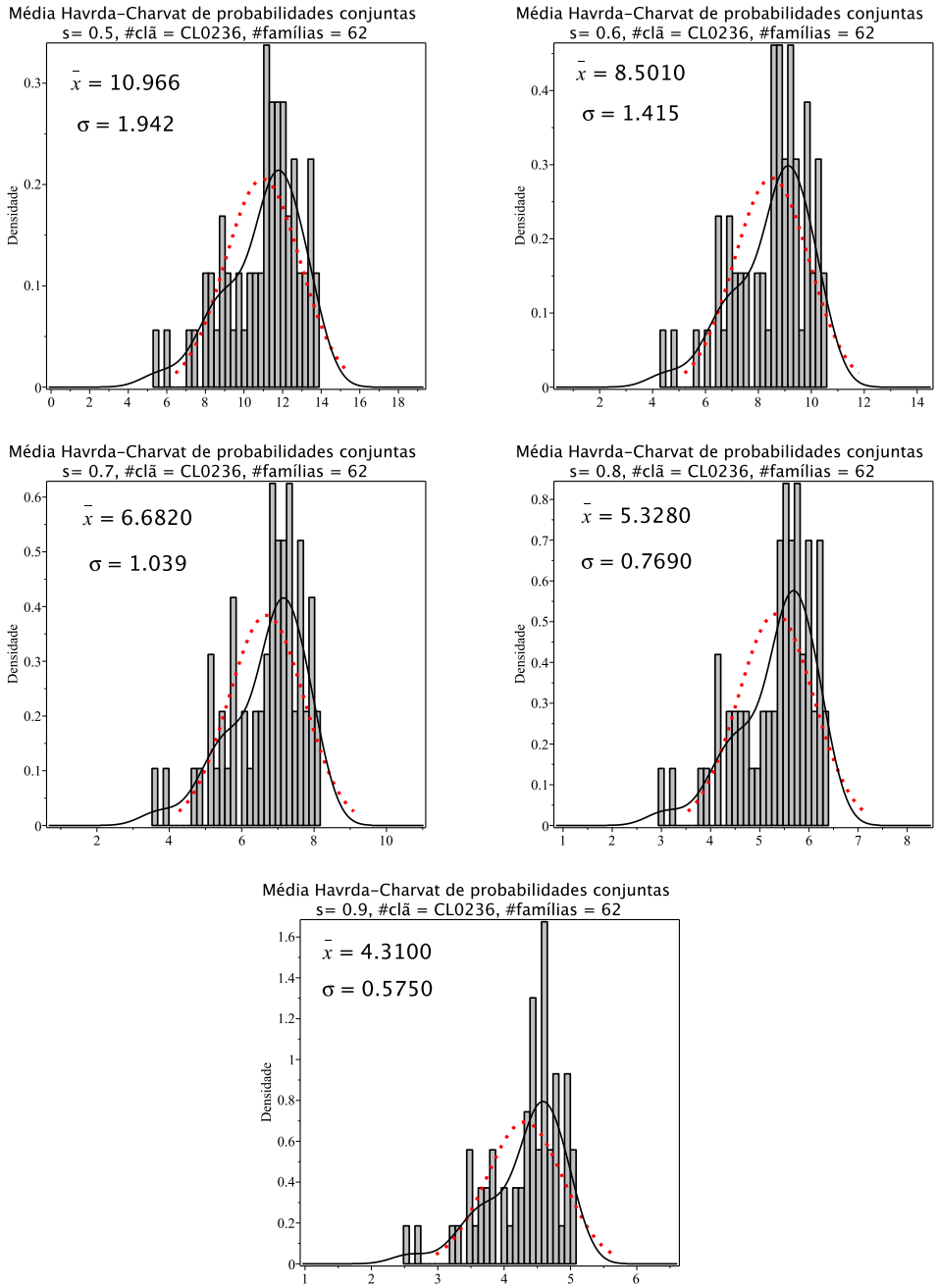

Figura 8.99: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9 do clã CL0236. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

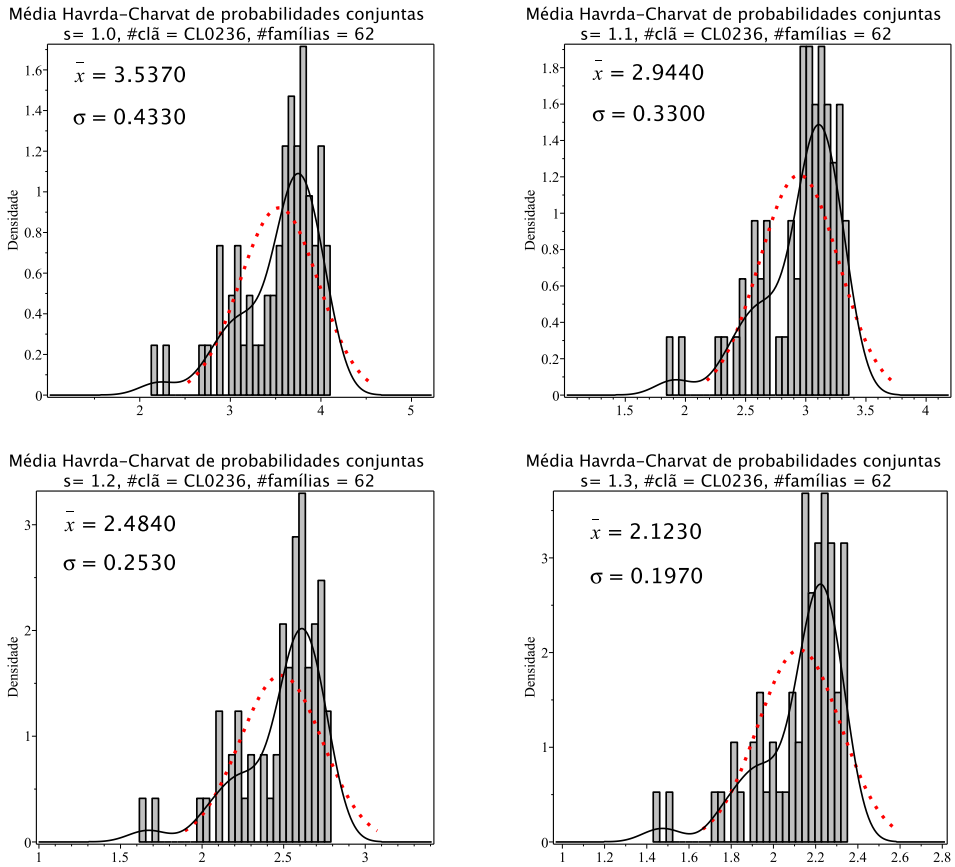

Figura 8.100: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0 (limite Gibbs-Shannon), 1.1, 1.2 e 1.3 do clã CL0236. A curva sólida em preto é a melhor curva ajustada ao histograma. A curva vermelha pontilhada corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

8.4 Símbolo de Jaccard

Existem alguns processos que são determinísticos e que antecedem os cálculos e a geração dos histogramas das médias de Jaccard da entropia de Havrda-Charvat que verificam a existência de algum tipo de relação na ocorrência dos aminoácidos nos pares de colunas e essa relação pode ser medida através do cálculo da Informação Mútua. Para cada uma das entropias apresentadas previamente nesse trabalho, tem-se a Informação Mútua da seguinte forma:

- Sharma-Mittal

$$M_{jk}^{(SM)}(r,s) = \frac{1}{1-r}\left(1 - \left(\frac{\sum_a \sum_b \Big(p_{jk}\big(a,b\big)\Big)^s}{\sum_a \sum_b \Big(p_j\big(a\big)p_k\big(b\big)\Big)^s}\right)^{\frac{1-r}{1-s}}\right) \tag{8.4}$$

- Havrda-Charvat

$$M_{jk}^{(HC)}(s) = \lim_{r\to s} M_{jk}^{(HC)}(r,s) = \frac{1}{1-s}\left(1 - \frac{\sum_a \sum_b \Big(p_{jk}\big(a,b\big)\Big)^s}{\sum_a \sum_b \Big(p_j\big(a\big)p_k\big(b\big)\Big)^s}\right) \tag{8.5}$$

O cálculo da distância de Informação é dado pela diferença entre a o valor de entropia de um par de colunas e o valor informação mútua do mesmo par:

$$d_{jk}(r,s) = (SM)_{jk}(r,s) - M_{jk}^{(SM)}(r,s), \tag{8.6}$$

onde,

$$\begin{aligned} (SM)_{jk}(r,s) &\geq 0; \\ (M)_{jk}^{(SM)}(r,s) &\geq 0; \\ (SM)_{jk}(r,s) - (M)_{jk}^{(SM)}(r,s) &\geq 0. \end{aligned} \tag{8.7}$$

A entropia de Jaccard [60] de Havrda-Charvat é obtida pela normalização da distância de Informação. Temos, então:

$$J_{jk}^{(HC)}(s) = 1 - \frac{M_{jk}^{(SM)}(s)}{(HC)_{jk}(s)}, \tag{8.8}$$

Das equações (8.7) e (8.8) temos: $0 \geq J_{jk}^{(HC)}(r,s) \geq 1$.

A média da entropia de Jaccard de Havrda-Charvat para todos os pares de colunas do bloco (80×80) é calculada da seguinte forma:

$$J_{jk}^{(HC)}(s) = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} J_{jk}^{(HC)}(s)$$
$$J_{jk}^{(HC)}(s) = \frac{2}{80(80-1)} \sum_{j=1}^{79} \sum_{k=j+1}^{80} J_{jk}^{(HC)}(s) \tag{8.9}$$
$$J_{jk}^{(HC)}(s) = \frac{2}{3160} \sum_{j=1}^{79} \sum_{k=j+1}^{80} J_{jk}^{(HC)}(s)$$

As Figuras (8.101), (8.102) e (8.103) representam os histogramas de Jaccard associado aos valores de parâmetro s, para o caso com o bloco representativo (80×80), respectivamente. Todos os histogramas são de densidade, o que demonstra que a área de cada barra é proporcional ao número de elementos (valores de entropia dos pares de colunas) presentes no intervalo de valores (largura da barra), de forma que a área total (a soma das áreas de todas as barras) é igual a 1. Logo, com as devidas aproximações e suavizações, obtemos uma curva que se ajusta à distribuição dos dados. As curvas ajustadas aos histogramas são apresentadas como linhas em vermelho, enquanto as linhas em azul representam as curvas gaussianas construídas com a média e o desvio padrão das distribuições da entropia.

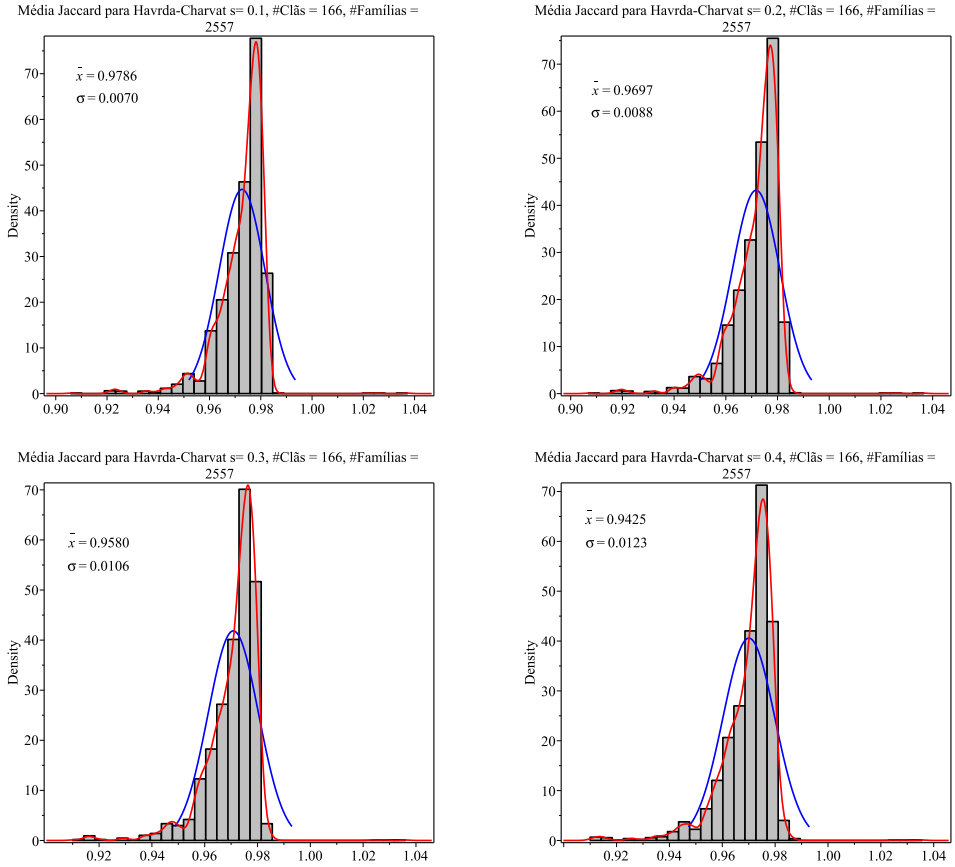

Figura 8.101: Histogramas de densidade das médias de Jaccard da entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.1, 0.2, 0.3 e 0.4. A curva em vermelho é a melhor curva ajustada ao histograma. A curva azul corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

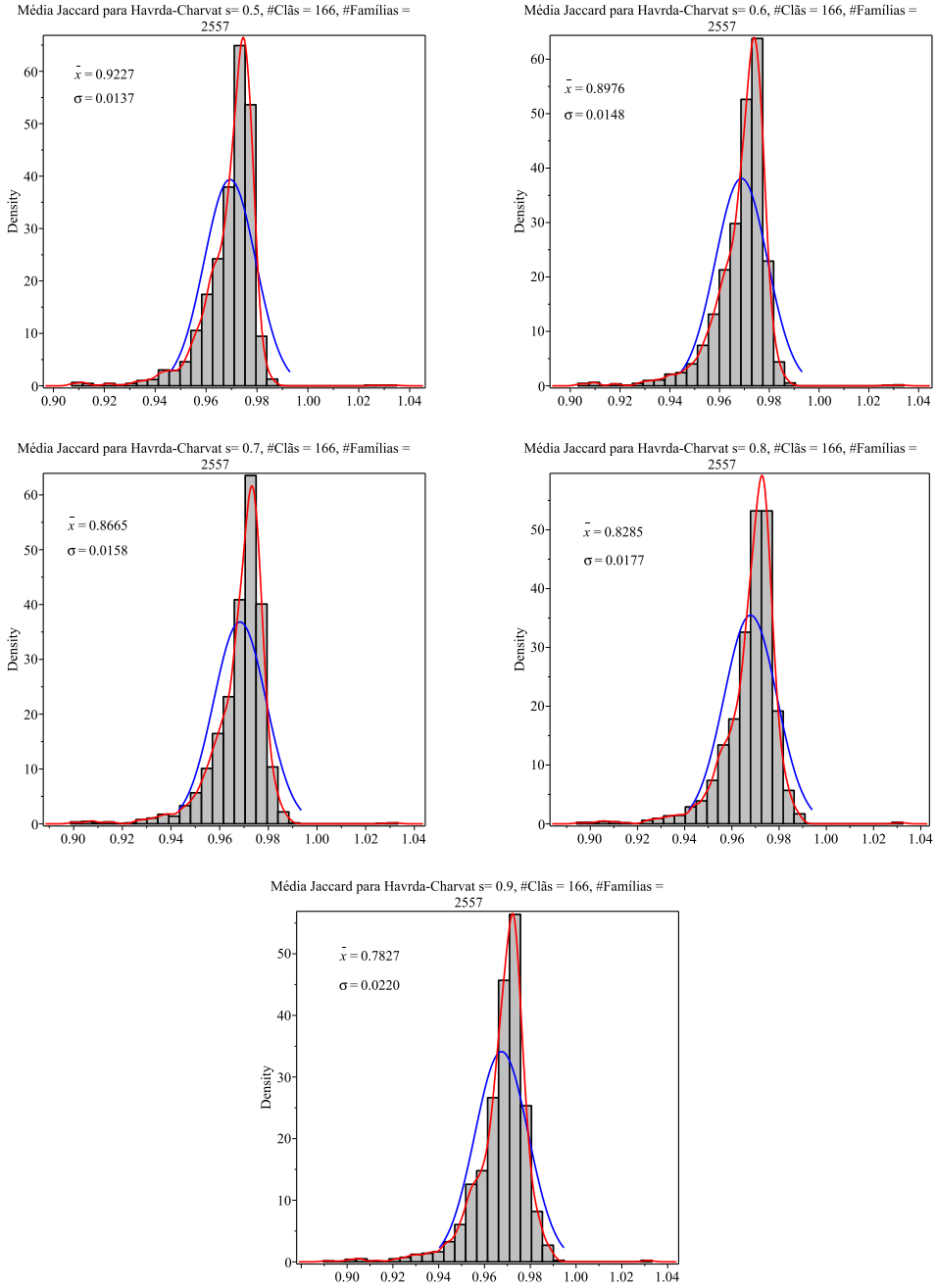

Figura 8.102: Histogramas de densidade das médias de Jaccard da entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 0.5, 0.6, 0.7, 0.8 e 0.9. A curva em vermelho é a melhor curva ajustada ao histograma. A curva azul corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

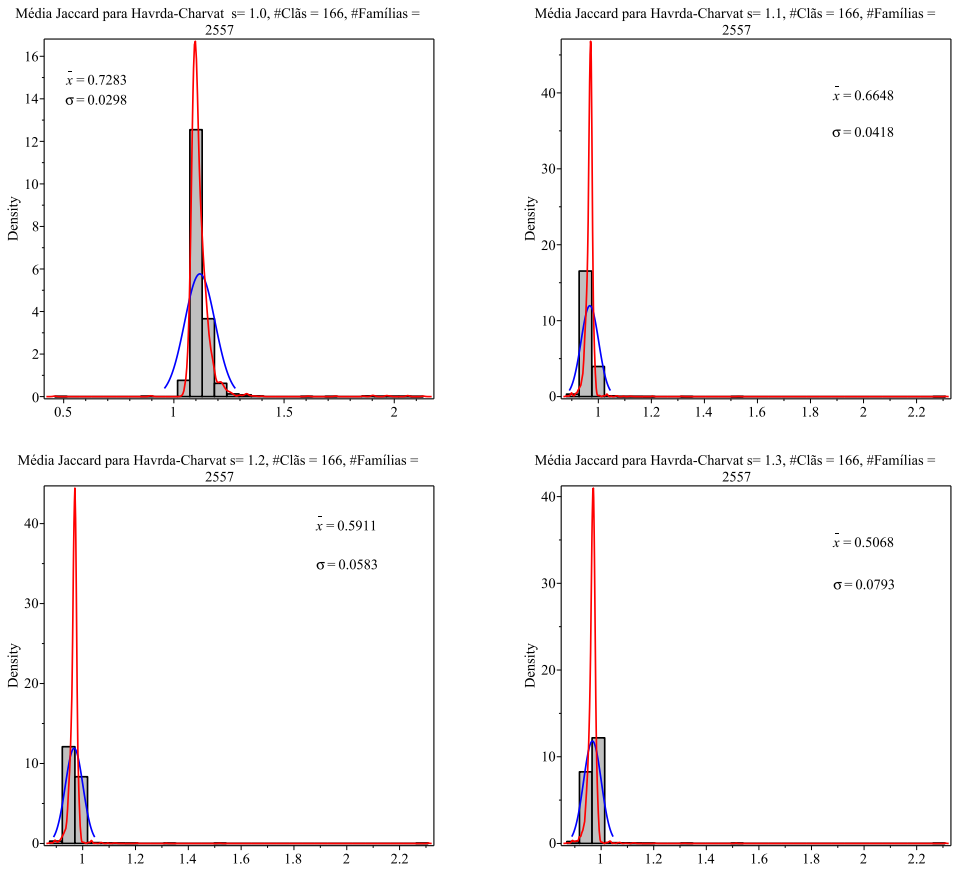

Figura 8.103: Histogramas de densidade das médias de entropia Havrda-Charvat de probabilidade conjunta para o parâmetro s igual 1.0(Limite Gibbs-Shannon), 1.1, 1.2 e 1.3. A curva em vermelho é a melhor curva ajustada ao histograma. A curva azul corresponde à curva gaussiana construída com o valor médio e o desvio padrão da distribuição das médias de entropia.

É interessante observar o comportamento dos histogramas em relação aos valores do parâmetro s (Figuras 8.101, 8.102 e 8.103). Nota-se que para os valores do parâmetro s igual a 1.0, 1.1, 1.2 e 1.3 há algum tipo de anomalia, que faz os histogramas bastante diferentes dos histogramas dos demais parâmetros.

8.5 Revisão bibliográfica

As famílias de domínios são grupos de domínios de proteínas que compartilham semelhanças estruturais e funcionais. A caracterização dessas famílias envolvem alguns processos como identificação e classificação de seus membros com base em vários critérios, como similaridade de sequência, estrutura, função e relações evolutivas. Os domínios de proteínas são as unidades básicas de proteínas que podem dobrar, funcionar e evoluir de forma independente. O conhecimento dos domínios de proteínas é fundamental para a classificação de proteínas, compreendendo suas

funções biológicas, analizando os seus mecanismos evolutivos juntamente com o design das proteínas. Assim, nas duas últimas décadas, várias abordagens de identificação de domínios de proteínas foram desenvolvidas e vários bancos de dados de domínios de proteínas também foram construídos. Existem vários métodos para caracterizar famílias de domínios de proteínas. Abaixo estão listados alguns destes métodos mais comumente utilizados [78].

Os métodos baseados em sequência são métodos que usam similaridade de sequência para identificar famílias de domínio. Uma abordagem comum é usar algoritmos de alinhamento de sequência para comparar as sequências de aminoácidos de diferentes proteínas e identificar regiões conservadas. Essas regiões conservadas provavelmente contêm domínios funcionais. Por exemplo, o banco de dados Pfam usa uma abordagem baseada em modelo oculto de Markov (HMM) para identificar domínios conservados em sequências de proteínas. Exemplos de tais métodos incluem BLAST (Basic Local Alignment Search Tool) [79] e HMMER (Hidden Markov Model-based search tool) [80].

Os métodos baseados em estrutura usam a estrutura 3D de proteínas para identificar famílias de domínio. Uma abordagem é utilizar algoritmos de alinhamento estrutural para comparar as estruturas de diferentes proteínas e identificar motivos estruturais comuns. Outra abordagem é usar algoritmos de agrupamento para agrupar proteínas com base em suas semelhanças estruturais ou similaridade estruturais para agruparem as proteínas em famílias. Exemplos de tais métodos incluem DaliLite (Alinhamento de matriz de distância com locais de ligação de ligante), CE (extensão combinatória) e TM-align (Alinhamento de modelagem baseado em modelo) [81]. Por exemplo, os bancos de dados SCOP e CATH usam abordagens baseadas em estrutura para classificar proteínas em diferentes dobras e famílias de domínio [82].

Os métodos baseados em função utilizam a similaridade funcional para agrupar proteínas em famílias. Exemplos de tais métodos incluem InterProScan (Uma ferramenta que combina vários métodos para identificar os domínios funcionais)[83] e PfamScan (Uma ferramenta que resposável por buscar os domínios de proteínas conservadas em um banco de dados de sequência)[84].

Métodos híbridos combinam abordagens baseadas em sequência, estrutura e função para identificar famílias de domínio. Por exemplo, alguns métodos usam uma combinação de algoritmos de alinhamento de sequência e alinhamento estrutural para identificar regiões conservadas com estruturas e funções semelhantes.

Em geral, a escolha do método para caracterizar as famílias de domínio depende dos dados disponíveis e da questão de pesquisa que está sendo abordada. Os métodos baseados em sequência são frequentemente usados para realizar análises em larga escala, enquanto os métodos baseados em estrutura são mais apropriados para análises detalhadas de estruturas de proteínas. Os métodos baseados em função podem fornecer informações sobre os papéis biológicos das famílias de domínio, enquanto os métodos híbridos podem fornecer uma caracterização mais abrangente das famílias de domínio.

Os Bancos de dados de proteínas são coleções de informações sobre proteínas, suas sequências, estruturas e funções. Esses bancos de dados desempenham um papel crucial na pesquisa de bioinformática sendo usados para analisar, comparar e anotar dados de proteínas. Nos parágrafos estão alguns dos bancos de dados de proteínas mais populares.

O RCSB Ligand Explorer é um banco de dados de pequenas moléculas e suas interações com proteínas. Ele fornece informações sobre os locais de ligação, afinidades de ligação e propriedades estruturais dos ligantes o acesso pode ser realizado através do site: http://ligand-expo.rcsb.org/ [85].

O Protein Information Resource (PIR) é um recurso abrangente para a sequência de proteínas e dados estruturais. Ele fornece anotações detalhadas e informações funcionais para uma ampla gama de proteínas pode ser acesso pelo site: https://proteininformationresource.org/ [86].

O repositório SWISS-MODEL é um banco de dados de modelos de estruturas de proteínas gerados usando o pipeline de modelagem de homologia SWISS-MODEL. Ele fornece modelos de proteínas de alta qualidade que não foram determinadas experimentalmente para todas as sequência no UniProtKB. O SWISS-MODEL, pode acessado pelo site: https://swissmodel.expasy.org/repository

O InterPro é um banco de dados de famílias de proteínas, domínios e locais funcionais. Ele integra informações de vários bancos de dados de proteínas como o Pfam, PROSITE, PRINTS, ProDom e SMART, entre outros bancos, também pode prever a função da proteína com base na análise de sequência, o acesso pode ser através do site: https://www.ebi.ac.uk/interpro/ [87].

O PROSITE é um banco de dados de famílias de proteínas, domínios e locais funcionais. As entradas do PROSITE incluem uma descrição do domínio ou local,

uma matriz de perfil que pode ser usada para pesquisar o domínio ou local em outras proteínas e um conjunto de exemplos anotados de proteínas que contêm o domínio ou local[88].

O banco de dados de domínio conservado (CDD) é um banco de dados de famílias e domínios de proteínas que é construído usando vários métodos HMM de perfil e alinhamento de sequência. O CDD inclui domínios de várias fontes, incluindo Pfam, SMART, COG e outros [89].

SUPERFAMILY é um banco de dados que classifica as proteínas em superfamílias estruturais e evolutivas com base na presença de domínios e seu arranjo. As entradas do SUPERFAMILY incluem informações sobre a arquitetura do domínio, relacionamentos evolutivos e anotações funcionais [90].

O Gene Ontology (GO) é um banco de dados de anotações padronizadas para produtos de genes, incluindo proteínas. Ele fornece um vocabulário para descrever funções de proteínas, localizações celulares e processos biológicos, acesso é realizado pelo site: http://geneontology.org/ [91].

O Gene3D é um banco de dados que fornece anotações estruturais e funcionais para domínios de proteínas. Ele usa uma combinação de sequência e informações estruturais para classificar as proteínas em diferentes famílias e superfamílias [92].

Esses bancos de dados são recursos úteis para pesquisadores que estudam a estrutura, função e a evolução das proteínas. Eles podem ser usados para identificar funções potenciais de proteínas com base em seu conteúdo de domínio, para prever os efeitos de mutações na função da proteína e para projetar experimentos para testar hipóteses sobre a estrutura e função da proteína. Cada banco de dados tem suas particularidades, e os pesquisadores costumam usar vários bancos de dados em suas análises para obter uma visão abrangente dos dados de proteínas.

Referências Bibliográficas

[1] MONDAINI, R. P., SIMÃO, C. "Khinchin–Shannon Generalized Inequalities for "Non-additive" Entropy Measures", *Trends in Biomathematics: Mathematical Modeling for Health, Harvesting, and Population Dynamics*, pp. 177–190, 2019.

[2] MONDAINI, R., DE ALBUQUERQUE NETO, S. "The Statistical Analysis of Protein Domain Family Distributions via Jaccard Entropy Measures". In: *International Symposium on Mathematical and Computational Biology*, pp. 169–207. Springer, 2019.

[3] DE ALBUQUERQUE NETO, S. C. *MEDIDAS DE ENTROPIA ALTERNATIVAS PARA CARACTERIZAC AO DA EXISTÊNCIA DE CLAS EM FAMÍLIAS DE DOMÍNIOS DE PROTEÍNAS*. Tese de Doutorado, Universidade Federal do Rio de Janeiro, 2018.

[4] MONDAINI, R., DE ALBUQUERQUE NETO, S. C. "Entropy measures and the statistical analysis of protein family classification". In: *BIOMAT 2015: International Symposium on Mathematical and Computational Biology*, pp. 192–209. World Scientific, 2016.

[5] MONDAINI, R., DE ALBUQUERQUE NETO, S. C. "The pattern recognition of probability distributions of amino acids in protein families", *Mathematical Biology and Biological Physics (BIOMAT 2016)*, pp. 29–50, 2017.

[6] MONDAINI, R., DE ALBUQUERQUE NETO, S. "Stochastic assessment of protein databases by generalized entropy measures", *Trends in Biomathematics: Modeling, Optimization and Computational Problems*, pp. 91–105, 2018.

[7] MONDAINI, R., DE ALBUQUERQUE NETO, S. C. "Towards a Thermostatistics of the Evolution of Protein Domains Through the Formation of Families and Clans", *Trends in Biomathematics: Mathematical Modeling for Health, Harvesting, and Population Dynamics*, pp. 139–152, 2019.

[8] ERMES M. SILVA, E. M. S. *Estatística. Para Os Cursos De Economia, Administração E Ciências Contábeis - Volume 1.* Atlas, 1999. ISBN: 8522422362,9788522422364.

[9] WATSON, J., CRICK, F. *A dupla hélice: Como descobri a estrutura do DNA.* Zahar, 2014.

[10] LANCIA, G. "Applications to computational molecular biology", *Handbook on Modeling for Discrete Optimization*, v. 88, pp. 270–304, 2004.

[11] SIDMAN, K. E., GEORGE, D. G., BARKER, W. C., et al. "The protein identification resource (PIR)", *Nucleic acids research*, v. 16, n. 5, pp. 1869–1871, 1988.

[12] ZHANG, Y., LU, S., ZHOU, X., et al. "Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree, k-nearest neighbors, and support vector machine", *Simulation*, v. 92, n. 9, pp. 861–871, 2016.

[13] SHARMA, B. D., MITTAL, D. P. "New non-additive measures of entropy for discrete probability distributions", *J. Math. Sci*, v. 10, pp. 28–40, 1975.

[14] HAVRDA, J., CHARVÁT, F. "Quantification method of classification processes. Concept of structural a-entropy", *Kybernetika*, v. 3, n. 1, pp. 30–35, 1967.

[15] FERRAR, I., SCHEID, N. *História do DNA e educação científica.* Livraria da Física, 2006.

[16] MAHAN, B., MYERS, R. *Química um Curso Universitário.* Livraria da Química, 1995.

[17] NUNES, P. "Rivalidades produtivas–disputas e brigas que impulsionaram a Ciência e a Tecnologia, de Michael White", *Educação Unisinos*, v. 8, n. 15, pp. 297–303, 2004.

[18] MAGALHÃES DE OLIVEIRA YAMAZAKI, R., CHOITI YAMAZAKI, S., MULINARI STUANI, G., et al. "História da biologia e sua articulação com uma atividade experimental: extração da molécula de DNA", *Enseñanza de las Ciencias*, , n. Extra, pp. 3815–3820, 2017.

[19] LEVENE, P. A., BASS, L. W. *Nucleic acids.* Chemical Catalog Company New York, 1931.

[20] MÉTHOT, P.-O. "Bacterial transformation and the origins of epidemics in the interwar period: The epidemiological significance of Fred Griffith's "Transforming Experiment"", *Journal of the History of Biology*, v. 49, n. 2, pp. 311–358, 2016.

[21] ROBERTS, K., ALBERTS, B., JOHNSON, A., et al. *Molecular biology of the cell*. W. W. Norton & Company, 2014. ISBN: 0815345240, 978-0815345244.

[22] TWYMAN, M. *Principles of Proteomics*. Taylor & Francis, 2004. ISBN: 9781859962732, 1859962734.

[23] DE SOUSA, M., FONTES, W., RICART, C. "Análise de Proteomas: O des-pertar da era pós-genômica", *Revista on line-Biotecnologia Ciência e Desenvolvimento*, v. 7, pp. 12–14, 2003.

[24] TAMA, F., GADEA, F., MARQUES, O., et al. "Building-block approach for determining low-frequency normal modes of macromolecules", *Proteins: Structure, Function, and Bioinformatics*, v. 41, n. 1, pp. 1–7, 2000.

[25] PANDEY, A., MANN, M. "Proteomics to study genes and genomes", *Nature*, v. 405, pp. 837–846, 2000. doi: https://doi.org/10.1038/35015709.

[26] NELSON, D., COX, M. *Principles of Biochemistry*. Worth Publishers Inc.,U.S., 2004. ISBN: 0716743396, 978-0716743392.

[27] STRYER, L., TYMOCZKO, J., BERG, J. *Biochemistry*. W.H.Freeman & Co Ltd, 2002. ISBN: 0716746840, 9780716746843.

[28] KELLENBERGER, E., MULLER, P., SCHALON, C., et al. "sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank", *Journal of chemical information and modeling*, v. 46, n. 2, pp. 717–727, 2006.

[29] HUNTER, C. "Aromatic interactions in proteins, DNA and synthetic receptors", *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, v. 345, n. 1674, pp. 77–85, 1993.

[30] BARBIERI, M. "Evolution of the genetic code: the ribosome-oriented model", *Biological Theory*, v. 10, n. 4, pp. 301–310, 2015.

[31] BUSCH, J., FERRARI, P., FLESIA, A., et al. "Testing statistical hypothesis on random trees and applications to the protein classification problem", *The Annals of applied statistics*, v. 3, n. 2, pp. 542–563, 2009.

[32] FARINHA, C. "Enrolamento (Folding) de proteínas", *Revista de Ciência Elementar*, v. 2, 10 2014. doi: 10.24927/rce2014.305.

[33] WILTGEN, M. "Algorithms for structure comparison and analysis: homology modelling of proteins", *Encyclopedia of Bioinformatics and Computational Biology; Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C., Eds*, pp. 38–61, 2018.

[34] GODOI C., V., OLIVEIRA J., A. B., CHAHINE, J., et al. "Introdução ao problema de enovelamento de proteínas: uma abordagem utilizando modelos computacionais simplificados", *Revista Brasileira de Ensino de Física*, v. 40, 2018.

[35] "Protein Data Bank (PDB)". `https://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/index.html`, 1971 até 2011.

[36] STERLING, T., IRWIN, J. J. "ZINC 15–ligand discovery for everyone", *Journal of chemical information and modeling*, v. 55, n. 11, pp. 2324–2337, 2015.

[37] CONSORTIUM, U. "Ongoing and future developments at the Universal Protein Resource", *Nucleic acids research*, v. 39, n. suppl_1, pp. D214–D219, 2010.

[38] BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., et al. "GenBank", *Nucleic acids research*, v. 39, n. suppl_1, pp. D32–D37, 2010.

[39] LEINONEN, R., SUGAWARA, H., SHUMWAY, M., et al. "The sequence read archive", *Nucleic acids research*, v. 39, n. suppl_1, pp. D19–D21, 2010.

[40] ALTSCHUL, S. F., BOGUSKI, M. S., GISH, W., et al. "Issues in searching molecular sequence databases", *Nature genetics*, v. 6, n. 2, pp. 119–129, 1994.

[41] EDDY, S. R. "Hidden markov models", *Current opinion in structural biology*, v. 6, n. 3, pp. 361–365, 1996.

[42] KROGH, A., BROWN, M., MIAN, I. S., et al. "Hidden Markov models in computational biology. Applications to protein modeling", *Journal of molecular biology*, v. 235, n. 5, pp. 1501–1531, 1994.

[43] FINN, R. D., BATEMAN, A., CLEMENTS, J., et al. "Pfam: the protein families database", *Nucleic acids research*, v. 42, n. D1, pp. D222–D230, 2014.

[44] PUNTA, M., COGGILL, P. C., EBERHARDT, R. Y., et al. "The Pfam protein families database", *Nucleic acids research*, v. 40, n. D1, pp. D290–D301, 2012.

[45] SAMMUT, S. J., FINN, R. D., BATEMAN, A. "Pfam 10 years on: 10 000 families and still growing", *Briefings in bioinformatics*, v. 9, n. 3, pp. 210–219, 2008.

[46] EL-GEBALI, S., MISTRY, J., BATEMAN, A., et al. "The Pfam protein families database in 2019", *Nucleic acids research*, v. 47, n. D1, pp. D427–D432, 2019.

[47] FINN, R. D., MISTRY, J., SCHUSTER-BÖCKLER, B., et al. "Pfam: clans, web tools and services", *Nucleic acids research*, v. 34, n. suppl_1, pp. D247–D251, 2006.

[48] FINN, R. D., TATE, J., MISTRY, J., et al. "The Pfam protein families database", *Nucleic acids research*, v. 36, n. suppl_1, pp. D281–D288, 2007.

[49] CANTELLI, G., BATEMAN, A., BROOKSBANK, C., et al. "The European Bioinformatics Institute (EMBL-EBI) in 2021", *Nucleic Acids Research*, v. 50, n. D1, pp. D11–D19, 11 2021. ISSN: 0305-1048. doi: 10.1093/nar/gkab1127. Disponível em: <https://doi.org/10.1093/nar/gkab1127>.

[50] BLUM, M., CHANG, H., CHUGURANSKY, S., et al. "The InterPro protein families and domains database: 20 years on", *Nucleic Acids Research*, v. 49, n. D1, pp. D344–D354, 11 2020. ISSN: 0305-1048. doi: 10.1093/nar/gkaa977. Disponível em: <https://doi.org/10.1093/nar/gkaa977>.

[51] STEINEGGER, M., SÖDING, J. "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets", *Nature biotechnology*, v. 35, n. 11, pp. 1026–1028, 2017.

[52] STEINEGGER, M., SÖDING, J. "Clustering huge protein sequence sets in linear time", *Nature communications*, v. 9, n. 1, pp. 1–8, 2018.

[53] ALTSCHUL, S. F., MADDEN, T. L., SCHÄFFER, A. A., et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic acids research*, v. 25, n. 17, pp. 3389–3402, 1997.

[54] SÖDING, J. "Protein homology detection by HMM–HMM comparison", *Bioinformatics*, v. 21, n. 7, pp. 951–960, 2005.

[55] FINN, R. D., CLEMENTS, J., EDDY, S. R. "HMMER web server: interactive sequence similarity searching", *Nucleic acids research*, v. 39, n. suppl_2, pp. W29–W37, 2011.

[56] HAUSSLER, D., KROGH, A. "Protein alignment and clustering". In: *conference Neural Networks for Computing*, 1992.

[57] HAUSSLER, D., KROGH, A., MIAN, I. S., et al. "Protein modeling using hidden Markov models: Analysis of globins". In: *[1993] Proceedings of the Twenty-sixth Hawaii International Conference on System Sciences*, v. 1, pp. 792–802. IEEE, 1993.

[58] SONNHAMMER, E. L., EDDY, S. R., DURBIN, R. "Pfam: a comprehensive database of protein domain families based on seed alignments", *Proteins: Structure, Function, and Bioinformatics*, v. 28, n. 3, pp. 405–420, 1997.

[59] LIPMAN, D. J., PEARSON, W. R. "Rapid and sensitive protein similarity searches", *Science*, v. 227, n. 4693, pp. 1435–1441, 1985.

[60] CARELS, N., MONDAINI, C. F., MONDAINI, R. P. "Entropy measures based method for the classification of protein domains into families and clans". In: *BIOMAT 2013: International Symposium on Mathematical and Computational Biology*, pp. 209–218. World Scientific, 2014.

[61] GRAVETTER, F. J., WALLNAU, L. B., FORZANO, L.-A. B., et al. *Essentials of statistics for the behavioral sciences*. Cengage Learning, 2020.

[62] MEYER, P. L. *Introductory probability and statistical applications*. Oxford and IBH Publishing, 1965.

[63] MAGALHÃES, N. M. *Probabilidade e variáveis aleatórias*. Edusp, 2006.

[64] BATSCHELET, E. *Introduction to mathematics for life scientists*. Springer Science & Business Media, 2012.

[65] ATCHLEY, W. R., WOLLENBERG, K. R., FITCH, W. M., et al. "Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis", *Molecular biology and evolution*, v. 17, n. 1, pp. 164–178, 2000.

[66] DURBIN, R., S. E. R. K. A. M. G. *Biological Sequence Analysis Probabilistic Models of Proteins and Nucleic Acids*, v. 2. Cambridge university press, 1998.

[67] BORODOVSKY, M., EKISHEVA, S. *Problems and solutions in biological sequence analysis*, v. 1. Cambridge University Press, 2006. ISBN: 0521847540, 9780511335129.

[68] VOLKENSTEIN, M. V. *Entropy and information*, v. 1. Springer Science & Business Media, 2009.

[69] BUSTAMANTE, C. D., FLEDEL-ALON, A., WILLIAMSON, S., et al. "Natural selection on protein-coding genes in the human genome", *Nature*, v. 437, n. 7062, pp. 1153–1157, 2005.

[70] MARTINS, R. A. "Mayer e a conservação da energia", *Cadernos de História e Filosofia da ciência*, v. 6, pp. 63–95, 1984.

[71] VAN W., G., SONNTAG, R. E., BORGNAKKE, C. *Fundamentos da termodinâmica clássica*. Editora Blucher, 1994.

[72] MONDAINI, R. P. *Trends in Biomathematics: Modeling Cells, Flows, Epidemics, and the Environment*. Springer, 2020.

[73] RÉNYI, A. "On Measures of Entropy and Information". In: Neyman, J. (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, v. 1, pp. 547–561, Berkeley, California, USA, 1961. University of California Press.

[74] LANDSBERG, P. T., VEDRAL, V. "Distributions and Channel Capacities in Generalized Statistical Mechanics", *Physics Letters A*, v. 247, n. 3, pp. 211–217, 1998. ISSN: 0375-9601. doi: 10.1016/S0375-9601(98)00500-3.

[75] OIKONOMOU, T. "Properties of the "non-extensive Gaussian" entropy", *Physica A: Statistical Mechanics and its Applications*, v. 381, pp. 155–163, 2007. ISSN: 0378-4371. doi: 10.1016/j.physa.2007.03.010.

[76] MONDAINI, R. P., DE ALBUQUERQUE NETO, S. C. "A Jaccard-like Symbol and its Usefulness in the Derivation of Amino Acid Distributions in Protein Domain Families". In: *International Symposium on Mathematical and Computational Biology*, pp. 201–220. Springer, 2020.

[77] PEARSON, K. "Contributions to the mathematical theory of evolution", *Philosophical Transactions of the Royal Society of London. A*, v. 185, pp. 71–110, 1894.

[78] BULJAN, M., BATEMAN, A. "The evolution of protein domain families", *Biochemical Society Transactions*, v. 37, n. 4, pp. 751–755, 2009.

[79] KONC, J., JANEŽIČ, D. "ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment", *Bioinformatics*, v. 26, n. 9, pp. 1160–1168, 2010.

[80] LIU, J., ROST, B. "Sequence-based prediction of protein domains", *Nucleic acids research*, v. 32, n. 12, pp. 3522–3530, 2004.

[81] ZHOU, H., SKOLNICK, J. "Template-based protein structure modeling using TASSERVMT", *Proteins: Structure, Function, and Bioinformatics*, v. 80, n. 2, pp. 352–361, 2012.

[82] WANG, Y., ZHANG, H., ZHONG, H., et al. "Protein domain identification methods and online resources", *Computational and structural biotechnology journal*, v. 19, pp. 1145–1153, 2021.

[83] QUEVILLON, E., SILVENTOINEN, V., PILLAI, S., et al. "InterProScan: protein domains identifier", *Nucleic acids research*, v. 33, n. suppl_2, pp. W116–W120, 2005.

[84] LI, W., COWLEY, A., ULUDAG, M., et al. "The EMBL-EBI bioinformatics web and programmatic tools framework", *Nucleic acids research*, v. 43, n. W1, pp. W580–W584, 2015.

[85] ZARDECKI, C., DUTTA, S., GOODSELL, D. S., et al. "RCSB Protein Data Bank: A resource for chemical, biochemical, and structural explorations of large and small biomolecules". 2016.

[86] WU, C. H., YEH, L.-S. L., HUANG, H., et al. "The protein information resource", *Nucleic acids research*, v. 31, n. 1, pp. 345–347, 2003.

[87] BLUM, M., CHANG, H.-Y., CHUGURANSKY, S., et al. "The InterPro protein families and domains database: 20 years on", *Nucleic acids research*, v. 49, n. D1, pp. D344–D354, 2021.

[88] HULO, N., BAIROCH, A., BULLIARD, V., et al. "The PROSITE database", *Nucleic acids research*, v. 34, n. suppl_1, pp. D227–D230, 2006.

[89] LU, S., WANG, J., CHITSAZ, F., et al. "CDD/SPARCLE: the conserved domain database in 2020", *Nucleic acids research*, v. 48, n. D1, pp. D265–D268, 2020.

[90] WILSON, D., MADERA, M., VOGEL, C., et al. "The SUPERFAMILY database in 2007: families and functions", *Nucleic acids research*, v. 35, n. suppl_1, pp. D308–D313, 2007.

[91] CONSORTIUM, G. O. "The Gene Ontology (GO) database and informatics resource", *Nucleic acids research*, v. 32, n. suppl_1, pp. D258–D261, 2004.

[92] LEES, J., YEATS, C., PERKINS, J., et al. "Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis", *Nucleic Acids Research*, v. 40, n. D1, pp. D465–D471, 12 2011. ISSN: 0305-1048. doi: 10.1093/nar/gkr1181. Disponível em: <https://doi.org/10.1093/nar/gkr1181>.