MEDIDAS DE ENTROPIA ALTERNATIVAS PARA CARACTERIZAÇÃO DA EXISTÊNCIA DE CLÃS EM FAMÍLIAS DE DOMÍNIOS DE PROTEÍNAS

Simão Coutinho de Albuquerque Neto

Tese de Doutorado 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 Doutor em Engenharia de Sistemas e Computação.

Orientador: Rubem Pinto Mondaini

Rio de Janeiro
Outubro de 2018

MEDIDAS DE ENTROPIA ALTERNATIVAS PARA CARACTERIZAÇÃO DA EXISTÊNCIA DE CLÃS EM FAMÍLIAS DE DOMÍNIOS DE PROTEÍNAS

Simão Coutinho de Albuquerque Neto

TESE 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 DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Rubem Pinto Mondaini, D.Sc.

Prof. José Abdalla Helayël Neto, Ph.D.

Prof. Marco Antonio von Krüger, Ph.D.

Prof. Heraldo Luis Silveira de Almeida, D.Sc.

Prof. Claudia Maria Lima Werner, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
OUTUBRO DE 2018

Albuquerque Neto, Simão Coutinho de

Medidas de Entropia Alternativas para Caracterização da Existência de Clãs em Famílias de Domínios de Proteínas/Simão Coutinho de Albuquerque Neto. – Rio de Janeiro: UFRJ/COPPE, 2018.

XIV, 104 p.: il.; 29, 7cm.

Orientador: Rubem Pinto Mondaini

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2018.

Referências Bibliográficas: p. 101 – 104.

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

À minha família.

Agradecimentos

Aos meus pais, por todo apoio e carinho dados em todos os momentos de minha vida.
Ao meu orientador, Professor Rubem Mondaini, pelo curso de Controle Ótimo lecionado por ele na Escola Politécnica da UFRJ, que fez despertar em mim o interesse pela pesquisa científica, pela oportunidade dada de fazer os cursos de mestrado e de doutorado, pelo acompanhamento experiente das etapas necessárias à consecução desta tese, por sua grande experiência como pesquisador profissional, e por todos os conselhos e incentivos que visam a mostrar-me o que significa o trabalho de um cientista.
Aos meus amigos, que sempre me deram motivação.
À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte financeiro dado nos últimos anos.

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

MEDIDAS DE ENTROPIA ALTERNATIVAS PARA CARACTERIZAÇÃO DA EXISTÊNCIA DE CLÃS EM FAMÍLIAS DE DOMÍNIOS DE PROTEÍNAS

Simão Coutinho de Albuquerque Neto

Outubro/2018

Orientador: Rubem Pinto Mondaini

Programa: Engenharia de Sistemas e Computação

O sequenciamento e catalogação de proteínas é uma iniciativa que busca uma classificação completa e precisa dos domínios protéicos. É de extrema importância na determinação das causas de doenças genéticas e na descoberta de tratamentos eficazes para as mesmas. Os domínios que apresentam similaridades em suas sequências, estruturas e funções são agrupados em famílias. Baseado na opinião de especialistas em relação à similaridade de funções, algumas famílias são reunidas em clãs. No presente trabalho, a estatística ANOVA é usada para a verificação da legitimidade da classificação de famílias de proteínas em clãs. Os resultados com blocos (100×200) e (100×100) representativos das famílias são apresentados e discutidos.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

ALTERNATIVE ENTROPY MEASURES FOR THE CHARACTERIZATION OF THE EXISTENCE OF CLANS IN PROTEIN DOMAIN FAMILIES

Simão Coutinho de Albuquerque Neto

October/2018

Advisor: Rubem Pinto Mondaini

Department: Systems Engineering and Computer Science

The sequencing and cataloging of proteins is an initiative that seeks a complete and accurate classification of protein domains. It is really important in the determination of the cause of genetic diseases and in the discovery of effective treatments for them. The domains which have similar sequences, structures and functions are grouped into families. Based on the opinion of expert biologists regarding the similarity of functions, some families are collected in clans. On the present work the ANOVA statistics is used to verify the legitimacy of the classification of protein families in clans. The results with the (100×200) and (100×100) representative family blocks are presented and discussed.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A existência do que chamamos de "vida" deu-se quando, sob diversas condições favoráveis e talvez de forma aleatória, átomos começaram a se combinar em moléculas e estas em estruturas maiores, as macromoléculas, capazes de se reproduzirem e multiplicarem, transmitindo suas características. Devido aos mais diversos tipos de combinações e mutações, pequenas cadeias de aminoácidos aumentaram gradativamente de tamanho e se uniram, dando origem a estruturas estáveis e de maior complexidade, denominadas proteínas.

As proteínas são macromoléculas de extrema importância que exercem diversos tipos de funções nos organismos dos seres vivos, como por exemplo, os anticorpos, as enzimas e as proteínas transportadoras. A evolução da vida em termos de diversidade de espécies, propriedades e características específicas de um espécime e desenvolvimento de novas habilidades está diretamente relacionada à evolução da complexidade e da capacidade de realizar novas funções das proteínas. Novas estruturas estáveis podem acarretar tanto em vantagens quanto desvantagens para o indivíduo. O mal da vaca louca é um exemplo de doença relacionada à replicação de proteínas com estruturas alteradas após serem infectadas por príons (*proteinaceous infectious particles*), proteínas que se enovelaram de forma anormal. Recentemente, um novo tipo de príon foi sintetizado em laboratório [1] com o objetivo de encontrar um possível tratamento para estas proteínas que não são neutralizadas pelo sistema imunológico.

As proteínas são formadas por uma ou mais estruturas independentemente estáveis, denominadas domínios. Durante as últimas décadas, diversos grupos científicos têm analizado, catalogado e armazenado os domínios em volumosos bancos de dados. Aqueles domínios cujas funções e sequências de aminoácidos sejam verificadas como similares, são agrupados em famílias. A classificação em famílias implica que os domínios têm algum tipo de relação evolutiva entre si: ou vêm de um ancestral em comum ou alguns destes são descendentes de outros da mesma família, que por sua vez tiveram sua origem a partir de mais outros também da mesma família, como um

grande braço de uma árvore genealógica.

Algumas destas famílias, por sua vez, são reunidas em grupos denominados clãs, conforme a similaridade de suas funções, porém uma ancestralidade comum, determinada pela semelhança de suas sequências, não é detectada.

As informações contidas nestes bancos de dados devem auxiliar na construção de uma teoria que descreva os processos de enovelamento e desenovelamento de proteínas. Uma ideia fundamental aqui apresentada é que os domínios de proteínas não evoluem independentemente, mas sim fazem parte de um processo de formação de famílias de proteínas (PFFP — *Protein Family Formation Process*) [2]. As distribuições dos aminoácidos nas famílias podem ser analizadas através de métodos de reconhecimentos de padrões e, ao considerarmos diversas versões do banco de dados, com a sua evolução ao longo do tempo, podemos estudar este processo estocástico através da formulação de uma equação mestre ou de uma equação de Fokker-Planck. Esforços nestes sentidos já têm sido realizados ao longo dos últimos anos [3–7].

Uma teoria de formação e evolução de proteínas elucidaria as condições iniciais necessárias para descrever a dinâmica molecular de protomoléculas primitivas até as proteínas atuais e faria previsões para o processo de enovelamento. Na ausência desta teoria podemos apenas obter um processo de polimerização que leva à formação de peptídeos contendo apenas aminoácidos iguais [3]. Seja por exemplo um aminoácido com uma cadeia lateral R como na Figura 1.1, cuja formação atômica pode ser escrita como:

$$a_1 = H_2NC_\alpha HRCOOH$$

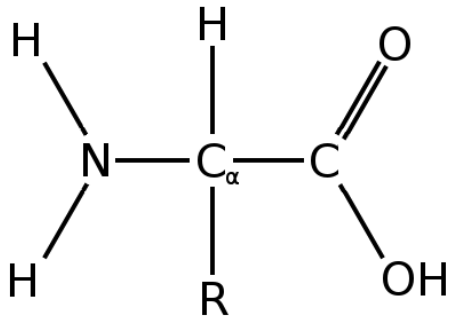


Figura 1.1: Exemplo de um aminoácido com cadeia lateral R.

Na Figura 1.2 temos a associação de dois aminoácidos com cadeias laterais R_1 e R_2. A união dos dois monômeros é uma polimerização de condensação, uma vez que além do peptídeo temos a formação de uma molécula de água.

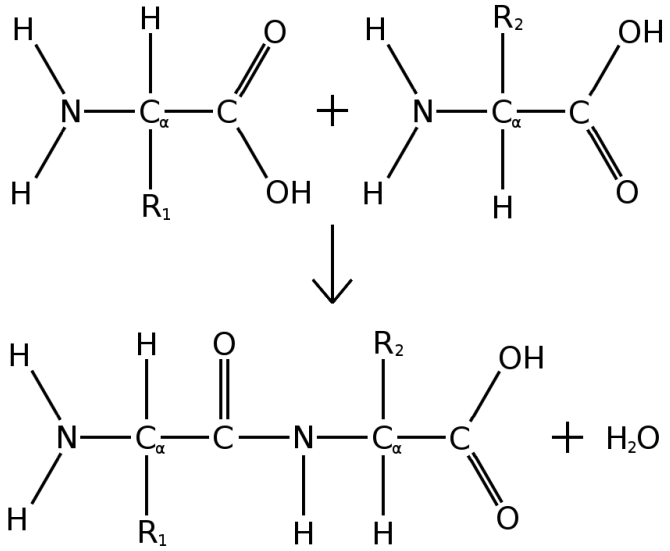


Figura 1.2: Ligação peptídica entre dois aminoácidos com cadeias laterais R_1 e R_2.

Para dois monômeros iguais, o polímero resultante apresenta a seguinte cadeia:

$$a_2 = H(HNC_\alpha HRCO)_2OH$$

A união de três aminoácidos idênticos cria a cadeia

$$a_3 = H(HNC_\alpha HRCO)_3OH$$

e duas moléculas de água (Figura 1.3). Já ao associarmos a cadeia a_2 com mais um aminoácido contendo o mesmo resíduo R, temos a formação da mesma cadeia a_3 e a liberação de apenas uma molécula de água (Figura 1.4).

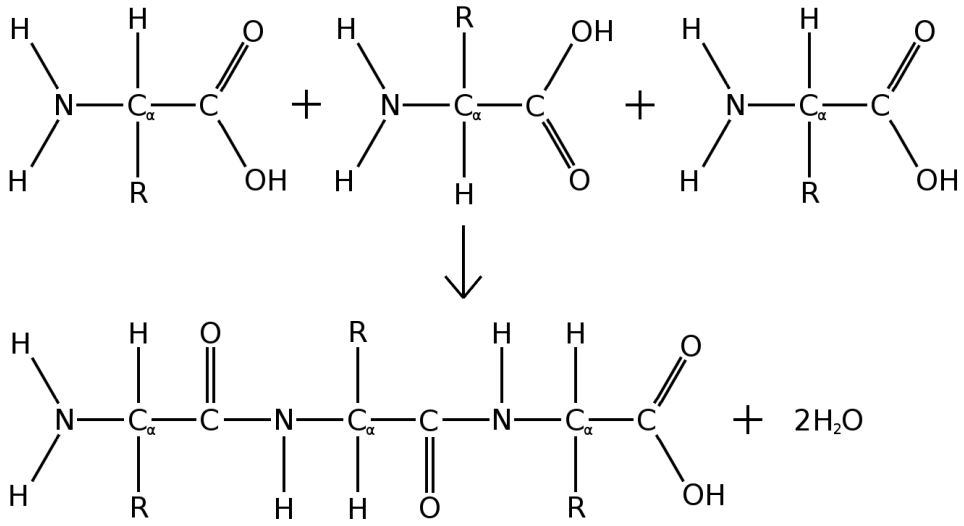


Figura 1.3: Ligação peptídica entre três aminoácidos iguais.

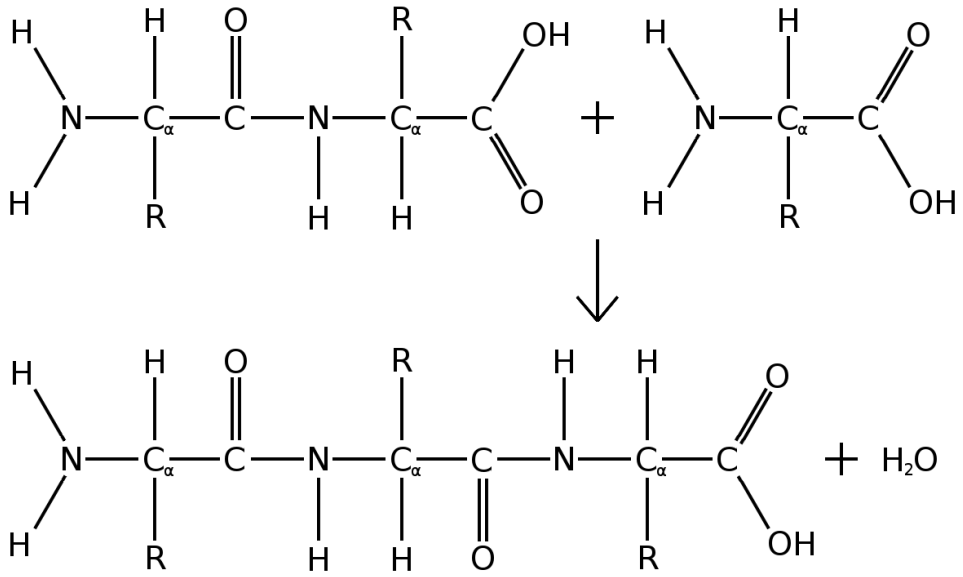


Figura 1.4: Ligação peptídica entre a cadeia a_2 com um aminoácido com o mesmo resíduo R.

A associação de n aminoácidos com cadeia a_1 resulta na formação de um peptídeo com cadeia a_n e na liberação de $(n-1)$ moléculas de água (Figura 1.5):

$$na_1 = a_n + (n-1)H_2O$$

com

$$a_n = H(HNC_\alpha HRCO)_n OH$$

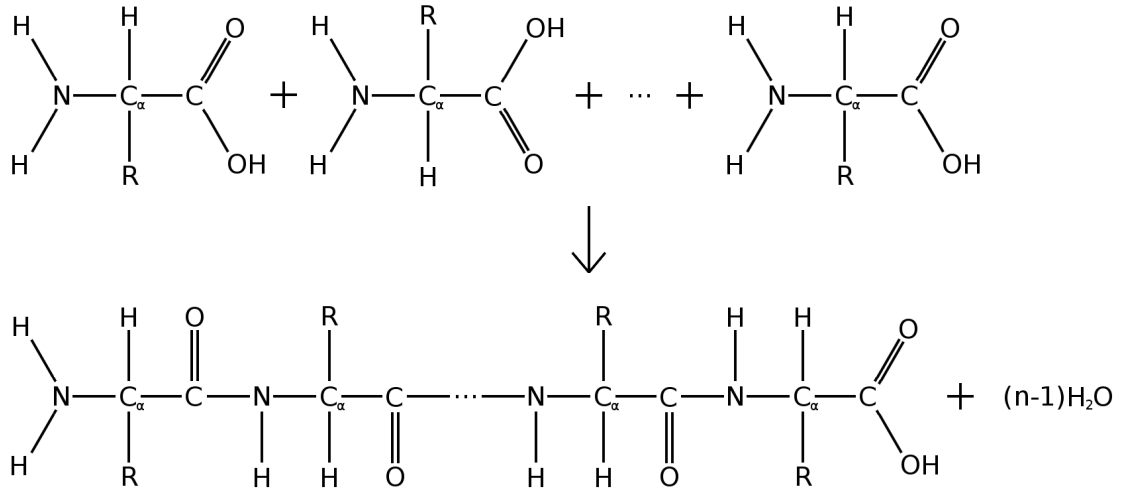


Figura 1.5: Ligação peptídica entre n aminoácidos iguais.

Já a ligação de uma cadeia formada por $(n-1)$ aminoácidos com cadeia a_1 com mais um aminoácido idêntico, forma uma cadeia a_n e uma molécula de água:

$$a_{n-1} + a_1 = a_n + H_2O$$

Uma cadeia formada por j aminoácidos iguais pode ser escrita como:

$$a_j = H(HNC_\alpha HRCO)_j OH$$

A associação de n cadeias a_j forma uma cadeia na_j e libera $(n-1)$ moléculas de água:

$$na_j = a_{jn} + (n-1)H_2O$$

Uma ligação peptídica entre uma cadeia formada por j aminoácidos e outra com k aminoácidos pode ser escrita como:

$$a_j + a_k = a_{j+k} + H_2O$$

E a associação de n cadeias a_j com m cadeias a_k:

$$\begin{aligned} na_j + ma_k &= a_{jn} + (n-1)H_2O + a_{km} + (m-1)H_2O \\ &= a_{jn} + a_{km} + (n+m-2)H_2O \\ &= a_{jn+km} + (n+m-1)H_2O \end{aligned}$$

Com isso temos uma álgebra simples de associação de cadeias de aminoácidos. Porém, no caso real onde ocorrem combinações com os diferentes tipos de aminoácidos, com R_1, R_2, R_3, ..., R_n, a vida não é tão simples [3]. Como ainda não somos capazes de inferir a dinâmica de associação a partir de um modelo real, mais complexo, utilizamos modelos probabilísticos e estatística para analisar a regularidade na ocorrência de aminoácidos nas proteínas, de forma a obter informações que ajudem a elucidar como se dá o processo de polimerização.

O presente trabalho reune os resultados preliminares de testes estatísticos realizados para a verificação da validade da classificação de famílias de domínios de proteínas em clãs. A determinação de uma ancestralidade auxilia no entendimento do 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.

O capítulo 2 aborda brevemente as características químicas e biológicas de uma proteína, relacionando as sequências de aminoácidos à estrutura e à função protéica. No capítulo 3 tratamos sobre bancos de dados de domínios de proteínas, com especial ênfase no PFAM, o banco de dados utilizado nos testes aqui apresentados. O quarto capítulo contém a definição do espaço amostral e apresenta a notação utilizada no cálculo das probabilidades de distribuição dos aminoácidos nas famílias de domínios do PFAM. O capítulo 5 apresenta diversas medidas de entropia presentes na literatura

e alguns testes utilizando o banco de dados. No capítulo 6 discorremos sobre testes comparativos de tempo de processamento utilizando diferentes sistemas operacionais e computacionais. No sétimo capítulo demonstramos como é feita a formulação para o teste ANOVA e apresentamos os resultados utilizando blocos representativos (100×200) e (100×100) das famílias de domínios de proteínas. Por último, o capítulo 8 apresenta as conclusões e algumas sugestões para trabalhos futuros.

Capítulo 2

Proteínas – Funções e Estruturas

2.1 Proteínas

As proteínas são as moléculas mais importantes para a manutenção da vida, uma vez que são responsáveis por diversas funções celulares: catálise de reações químicas, transporte de substratos etc. A evolução na complexidade das proteínas está diretamente relacionada à evolução dos seres vivos. Diversas doenças, como os males de Alzheimer e de Parkinson, por exemplo, são resultantes da má formação de proteínas.

Os aminoácidos são os monômeros que constituem a proteína. Sua estrutura (Figura 2.1) consiste em um grupo amina (NH_2) e um grupo carboxila (COOH) conectados a um átomo de carbono, frequentemente nomeado como carbono alfa, C_α (por estar na posição central aos dois grupos e para diferenciá-lo do átomo de carbono da carboxila), que ainda tem como ligações um átomo de hidrogênio e uma cadeia lateral, R. Esta cadeia lateral (também chamada de radical ou resíduo) é o grande diferencial que identifica cada um dos 20 tipos de aminoácidos existentes no código genético (Tabela 2.1), podendo conter desde um simples átomo de hidrogênio (Glicina) até estruturas mais complexas com anéis aromáticos (Fenilalanina, Tirosina e Triptofano). Devido à sua diversidade, os diferentes resíduos apresentam também diferentes polaridades, o que é um fator crucial na estrutura tridimensional do estado enovelado da proteína. As cadeias de aminoácidos observadas na Natureza costumam ter entre dezenas a milhares de monômeros [8, 9].

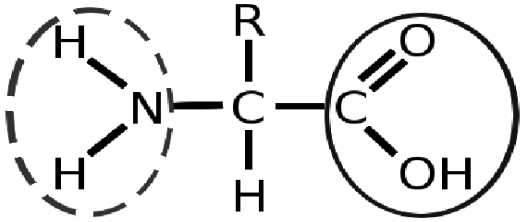


Figura 2.1: Estrutura de um aminoácido. O grupo amina (NH_2) está circulado por uma linha tracejada, enquanto o grupo carboxila (COOH) está circulado por uma linha contínua.

Tabela 2.1: Os 20 diferentes tipos de aminoácidos presentes no código genético. Adaptado de [8, 10, 11].

Aminoácido	**Códigos**		**Polaridade**
	3 letras	1 letra	
Alanina	Ala	A	apolar
Cisteina	Cys	C	apolar
Ácido aspártico	Asp	D	negativa
Ácido glutâmico	Glu	E	negativa
Fenilalanina	Phe	F	apolar
Glicina	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

As proteínas têm sua origem nos ribossomos das células através do processo de tradução do RNA mensageiro (RNAm), molécula portadora da sequência de nucleotídeos transcrita do DNA a ser sintetizada. Cada grupo de três nucleotídeos do RNAm (códons) designa um aminoácido (Tabela 2.2) que, por meio de uma reação química, tem o átomo de carbono do seu grupo carboxila unido por uma ligação covalente ao átomo de nitrogênio do grupo amina do aminoácido anterior do polipeptídeo em formação, e libera uma molécula de água no processo. A cadeia de aminoácidos formada apresenta a sequência central $NC_{\alpha}C$ que se repete ao longo de toda a proteína (Figura 2.2). Esta sequência $NC_{\alpha}CNC_{\alpha}C...NC_{\alpha}C$ é chamada de *espinha dorsal* da proteína.

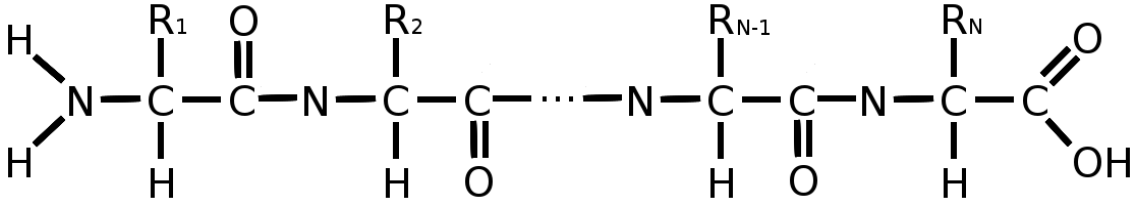


Figura 2.2: Exemplo de uma cadeia de aminoácidos.

Tabela 2.2: O código genético. Relação entre códons e aminoácidos. Adaptado de [12].

Primeira posição	Segunda posição								Terceira posição
	U		**C**		**A**		**G**		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	**U**
	UUC		UCC		UAC		UGC		**C**
	UUA	Leu	UCA		UAA	parada	UGA	parada	**A**
	UUG		UCG		UAG		UGG	Trp	**G**
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	**U**
	CUC		CCC		CAC		CGC		**C**
	CUA		CCA		CAA	Gln	CGA		**A**
	CUG		CCG		CAG		CGG		**G**
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	**U**
	AUC		ACC		AAC		AGC		**C**
	AUA		ACA		AAA	Lys	AGA	Arg	**A**
	AUG*	Met	ACG		AAG		AGG		**G**
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	**U**
	GUC		GCC		GAC		GGC		**C**
	GUA		GCA		GAA	Glu	GGA		**A**
	GUG*		GCG		GAG		GGG		**G**

*Além de codificar o aminoácido Metionina, o códon AUG especifica o início da sequência (proteína). O mesmo também ocorre com o códon GUG, porém numa frequência muito menor. Os códons UAA, UAG e UGA, por sua vez, codificam apenas o término da sequência, não especificando nenhum aminoácido.

2.2 Aspectos Conformacionais

A função de uma proteína está relacionada à sua conformação tridimensional que, por sua vez, se deve à sua sequência de aminoácidos. Esta sequência que forma a proteína é conhecida como sendo sua estrutura primária. As estruturas secundárias são relacionadas com as conformações tridimensionais de segmentos frequentemente observados nas proteínas, sendo as hélices-α e as folhas-β as estruturas mais comuns (Figura 2.3). A formação dessas estruturas ocorre devido a ligações de hidrogênio entre átomos de hidrogênio do grupo amina com o oxigênio do grupo carboxila da espinha dorsal da proteína.

A estrutura terciária corresponde a estrutura da sequência completa da cadeia de aminoácidos da molécula. É o efeito da proteína dobrar-se sobre si mesma, formando um aglomerado de estruturas secundárias. A estrutura quaternária é formada pela união de moléculas de proteínas enoveladas (estruturas terciárias), o que a faz adquirir uma conformação própria para desempenhar funções biológicas específicas.

As proteínas contêm um ou mais blocos básicos denominados *domínios*, que possuem estrutura e/ou funções particulares, e que podem ser encontrados (com alguma

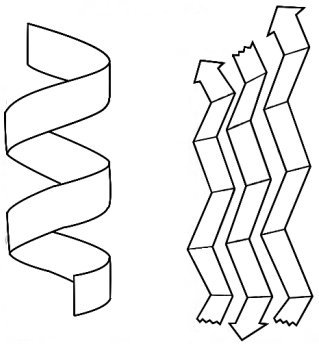


Figura 2.3: Exemplos de estruturas secundárias. Na esquerda uma hélice-α e na direita uma folha-β. Adaptado de [8].

alteração em sua sequência de aminoácidos) em proteínas com diferentes funções. Os domínios geralmente se enovelam independentemente do restante da proteína e em seguida a auxiliam a adquirir sua conformação definitiva. A evolução molecular se dá tanto pela criação de novos domínios, através da alteração na sequência de aminoácidos de um domínio preexistente que acarrete, como consequência, em uma outra estrutura tridimensional; quanto por novas combinações de domínios que resultem em conformações estáveis.

O processo de enovelamento ocorre devido à existência de diversos tipos de interações (hidrofóbicas e eletrostáticas, ligações covalentes, iônicas, de hidrogênio), ao auxílio de outras proteínas (chaperonas) e a outros fatores externos (como a viscosidade do solvente onde a proteína se encontra diluída). Devido a tantos fatores, até o presente momento não foi possível determinar quais destes seria o fator crucial na direção da sequência de aminoácidos até o estado nativo enovelado da proteína. Uma maior compreensão de tais relações é de grande valia na determinação dos fatores que levam uma proteína a adotar uma conformação diferente que resulte em patologias.

Capítulo 3

Bancos de Dados de Proteínas

Como dito no capítulo anterior, os domínios são estruturas que são encontradas em diferentes proteínas cujas funções podem ser similares ou não. A identificação dos domínios e, posteriormente, a decomposição de proteínas nestas sub-unidades têm grande importância na determinação da função biológica da molécula. Com o avanço científico nas últimas décadas, a classificação de proteínas através da identificação de similaridades em suas sequências de aminoácidos tornou-se possível [13].

Neste contexto, diversos grupos científicos surgiram ou passaram a se dedicar à identificação de sequências de aminoácidos e a armazená-las em grandes bancos de dados. Como exemplo, podemos citar: *Universal Protein resource* (UniProt) [14]; *Protein Research Foundation, Japan* (PRF) [15]; e *Protein Family Database* (PFAM) [16].

O banco de dados PFAM contém uma vasta gama de domínios de proteínas, famílias de domínios e clãs. Através de um processo de verificação de semelhança entre as sequências de aminoácidos de diferentes domínios, estes podem ser agrupados em famílias, de forma que é esperado que elas reunam indivíduos que evoluíram de algum ancestral em comum [17]. A similaridade entre as sequências dos membros de uma mesma família não implica necessariamente que estes têm funções similares [17].

Os clãs, introduzidos no PFAM a partir da versão 18.0 [18], são agrupamentos de duas ou mais famílias relacionadas, sendo como "superfamílias". Os domínios de um clã apresentam similaridades em suas sequências de aminoácidos, mas não com um nível de significância alto o suficiente para classificar todos como pertencentes a uma mesma família. A classificação em clãs auxilia na predição de funções e estruturas de famílias com funções desconhecidas ao relacionarmos com outras que contenham mais informações [19]. A Tabela 3.1 apresenta a evolução do banco de dados PFAM da versão 18.0 até a atual 31.0. Na versão 27.0 adotada no presente trabalho, estão registradas um total de 14831 famílias, mas apenas 4563 destas estão classificadas em clãs. Desde a versão 26.0 milhares de famílias com funções conhecidas cadastradas no PFAM contêm um redirecionamento para artigos na Wikipedia com informações

sobre sua função biológica [20].

Tabela 3.1: Evolução do banco de dados PFAM.

Banco de dados PFAM				
versão	ano	nº de famílias	nº de famílias class. em clãs	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

Algumas alterações são feitas a cada nova versão disponibilizada do PFAM, como: inclusão ou exclusão de domínios; criação ou destruição de famílias, podendo os domínios que formavam a família destruída serem distribuídos entre uma ou várias famílias; clãs podem ser criados ou destruídos, podendo as famílias de um clã destruído serem aglutinadas por um ou mais clãs. A justificativa para tais alterações se deve ao fato dos resultados computacionais serem constantemente analisados pelos biólogos especialistas que dão seu parecer sobre eles.

O PFAM foi criado com o objetivo não apenas de armazenar todos os domínios, mas também de acelerar a caracterização e classificação completa e precisa destes, de forma a ampliar a compreensão do espaço de sequências de proteínas [17, 21]. A identificação dos domínios na proteína no seu estado natural, dobrada sobre si mesma, é uma grande dificuldade, mas através do uso de solventes que fazem com que ela se desenovele, podemos determinar todos os aminoácidos presentes em sua estrutura primária, e sua sequência é comparada com as *sementes* (modelos que caracterizam as famílias de domínios armazenados previamente) contidas no banco de dados.

As sementes 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

justificasse a relação entre as diferentes sequências, chamado de *alinhamento*, foi desenvolvido. O processo de alinhamento consiste na busca do melhor resultado de correspondência correta entre os aminoácidos das sequências sobrepostas. Para tal, buracos ou vãos podem ser acrescentados às sequências. A justificativa para a utilização destes vãos se deve ao fato que durante o processo evolutivo, o que antes era um único domínio, ramificou-se em múltiplos indivíduos através da inclusão e/ou remoção de resíduos, além da substituição de alguns resíduos.

O alinhamento foi inicialmente feito entre sequências que se sabia serem biologicamente relacionadas e, posteriormente, para checar a incorporação de novos indivíduos suspeitos às famílias existentes. Todo o processo de verificar a estrutura de uma proteína, determinar sua composição química, identificar os domínios, alinhá-los com outros previamente caracterizados, 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.

No início da década de 1990, os primeiros trabalhos envolvendo a utilização de *modelos ocultos de Markov* (HMM — *hidden Markov models*) na modelagem de alinhamentos de múltiplas sequências de aminoácidos começaram a surgir [22, 23], o que inspirou e motivou diversos grupos a adotarem métodos com modelos probabilísticos, e na segunda metade da mesma década, no ano de 1997, a primeira versão do PFAM foi disponibilizada [24]. O uso de HMM possibilitou um grande incremento na velocidade de reconhecimento dos domínios.

As sementes foram criadas a partir da escolha pelos especialistas de indivíduos considerados mais relevantes para a caracterização das famílias. Os mais relevantes não são aqueles mais semelhantes entre si, mas sim aqueles que, sendo *a priori* reconhecidos como sendo pertencentes à mesma família, apresentam maior variedade. Após a seleção dos elementos, dá-se o processo de alinhamento. Com as sequências alinhadas são feitos perfis de modelos HMM, que são justamente as sementes que depois serão utilizadas para a identificação dos domínios nas sequências inseridas em versões subsequentes. Quando uma sequência não é reconhecida pelo software como sendo pertencente a uma dada família, mas é tida como tal pelo critério dos especialistas, ela é então não apenas incluída na família, mas também à semente. Feita a inclusão, todas as sequências cadastradas na família passam novamente pelo processo de análise para verificar se o "casamento" com a semente permanece relevante. Se a "pontuação" diminuir significantemente para uma determinada sequência, pode ocorrer desta ser também incluída à semente, dando reinício ao processo de verificação, ou ela pode ser removida da família. Para não haver falsos positivos ou falsos negativos na identificação de sequências, a semente deve ser construída com o máximo de qualidade possível.

Na Figura 3.1 temos um esquema de identificação de domínios em proteínas e sua organização em famílias. Para as 12 proteínas fictícias são identificados um total de sete diferentes tipos de 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.

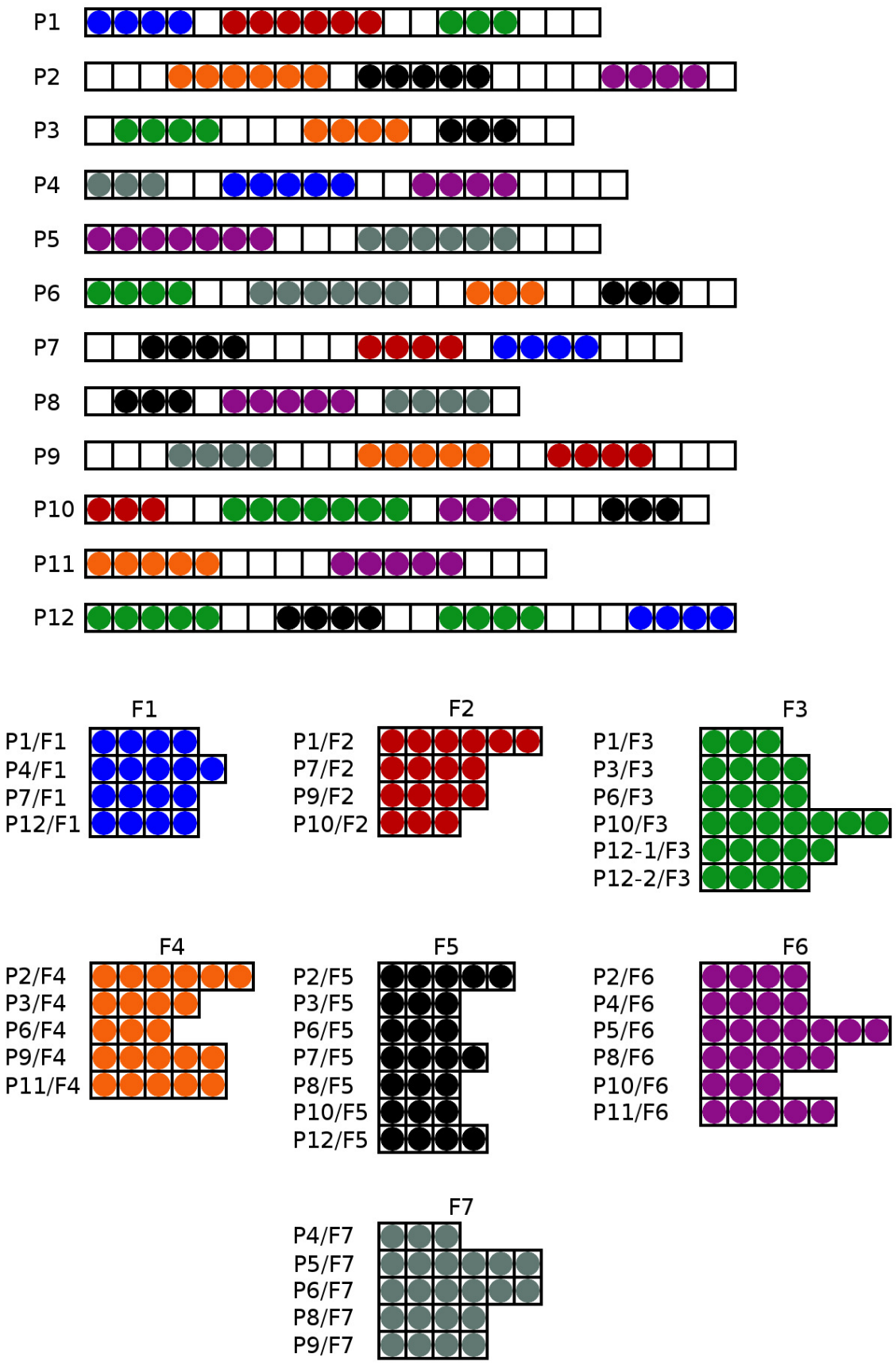


Figura 3.1: Exemplo de identificação de domínios em proteínas e formação de famílias de domínios.

Capítulo 4

O Espaço Amostral e sua Distribuição de Probabilidades

Para a análise proposta de verificação da distribuição de aminoácidos nas famílias de domínios de proteínas por coluna e por pares de colunas, restringimos nosso espaço amostral a clãs formados por famílias que contenham blocos representativos de m linhas (domínios) por n colunas (aminoácidos). Ou seja, blocos $(m \times n)$ são então recortados das famílias de domínios esquematizadas na Figura 3.1 do capítulo anterior. No presente trabalho, utilizamos a versão 27.0 do PFAM [20] e realizamos os trabalhos com dois diferentes blocos a fim de comparação: (100×100) e (100×200). Uma restrição adicional, não estritamente necessária, em relação à utilização de clãs que contenham ao menos cinco famílias que possam ser representadas por estes blocos também foi adotada.

A Figura 4.1 explicita como se dá a criacão dos blocos, com a remoção dos domínios (linhas) que contenham menos do que n aminoácidos e a exclusão dos aminoácidos sobressalentes quando um domínio tem mais do que os n aminoácidos. Uma família que não contenha um mínimo de m domínios com n aminoácidos é descartada, e um clã que não possua um mínimo de cinco famílias que possuam os blocos $(m \times n)$ também não é utilizado na criação do espaço amostral.

Dada uma coluna j, onde $j = 1, 2, \ldots, n$, a probabilidade de ocorrência do aminoácido a, de um total de 20 aminoácidos possíveis (a = A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y), é definida como

$$p_j(a) = \frac{n_j(a)}{m}\,, \quad \forall a\,, \tag{4.1}$$

sendo $n_j(a)$ a ocorrência total do aminoácido a na coluna j. Assim, pelos axiomas

Figura 4.1: Blocos $(m \times n)$ de aminoácidos representativos das famílias. Domínios com menos de n colunas são removidos (em vermelho) e os com mais de n colunas tem o excesso $(n+1$ em diante) removido (em azul).

de Kolmogorov [25], temos que

$$1 \geq p_j(a) \geq 0\,, \quad \forall a\,, \forall j\,, \tag{4.2}$$

e pela definição de probabilidade

$$\sum_{a=1}^{20} p_j(a) = \sum_{a=1}^{20} \frac{n_j(a)}{m} = \frac{1}{m} \sum_{a=1}^{20} n_j(a) = \frac{m}{m} = 1\,, \quad \forall j\,. \tag{4.3}$$

As probabilidades dos aminoácidos de cada uma das n colunas podem ser agrupadas em vetores coluna, da seguinte forma:

$$p_j = \begin{pmatrix} p_j(A) \\ p_j(C) \\ p_j(D) \\ \vdots \\ p_j(W) \\ p_j(Y) \end{pmatrix}\,, \tag{4.4}$$

resultando em n vetores p_j contendo 20 elementos (aminoácidos) cada.

Analogamente à definição de probabilidade de uma coluna (ou, probabilidade simples, como nos referiremos a ela daqui em diante), a probabilidade conjunta da ocorrência de um par de aminoácidos a, b em um par de colunas j, k pode ser definida como

$$p_{jk}(a,b) = \frac{n_{jk}(a,b)}{m}\,, \quad \forall a,b\,, \tag{4.5}$$

onde, $n_{jk}(a,b)$ é a ocorrência total do par de aminoácidos a,b no par de colunas j,k, com a,b = A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, $j = 1,2,\ldots,(n-1)$, $k = (j+1),(j+2),\ldots,n$. Temos para as probabilidades conjuntas que:

$$1 \geq p_{j,k}(a,b) \geq 0\,, \quad \forall a,b\,,\, \forall j,k\,, \tag{4.6}$$

$$\sum_{a=1}^{20}\sum_{b=1}^{20} p_{jk}(a,b) = \sum_{a=1}^{20}\sum_{b=1}^{20}\frac{n_{jk}(a,b)}{m} = \frac{1}{m}\sum_{a=1}^{20}\sum_{b=1}^{20} n_{jk}(a,b) = \frac{m}{m} = 1\,, \forall j,k\,. \tag{4.7}$$

Para um dado par de colunas j,k, as probabilidades conjuntas podem ser reunidas em matrizes com 400 elementos (20×20):

$$p_{jk} = \begin{pmatrix} p_{jk}(A,A) & p_{jk}(A,C) & p_{jk}(A,D) & \ldots & p_{jk}(A,W) & p_{jk}(A,Y) \\ p_{jk}(C,A) & p_{jk}(C,C) & p_{jk}(C,D) & \ldots & p_{jk}(C,W) & p_{jk}(C,Y) \\ p_{jk}(D,A) & p_{jk}(D,C) & p_{jk}(D,D) & \ldots & p_{jk}(D,W) & p_{jk}(D,Y) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{jk}(W,A) & p_{jk}(W,C) & p_{jk}(W,D) & \ldots & p_{jk}(W,W) & p_{jk}(W,Y) \\ p_{jk}(Y,A) & p_{jk}(Y,C) & p_{jk}(Y,D) & \ldots & p_{jk}(Y,W) & p_{jk}(Y,Y) \end{pmatrix} \tag{4.8}$$

Para um bloco $(m \times n)$, o número total de matrizes referente aos pares de colunas é igual a combinação das n colunas tomadas duas a duas:

$$\binom{n}{2} = \frac{n!}{2!\,(n-2)!} = \frac{n\cdot(n-1)}{2} \text{ pares de colunas}$$

Desta forma, para os blocos representativos 100×200 e 100×100, temos, respectivamente:

$$\binom{200}{2} = \frac{200\cdot 199}{2} = 19900 \text{ pares de colunas}$$

e

$$\binom{100}{2} = \frac{100\cdot 99}{2} = 4950 \text{ pares de colunas}$$

Por sua vez, as matrizes de probabilidades conjuntas podem ser organizadas em um

arranjo triangular, da seguinte forma:

$$P = \begin{array}{ccccccc} p_{12} & p_{13} & p_{14} & \cdots & p_{1\,n-2} & p_{1\,n-1} & p_{1\,n} \\ & p_{23} & p_{24} & \cdots & p_{2\,n-2} & p_{2\,n-1} & p_{2\,n} \\ & & p_{34} & \cdots & p_{3\,n-2} & p_{3\,n-1} & p_{3\,n} \\ & & & \ddots & \vdots & \vdots & \vdots \\ & & & & p_{n-3\,n-2} & p_{n-3\,n-1} & p_{n-3\,n} \\ & & & & & p_{n-2\,n-1} & p_{n-2\,n} \\ & & & & & & p_{n-1\,n} \end{array} \tag{4.9}$$

O número de ordem dessas matrizes na contagem da esquerda para a direita e de cima para baixo podem ser escritos como:

$$C_{jk} = j(n-1) - \frac{j(j-1)}{2} - (n-k) \tag{4.10}$$

Estes números também podem ser equivalentemente organizados em um arranjo triangular da seguinte forma:

$$C = \begin{array}{ccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & \ldots & (n-3) & (n-2) & (n-1) \\ & n & (n+1) & (n+2) & (n+3) & (n+4) & \ldots & (2n-5) & (2n-4) & (2n-3) \\ & & (2n-2) & (2n-1) & 2n & (2n+1) & \ldots & (3n-8) & (3n-7) & (3n-6) \\ & & & (3n-5) & (3n-4) & (3n-3) & \ldots & (4n-12) & (4n-11) & (4n-10) \\ & & & & (4n-9) & (4n-8) & \ldots & (5n-17) & (5n-16) & (5n-15) \\ & & & & & (5n-14) & \ldots & (6n-23) & (6n-22) & (6n-21) \\ & & & & & & \ddots & \vdots & \vdots & \vdots \\ & & & & & & & \frac{1}{2}(n^2-n-10) & \frac{1}{2}(n^2-n-8) & \frac{1}{2}(n+2)(n-3) \\ & & & & & & & & \frac{1}{2}(n^2-n-4) & \frac{1}{2}(n+1)(n-2) \\ & & & & & & & & & \frac{1}{2}n(n-1) \end{array} \tag{4.11}$$

Após esta apresentação dos conceitos de probabilidades simples e conjunta para a caracterização da distribuição dos aminoácidos nos blocos $(m \times n)$, o teorema de Bayes é então expresso como

$$p_{jk}(a|b) = \frac{p_{kj}(b|a)p_j(a)}{p_k(b)}\,, \tag{4.12}$$

onde, $p_{jk}(a|b)$ e $p_{kj}(b|a)$ 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. O teorema de Bayes relaciona as probabilidades *a priori* (probabilidades simples) com as probabilidades *a posteriori* (probabilidades condicionais) [26]. A probabilidade conjunta de um par de colunas j, k pode ser expressa em função das probabilidades condicionais e simples da seguinte forma:

$$p_{jk}(a,b) = \underbrace{p_{jk}(a|b)p_k(b) = p_{kj}(b|a)p_j(a)}_{\text{Bayes}} = p_{kj}(b,a) \tag{4.13}$$

Caso duas colunas sejam independentes entre si, ou seja, não havendo um vínculo entre a ocorrência dos aminoácidos em uma destas colunas em relação a ocorrência dos aminoácidos na outra, as probabilidades condicionais são iguais às probabilidades simples. Logo, a probabilidade conjunta do par de colunas j, k, para este caso, fica caracterizada como o produto entre as probabilidades simples de cada uma das colunas:

$$p_{jk}(a,b) = p_j(a)p_k(b)\,. \tag{4.14}$$

Como dito anteriormente, cada linha do bloco representativo de uma família é originária de um domínio. A conformação adotada pelo domínio está relacionada com a sua sequência de aminoácidos, de forma que é natural supor que haja uma dependência entre as ocorrências dos aminoácidos nos pares de colunas. Portanto, caso haja pares de colunas que sejam independentes, estes devem ser reconhecidos através da análise da distribuição dos aminoácidos e não por uma suposição *a priori* de independência.

Após as restrições impostas de famílias com blocos (100×200) e clãs constituidos com um mínimo de 5 famílias (Tabela 4.1), restam apenas 1069 famílias como elementos do espaço amostral para o teste estatístico. Já para a restrição de blocos (100×100), após a restrição de clãs contendo um mínimo de 5 famílias (Tabela 4.2), nos restam apenas 2180 famílias.

Tabela 4.1: Restrições adotadas para a análise estatística ANOVA.

Restrições	nº de famílias class. em 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 4.2: Restrições adotadas para a análise estatística ANOVA.

Restrições	nº de famílias class. em clãs	nº de clãs
nenhuma	4563	515
blocos 100×100, um bloco por família	2525	393
clãs com 5 ou mais famílias	2180	146

Capítulo 5

Medidas de Entropia – Famílias Sharma-Mittal e a Entropia de Jaccard

Após a introdução dos blocos $(m \times n)$ representativos das famílias de domínios de proteínas, desejamos saber a distribuição dos aminoácidos ao longo das colunas e dos pares de colunas. A adoção de medidas de entropia é de grande valia uma vez que quantificam a incerteza das distribuições de probabilidade. Assim, teremos uma grandeza escalar por coluna ou por par de colunas ao invés de um vetor ou uma matriz de probabilidades, respectivamente. Neste capítulo introduzimos algumas destas medidas presentes na literatura que serão utilizadas neste trabalho. A entropia Sharma-Mittal [27] de dois parâmetros é definida para as distribuições de probabilidade simples e probabilidade conjunta, respectivamente, como:

$$(SM)_j(r,s) = -\frac{1}{1-r}\left(1 - \left(\sum_a \Big(p_j(a)\Big)^s\right)^{\frac{1-r}{1-s}}\right) \tag{5.1}$$

$$(SM)_{jk}(r,s) = -\frac{1}{1-r}\left(1 - \left(\sum_a \sum_b \Big(p_{jk}(a,b)\Big)^s\right)^{\frac{1-r}{1-s}}\right) \tag{5.2}$$

onde, r e s são parâmetros adimensionais.

A partir da entropia de Sharma-Mittal podemos obter um conjunto de entropias contendo apenas um parâmetro e a entropia de Shannon, livre de parâmetros, ao tomarmos certos limites. É importante salientar que tanto nas expressões da entropia de Sharma-Mittal acima quanto nas que serão apresentadas a seguir, uma constante dimensional c é suprimida sem perda de generalidade, uma vez que para nossas aplicações podemos considerá-las como adimensionais. Desta forma, todas as grandezas aqui tratadas são consideradas sem dimensões. A entropia, como definida

em Termodinâmica, é uma grandeza física de energia dividida por temperatura (M L^2 T^{-2}/Θ), enquanto que na teoria de comunicações é tratada como adimensional, e geralmente medida em bits.

A entropia de Havrda-Charvat [28] é obtida a partir da entropia de Sharma-Mittal ao tomarmos o limite do parâmetro r tendendo a s. Para as distribuições de probabilidade simples e conjunta temos, respectivamente:

$$(HC)_j(s) = \lim_{r\to s}(SM)_j(r,s) = -\frac{1}{1-s}\left(1 - \sum_a \Big(p_j(a)\Big)^s\right) \tag{5.3}$$

$$(HC)_{jk}(s) = \lim_{r\to s}(SM)_{jk}(r,s) = -\frac{1}{1-s}\left(1 - \sum_a\sum_b \Big(p_{jk}(a,b)\Big)^s\right) \tag{5.4}$$

A entropia de Landsberg-Vedral [29], uma normalização da Havrda-Charvat, é obtida da seguinte forma:

$$\begin{aligned}(LV)_j(s) = \quad \lim_{r\to 2-s}(SM)_j(r,s) &= -\frac{1}{1-s}\left(\frac{1 - \sum_a \Big(p_j(a)\Big)^s}{\sum_a \Big(p_j(a)\Big)^s}\right) \\ &= \frac{(HC)_j(s)}{\sum_a \Big(p_j(a)\Big)^s}\end{aligned} \tag{5.5}$$

$$\begin{aligned}(LV)_{jk}(s) = \quad \lim_{r\to 2-s}(SM)_{jk}(r,s) &= -\frac{1}{1-s}\left(\frac{1 - \sum_a\sum_b \Big(p_{jk}(a,b)\Big)^s}{\sum_a\sum_b \Big(p_{jk}(a,b)\Big)^s}\right) \\ &= \frac{(HC)_{jk}(s)}{\sum_a\sum_b \Big(p_{jk}(a,b)\Big)^s}\end{aligned} \tag{5.6}$$

A medida de entropia de Renyi [30] é obtida tomando o limite de r tendendo a 1:

$$R_j(s) = \lim_{r\to 1}(SM)_j(r,s) = \frac{1}{1-s}\ln\left(\sum_a \Big(p_j(a)\Big)^s\right) \tag{5.7}$$

$$R_{jk}(s) = \lim_{r\to 1}(SM)_{jk}(r,s) = \frac{1}{1-s}\ln\left(\sum_a\sum_b \Big(p_{jk}(a,b)\Big)^s\right) \tag{5.8}$$

A entropia Shannon [31–33], por sua vez, surge através do limite do parâmetro s tendendo a 1 de qualquer uma das três entropias de um único parâmetro definidas

acima:

$$\lim_{s\to 1}(HC)_j(s) = \lim_{s\to 1}(LV)_j(s) = \lim_{s\to 1} R_j(s) = S_j \tag{5.9}$$

$$\lim_{s\to 1}(HC)_{jk}(s) = \lim_{s\to 1}(LV)_{jk}(s) = \lim_{s\to 1} R_{jk}(s) = S_{jk} \tag{5.10}$$

onde,

$$S_j = -\sum_a p_j(a)\ln\big(p_j(a)\big) \tag{5.11}$$

$$S_{jk} = -\sum_a\sum_b p_{jk}(a,b)\ln\big(p_{jk}(a,b)\big) \tag{5.12}$$

Na Figura 5.1, temos os gráficos destas entropias de um parâmetro para um exemplo em que existam eventos dicotômicos com probabilidades $p_1 = p$ e $p_2 = (1-p)$. As Figuras 5.1(a), (c) e (e) contêm as superfícies das entropias de Havrda-Charvat, Landsberg-Vedral e Renyi, respectivamente. A curva destacada em preto sobre as superfícies corresponde à entropia de Shannon. As Figuras 5.1 (b), (d) e (f) apresentam os cortes longitudinais feitos nas superfícies correspondentes, com os valores do parâmetro s iguais a 0.1, 0.5, 0.9, 1 (Shannon), 1.1, 1.5 e 1.9.

A entropia de uma distribuição de probabilidade conjunta pode ser expressa por uma relação entre as entropias das distribuições de probabilidade simples da seguinte forma:

$$\begin{aligned}
(SM)_{jk}(r,s) &= -\frac{1}{1-r}\left(1-\left(\sum_a\sum_b\Big(p_{jk}(a,b)\Big)^s\right)^{\frac{1-r}{1-s}}\right)\\
&= -\frac{1}{1-r}\left(1-\left(\sum_a\sum_b\Big(p_j(a)\Big)^s\Big(p_{kj}(b|a)\Big)^s\right)^{\frac{1-r}{1-s}}\right)\\
&= -\frac{1}{1-r}\left(1-\left(\sum_c\Big(p_j(c)\Big)^s\sum_a\sum_b\hat{p}_j(a)\Big(p_{kj}(b|a)\Big)^s\right)^{\frac{1-r}{1-s}}\right)\\
&= -\frac{1}{1-r}\left(1-\left(\sum_c\Big(p_j(c)\Big)^s\right)^{\frac{1-r}{1-s}}\left(\sum_a\sum_b\hat{p}_j(a)\Big(p_{kj}(b|a)\Big)^s\right)^{\frac{1-r}{1-s}}\right)\\
&= -\frac{1}{1-r}\Big(1-\Big(1+(1-r)(SM)_j(r,s)\Big)\Big(1+(1-r)(SM)_{k|j}(r,s)\Big)\Big)
\end{aligned}$$

$$(SM)_{jk}(r,s) = (SM)_j(r,s)+(SM)_{k|j}(r,s)+(1-r)(SM)_j(r,s)(SM)_{k|j}(r,s) \tag{5.13}$$

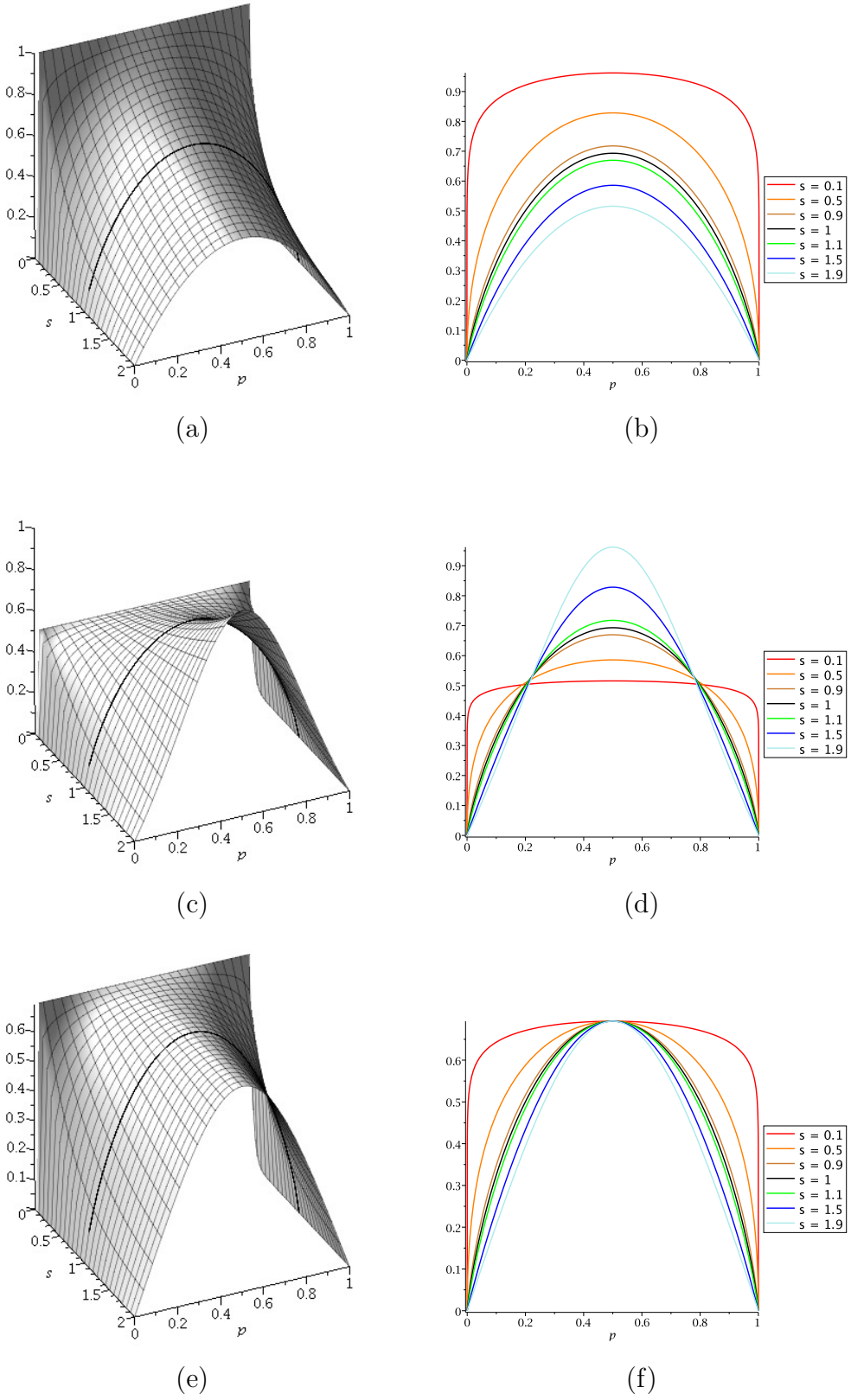


Figura 5.1: Gráficos das entropias de um parâmetro. Para um exemplo com $p_1 = p$ e $p_2 = (1 - p)$, as superfícies e cortes longitudinais, respectivamente, para Havrda-Charvat (a, b), Landsberg-Vedral (c, d) e Renyi (e, f). As curvas correspondentes ao limite Shannon ($s = 1$) são representadas em preto em todas as figuras.

onde

$$\hat{p}_j(a) \equiv \frac{\Big(p_j(a)\Big)^s}{\sum\limits_c \Big(p_j(c)\Big)^s} \,, \tag{5.14}$$

$$\sum_a \hat{p}_j(a) = 1 \,,$$

$$\hat{p}_j(a) \leq 1 \quad \forall a \,,$$

com $c =$ A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, e

$$(SM)_{k|j}(r,s) \equiv -\frac{1}{1-r}\left(1 - \left(\sum_a \sum_b \hat{p}_j(a)\Big(p_{kj}(b|a)\Big)^s\right)^{\frac{1-r}{1-s}}\right) , \tag{5.15}$$

é a entropia associada à probabilidade condicional $p_{kj}(b|a)$.

De acordo com Khinchin [32], para a entropia Shannon temos que $S_{k|j} \leq S_k$, uma vez que o conhecimento *a priori* do ocorrido na coluna j fornece informação ao evento da coluna k associada e portanto diminui a incerteza sobre a coluna k. Verificamos então se esta relação pode ser estendida para a entropia de Sharma-Mittal:

$$(SM)_{k|j}(r,s) \overset{?}{\leq} (SM)_k(r,s)$$

$$\frac{-1}{1-r}\left(1 - \left(\sum_a \sum_b \hat{p}_j(a)\Big(p_{kj}(b|a)\Big)^s\right)^{\frac{1-r}{1-s}}\right) \leq \frac{-1}{1-r}\left(1 - \left(\sum_b \Big(p_k(b)\Big)^s\right)^{\frac{1-r}{1-s}}\right)$$

Para $1 > r \geq s$:

$$\sum_a \sum_b \hat{p}_j(a)\Big(p_{kj}(b|a)\Big)^s \leq \sum_b \Big(p_k(b)\Big)^s$$

$$\sum_a \sum_b \left(\frac{\Big(p_j(a)\Big)^s}{\sum\limits_c \Big(p_j(c)\Big)^s}\right)\Big(p_{kj}(b|a)\Big)^s \leq \sum_b \Big(p_k(b)\Big)^s$$

$$\sum_a \sum_b \Big(p_j(a)\Big)^s \Big(p_{kj}(b|a)\Big)^s \leq \sum_c \Big(p_j(c)\Big)^s \sum_b \Big(p_k(b)\Big)^s$$

$$\sum_a \sum_b \Big(p_{jk}(a,b)\Big)^s \leq \sum_c \sum_b \Big(p_j(c)p_k(b)\Big)^s$$

Como a e c podem ser iguais aos mesmos elementos do mesmo domínio, temos que:

$$\sum_a \sum_b \Big(p_{jk}(a,b)\Big)^s \leq \sum_a \sum_b \Big(p_j(a)p_k(b)\Big)^s$$

ocorrendo a igualdade quando os eventos da coluna j e k forem independentes. Assim, confirmamos que a entropia condicional é limitada pela entropia simples.

A equação (5.13) pode ser então limitada através da desigualdade:

$$(SM)_{jk}(r,s) \leq (SM)_j(r,s) + (SM)_k(r,s) + (1-r)(SM)_j(r,s)(SM)_k(r,s) \quad (5.16)$$

Tomando os devidos limites temos para as entropias de Havrda-Charvat, Landsberg-Vedral, Renyi e Shannon, respectivamente:

$$(HC)_{jk}(s) \leq (HC)_j(s) + (HC)_k(s) + (1-s)(HC)_j(s)(HC)_k(s) \ , \quad (5.17)$$

$$(LV)_{jk}(s) \leq (LV)_j(s) + (LV)_k(s) - (1-s)(LV)_j(s)(LV)_k(s) \ , \quad (5.18)$$

$$R_{jk}(s) \leq R_j(s) + R_k(s) \ , \quad (5.19)$$

e

$$S_{jk} \leq S_j + S_k \quad (5.20)$$

Como mencionado no capítulo anterior, consideramos que há um tipo de vínculo na ocorrência dos aminoácidos nos pares de colunas. Uma forma de medir esta dependência é através do cálculo da Informação Mútua, que pode ser definida em função das medidas de entropia. Então, para as entropias previamente introduzidas, temos:

- Sharma-Mittal

$$M_{jk}^{(SM)}(r,s) = \frac{1}{1-r}\left(1 - \left(\frac{\sum\limits_a \sum\limits_b \Big(p_{jk}(a,b)\Big)^s}{\sum\limits_a \sum\limits_b \Big(p_j(a)p_k(b)\Big)^s}\right)^{\frac{1-r}{1-s}}\right) \quad (5.21)$$

- Havrda-Charvat

$$M_{jk}^{(HC)}(s) = \lim_{r\to s} M_{jk}^{(SM)}(r,s) = \frac{1}{1-s}\left(1 - \frac{\sum\limits_a \sum\limits_b \Big(p_{jk}(a,b)\Big)^s}{\sum\limits_a \sum\limits_b \Big(p_j(a)p_k(b)\Big)^s}\right) \quad (5.22)$$

- Landsberg-Vedral

$$M_{jk}^{(LV)}(s) = \lim_{r\to 2-s} M_{jk}^{(SM)}(r,s) = -\frac{1}{1-s}\left(1 - \frac{\sum\limits_a \sum\limits_b \Big(p_j(a)p_k(b)\Big)^s}{\sum\limits_a \sum\limits_b \Big(p_{jk}(a,b)\Big)^s}\right) \quad (5.23)$$

$$= M_{jk}^{(HC)}(s) \cdot \frac{\sum\limits_a \sum\limits_b \Big(p_j(a)p_k(b)\Big)^s}{\sum\limits_a \sum\limits_b \Big(p_{jk}(a,b)\Big)^s}$$

- Renyi

$$M_{jk}^{(R)}(s) = \lim_{r\to 1} M_{jk}^{(SM)}(r,s) = -\frac{1}{1-s} \ln \left(\frac{\sum\limits_a \sum\limits_b \Big(p_{jk}(a,b)\Big)^s}{\sum\limits_a \sum\limits_b \Big(p_j(a)p_k(b)\Big)^s} \right) \tag{5.24}$$

- Shannon

$$M_{jk} = \lim_{s\to 1} M_{jk}^{(HC)}(s) = \lim_{s\to 1} M_{jk}^{(LV)}(s) = \lim_{s\to 1} M_{jk}^{(R)}(s)$$

$$= \sum_a \sum_b p_{jk}(a,b) \ln\big(p_{jk}(a,b)\big) - \sum_a \sum_b p_j(a)p_k(b) \ln\big(p_j(a)p_k(b)\big)$$

$$= -S_{jk} + S_j + S_k \tag{5.25}$$

A medida de Informação Mútua da entropia de Shannon, equação (5.25), pode também ser obtida através da divergência de Kullback-Leibler [34, 35], que é definida como:

$$(KL)_{jk}(b) = \sum_a p_{jk}(a|b) \log \left(\frac{p_{jk}(a|b)}{p_j(a)} \right) \tag{5.26}$$

Então temos que:

$$(KL)_{jk}(b) = \sum_a \frac{p_{jk}(a,b)}{p_k(b)} \log \left(\frac{p_{jk}(a,b)}{p_j(a)p_k(b)} \right) \tag{5.27}$$

e a Informação Mútua M_{jk} é dada por:

$$M_{jk} = \sum_a p_k(b)(KL)_{jk}(b) = \sum_a \sum_b p_{jk}(a,b) \log \left(\frac{p_{jk}(a,b)}{p_j(a)p_k(b)} \right) \tag{5.28}$$

que é o mesmo que a equação (5.25), c.q.d.

É fácil notar que quando um par de colunas for independente, o valor da Informação Mútua será igual a zero, porém pode haver uma combinação tal que a soma sobre todos os pares de aminoácidos da potência da probabilidade conjunta seja igual ao produto das somas de potência das probabilidades simples sem que exista necessariamente uma independência entre os pares de colunas [2].

Podemos agora introduzir o conceito de Distância de Informação, que é a diferença entre o valor de entropia de um par de colunas e o valor de Informação Mútua do mesmo par:

$$d_{jk}^{(SM)}(r,s) = (SM)_{jk}(r,s) - M_{jk}^{(SM)}(r,s)\,, \tag{5.29}$$

escrita de forma genérica em função da entropia de Sharma-Mittal, sendo as outras obtidas facilmente através dos limites correspondentes:

- Havrda-Charvat

$$d_{jk}^{(HC)}(s) = (HC)_{jk}(s) - M_{jk}^{(HC)}(s) \tag{5.30}$$

- Landsberg-Vedral

$$d_{jk}^{(LV)}(s) = (LV)_{jk}(s) - M_{jk}^{(LV)}(s) \tag{5.31}$$

- Renyi

$$d_{jk}^{(R)}(s) = R_{jk}(s) - M_{jk}^{(R)}(s) \tag{5.32}$$

- Shannon

$$d_{jk} = S_{jk} - M_{jk} \tag{5.33}$$

É fácil notar que para o caso em que duas colunas são independentes entre si, sua Informação Mútua é nula, sendo portanto a Distância de Informação igual ao valor de entropia. Assim sendo,

$$d_{jk}^{(SM)}(r,s) \geq (SM)_{jk}(r,s)\,. \tag{5.34}$$

Além disso, como estamos trabalhando com medidas de entropia e com a noção de distância, temos que atender os seguintes requisitos:

$$\begin{cases} (SM)_{jk}(r,s) \geq 0 \\ \\ M_{jk}^{(SM)}(r,s) \geq 0 \\ \\ d_{jk}^{(SM)}(r,s) \geq 0 \end{cases} \tag{5.35}$$

Desta forma, se para determinados valores de parâmetros alguma destas desigualdades for violada, restrições devem ser adotadas em relação à descrição do banco de dados pela medida de entropia utilizada. As equações (5.34) e (5.35) são então sintetizadas como:

$$0 \leq d_{jk}(r,s) = (SM)_{jk}(r,s) - M_{jk}^{(SM)}(r,s) \leq (SM)_{jk}(r,s) \tag{5.36}$$

Podemos agora apresentar a Entropia de Jaccard[36], obtida através da normalização da Distância de Informação:

$$J_{jk}^{(SM)}(r,s) = \frac{d_{jk}^{(SM)}(r,s)}{(SM)_{jk}(r,s)} \tag{5.37}$$

$$J_{jk}^{(SM)}(r,s) = 1 - \frac{M_{jk}^{(SM)}(r,s)}{(SM)_{jk}(r,s)} \tag{5.38}$$

escrita em função da entropia de Sharma-Mittal. Para as entropias restantes temos:

- Havrda-Charvat

$$J_{jk}^{(HC)}(s) = 1 - \frac{M_{jk}^{(HC)}(s)}{(HC)_{jk}(s)} \tag{5.39}$$

- Landsberg-Vedral

$$J_{jk}^{(LV)}(s) = 1 - \frac{M_{jk}^{(LV)}(s)}{(LV)_{jk}(s)} \tag{5.40}$$

$$= J_{jk}^{(HC)}(s) \cdot \sum_a \sum_b \Big(p_j(a)p_k(b)\Big)^s + 1 - \sum_a \sum_b \Big(p_j(a)p_k(b)\Big)^s$$

- Renyi

$$J_{jk}^{(R)}(s) = 1 - \frac{M_{jk}^{(R)}(s)}{R_{jk}(s)} \tag{5.41}$$

- Shannon

$$J_{jk} = 1 - \frac{M_{jk}}{S_{jk}} \tag{5.42}$$

$$= 2 - \frac{S_j + S_k}{S_{jk}}$$

Pela equação (5.36) temos que a entropia de Jaccard é definida para valores de parâmetros em que seja obedecida a desigualdade:

$$0 \leq J_{jk}^{(SM)}(r,s) \leq 1 \tag{5.43}$$

Os valores do parâmetro s correspondentes a uma Distância de Informação negativa não levam a uma caracterização útil obtida da Entropia de Jaccard, e portanto não são levados em consideração.

Um estudo do comportamento das curvas de Jaccard, foi feito utilizando distribuições hipotéticas de apenas dois aminoácidos (A e C) em um par de colunas. Os cálculos foram realizados com a Jaccard associada a entropia de Havrda-Charvat (eq. (5.39)). Os seguintes casos foram analizados:

(a) Distribuição dos aminoácidos na primeira coluna: $p_1(A) = 0.5$, $p_1(C) = 0.5$.

Distribuição dos aminoácidos na segunda coluna: $p_2(A) = 0.5$, $p_2(C) = 0.5$.

Configurações possíveis das probabilidades conjuntas:

I)

$$\begin{cases} p_{12}(A,A) = 0.5 \\ p_{12}(A,C) = 0 \\ p_{12}(C,A) = 0 \\ p_{12}(C,C) = 0.5 \end{cases}$$

II)

$$\begin{cases} p_{12}(A,A) = 0.4 \\ p_{12}(A,C) = 0.1 \\ p_{12}(C,A) = 0.1 \\ p_{12}(C,C) = 0.4 \end{cases}$$

III)

$$\begin{cases} p_{12}(A,A) = 0.3 \\ p_{12}(A,C) = 0.2 \\ p_{12}(C,A) = 0.2 \\ p_{12}(C,C) = 0.3 \end{cases}$$

IV)

$$\begin{cases} p_{12}(A,A) = 0.2 \\ p_{12}(A,C) = 0.3 \\ p_{12}(C,A) = 0.3 \\ p_{12}(C,C) = 0.2 \end{cases}$$

V)

$$\begin{cases} p_{12}(A,A) = 0.1 \\ p_{12}(A,C) = 0.4 \\ p_{12}(C,A) = 0.4 \\ p_{12}(C,C) = 0.1 \end{cases}$$

VI)

$$\begin{cases} p_{12}(A,A) = 0 \\ p_{12}(A,C) = 0.5 \\ p_{12}(C,A) = 0.5 \\ p_{12}(C,C) = 0 \end{cases}$$

(b) Distribuição dos aminoácidos na primeira coluna: $p_1(A) = 0.6$, $p_1(C) = 0.4$

Distribuição dos aminoácidos na segunda coluna: $p_2(A) = 0.4$, $p_2(C) = 0.6$

Configurações possíveis das probabilidades conjuntas:

I)

$$\begin{cases} p_{12}(A,A) = 0.4 \\ p_{12}(A,C) = 0.2 \\ p_{12}(C,A) = 0 \\ p_{12}(C,C) = 0.4 \end{cases}$$

II)

$$\begin{cases} p_{12}(A,A) = 0.3 \\ p_{12}(A,C) = 0.3 \\ p_{12}(C,A) = 0.1 \\ p_{12}(C,C) = 0.3 \end{cases}$$

III)

$$\begin{cases} p_{12}(A,A) = 0.2 \\ p_{12}(A,C) = 0.4 \\ p_{12}(C,A) = 0.2 \\ p_{12}(C,C) = 0.2 \end{cases}$$

IV)

$$\begin{cases} p_{12}(A,A) = 0.1 \\ p_{12}(A,C) = 0.5 \\ p_{12}(C,A) = 0.3 \\ p_{12}(C,C) = 0.1 \end{cases}$$

V)

$$\begin{cases} p_{12}(A,A) = 0 \\ p_{12}(A,C) = 0.6 \\ p_{12}(C,A) = 0.4 \\ p_{12}(C,C) = 0 \end{cases}$$

(c) Distribuição dos aminoácidos na primeira coluna: $p_1(A) = 0.7$, $p_1(C) = 0.3$

Distribuição dos aminoácidos na segunda coluna: $p_2(A) = 0.3$, $p_2(C) = 0.7$

Configurações possíveis das probabilidades conjuntas:

I)

$$\begin{cases} p_{12}(A,A) = 0.3 \\ p_{12}(A,C) = 0.4 \\ p_{12}(C,A) = 0 \\ p_{12}(C,C) = 0.3 \end{cases}$$

II)

$$\begin{cases} p_{12}(A,A) = 0.2 \\ p_{12}(A,C) = 0.5 \\ p_{12}(C,A) = 0.1 \\ p_{12}(C,C) = 0.2 \end{cases}$$

III)

$$\begin{cases} p_{12}(A,A) = 0.1 \\ p_{12}(A,C) = 0.6 \\ p_{12}(C,A) = 0.2 \\ p_{12}(C,C) = 0.1 \end{cases}$$

IV)

$$\begin{cases} p_{12}(A,A) = 0 \\ p_{12}(A,C) = 0.7 \\ p_{12}(C,A) = 0.3 \\ p_{12}(C,C) = 0 \end{cases}$$

(d) Distribuição dos aminoácidos na primeira coluna: $p_1(A) = 0.8$, $p_1(C) = 0.2$

Distribuição dos aminoácidos na segunda coluna: $p_2(A) = 0.2$, $p_2(C) = 0.8$

Configurações possíveis das probabilidades conjuntas:

I)

$$\begin{cases} p_{12}(A,A) = 0.2 \\ p_{12}(A,C) = 0.6 \\ p_{12}(C,A) = 0 \\ p_{12}(C,C) = 0.2 \end{cases}$$

II)

$$\begin{cases} p_{12}(A,A) = 0.1 \\ p_{12}(A,C) = 0.7 \\ p_{12}(C,A) = 0.1 \\ p_{12}(C,C) = 0.1 \end{cases}$$

III)

$$\begin{cases} p_{12}(A,A) = 0 \\ p_{12}(A,C) = 0.8 \\ p_{12}(C,A) = 0.2 \\ p_{12}(C,C) = 0 \end{cases}$$

(e) Distribuição dos aminoácidos na primeira coluna: $p_1(A) = 0.9$, $p_1(C) = 0.1$

Distribuição dos aminoácidos na segunda coluna: $p_2(A) = 0.1$, $p_2(C) = 0.9$

Configurações possíveis das probabilidades conjuntas:

I)

$$\begin{cases} p_{12}(A,A) = 0 \\ p_{12}(A,C) = 0.9 \\ p_{12}(C,A) = 0.1 \\ p_{12}(C,C) = 0 \end{cases}$$

II)

$$\begin{cases} p_{12}(A,A) = 0.1 \\ p_{12}(A,C) = 0.8 \\ p_{12}(C,A) = 0 \\ p_{12}(C,C) = 0.1 \end{cases}$$

(f) Distribuição dos aminoácidos na primeira coluna: $p_1(A) = 1$, $p_1(C) = 0$

Distribuição dos aminoácidos na segunda coluna: $p_2(A) = 0$, $p_2(C) = 1$

Configurações possíveis das probabilidades conjuntas:

I)

$$\begin{cases} p_{12}(A,A) = 0 \\ p_{12}(A,C) = 1 \\ p_{12}(C,A) = 0 \\ p_{12}(C,C) = 0 \end{cases}$$

Os gráficos de entropia Jaccard de Havrda-Charvat dos cinco primeiros casos são mostrados na figura (5.2). O sexto caso não é mostrado, uma vez que para esta distribuição de probabilidades a entropia Jaccard é igual a 0 para qualquer valor do parâmetro s. Para o primeiro caso, apesar das seis possíveis distribuições, temos três duplas semelhantes (I e VI, II e V, III e IV), o que resulta em curvas iguais para as duplas, como pode ser observado no primeiro gráfico.

Os resultados obtidos a partir destas curvas indicam que para trabalharmos com a entropia Jaccard associada à entropia de Havrda-Charvat, devemos trabalhar com o parâmetro no intervalo $0 < s \leq 1$. Na Figura 5.3 apresentamos as curvas de médias de entropia Jaccard associada à entropia de Havrda-Charvat de nove famílias pertencentes ao banco de dados PFAM v27.0 para os cortes 100×200. Como esperado, as curvas obtidas foram similares aos exemplos apresentados na Figura 5.2. Para cada valor do parâmetro, as médias sobre os pares de colunas são calculadas da seguinte forma:

$$(HC)(s) = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} (HC)_{jk}(s) \tag{5.44}$$

$$J^{(HC)}(s) = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} J_{jk}^{(HC)}(s) \tag{5.45}$$

As Tabelas 5.1, 5.2 e 5.3 a seguir, ajudam a esclarecer a restrição nos valores do parâmetro s na classificação das famílias e clãs do banco de dados PFAM. Já de antemão deve ser informado que a restrição $0 < s \leq 1$ garante que a entropia de Havrda-Charvat, a Informação Mútua e a Distância de Informação terão valores não negativos. Uma vez que a Distância de Informação é não negativa, a entropia de Jaccard também é.

Um segundo estudo foi feito em relação às distribuições no banco de dados dos valores das entropias de Havrda-Charvat de probabilidade conjunta e a de Jaccard associada. Para isso foram utilizadas as médias sobre os pares de colunas (equações (5.44) e (5.45)).

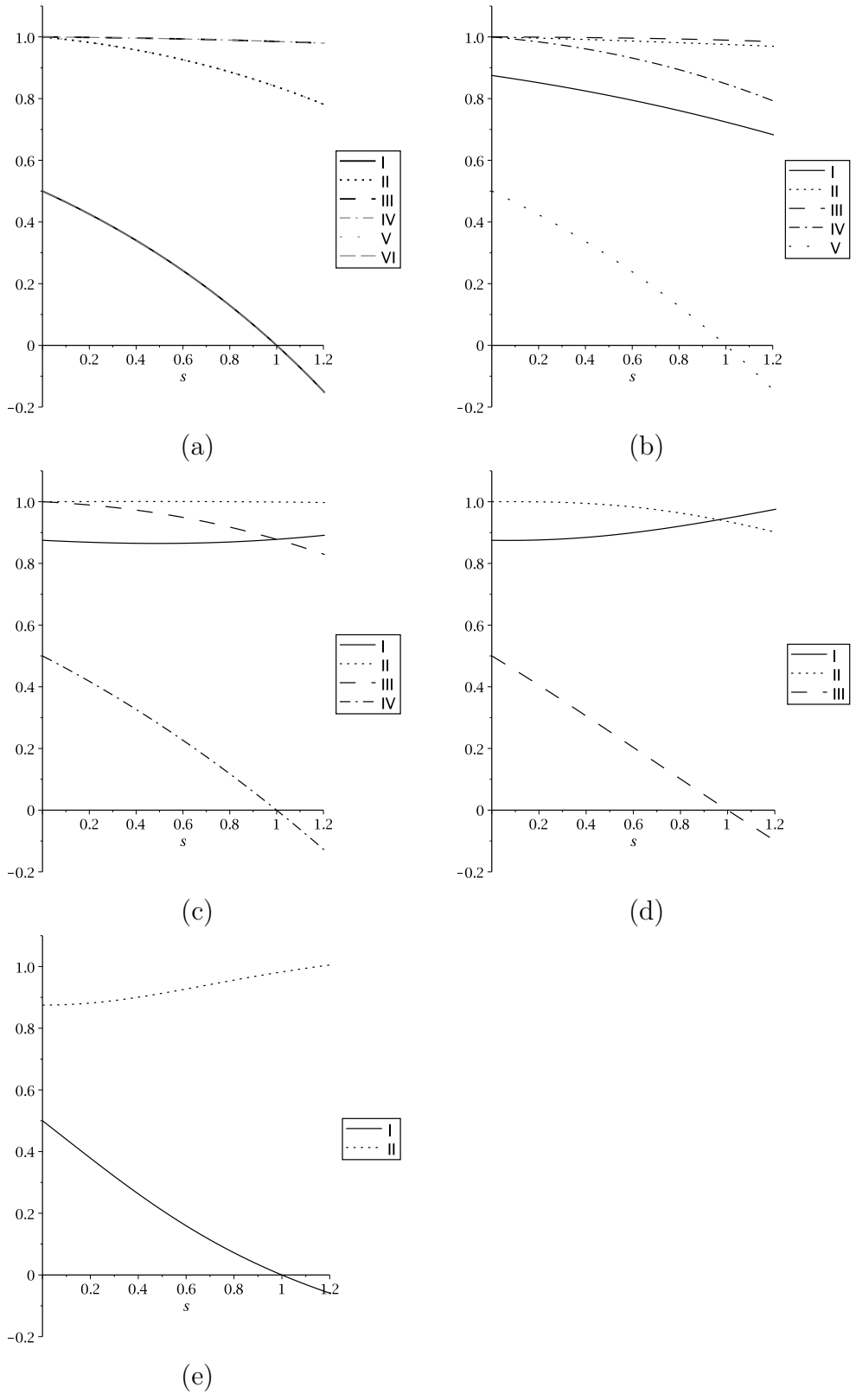


Figura 5.2: Curvas de entropia Jaccard de Havrda-Charvat contra o parâmetro s. O primeiro gráfico (a), referente ao primeiro caso apresenta apenas três curvas, IV, V e VI, que sobrepõem as curvas III, II e I, respectivamente.

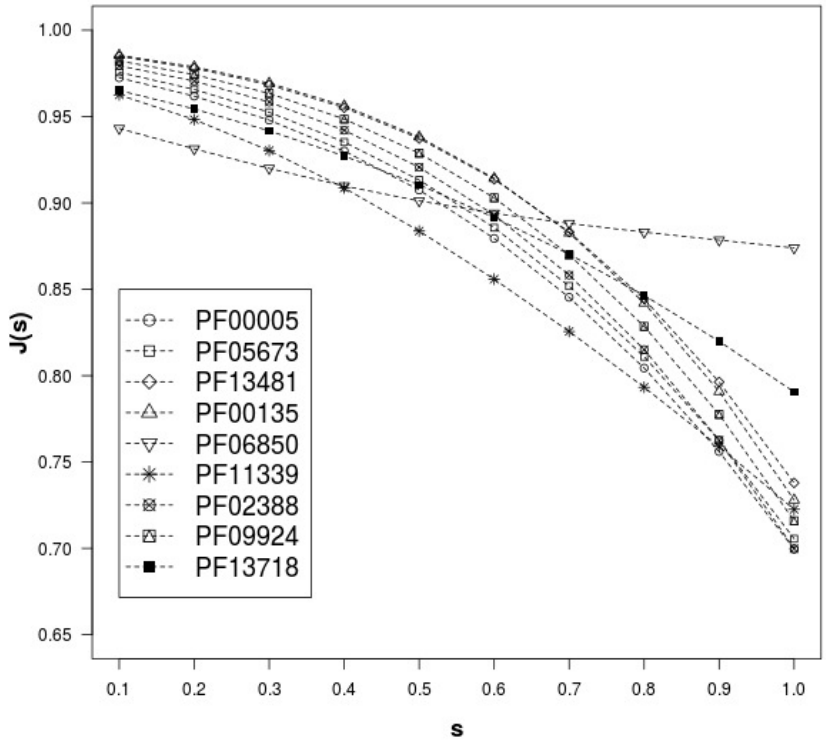


Figura 5.3: Curvas das médias de entropia Jaccard associada a Havrda-Charvat contra o parâmetro s. Famílias PF00005, PF05673 e PF13481 pertencentes ao clã CL0023, famílias PF00135, PF06850 e PF11339 pertencentes ao clã CL0028 e famílias PF02388, PF09924 e PF13718 pertencentes ao clã CL0257.

Tabela 5.1: Estudo de valores admissíveis da entropia de Havrda-Charvat, Informação Mútua e Distância de Informação para os blocos representativos 100×200 das famílias PF00005 e PF13481, pertencentes ao clã CL0023.

	PF00005			PF13481		
s	$H_{jk}(s)$	$M_{jk}(s)$	$d_{jk}(s)$	$H_{jk}(s)$	$M_{jk}(s)$	$d_{jk}(s)$
0.1	0	0	0	0	0	0
0.3	0	0	0	0	0	0
0.5	0	0	0	0	0	0
0.7	0	0	0	0	0	0
0.9	0	0	0	0	0	0
1.0	0	0	0	0	0	0
1.2	0	8	5	0	0	0
1.5	0	33	4741	0	0	193
1.7	0	55	9679	0	0	10120
1.9	0	65	12442	0	0	16317
2.0	0	69	13203	0	0	17485

As Figuras (5.4) e (5.5) apresentam histogramas das médias de Havrda-Charvat e da Jaccard associada para alguns valores do parâmetro s, para os casos com blocos representativos (100×200) e (100×100), respectivamente. Todos os histogramas

Tabela 5.2: Estudo de valores admissíveis da entropia de Havrda-Charvat, Informação Mútua e Distância de Informação para os blocos representativos 100×200 das famílias PF00135 e PF06850, pertencentes ao clã CL0028.

	PF00135			PF06850		
s	$H_{jk}(s)$	$M_{jk}(s)$	$d_{jk}(s)$	$H_{jk}(s)$	$M_{jk}(s)$	$d_{jk}(s)$
0.1	0	0	0	0	0	0
0.3	0	0	0	0	0	0
0.5	0	0	0	0	0	0
0.7	0	0	0	0	0	0
0.9	0	0	0	0	0	0
1.0	0	0	0	0	0	0
1.2	0	0	0	0	718	16
1.5	0	0	467	0	1708	38
1.7	0	0	14509	0	2351	61
1.9	0	0	19026	0	2898	192
2.0	0	0	19451	0	3139	309

Tabela 5.3: Estudo de valores admissíveis da entropia de Havrda-Charvat, Informação Mútua e Distância de Informação para os blocos representativos 100×200 das famílias PF02388 e PF09924, pertencentes ao clã CL0257.

	PF02388			PF09924		
s	$H_{jk}(s)$	$M_{jk}(s)$	$d_{jk}(s)$	$H_{jk}(s)$	$M_{jk}(s)$	$d_{jk}(s)$
0.1	0	0	0	0	0	0
0.3	0	0	0	0	0	0
0.5	0	0	0	0	0	0
0.7	0	0	0	0	0	0
0.9	0	0	0	0	0	0
1.0	0	0	0	0	0	0
1.2	0	0	0	0	0	0
1.5	0	0	3751	0	0	2143
1.7	0	0	12134	0	0	12179
1.9	0	0	15963	0	0	17337
2.0	0	1	16854	0	0	18317

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. As curvas ajustadas aos histogramas são apresentadas como linhas contínuas enquanto as linhas pontilhadas representam as curvas gaussianas construídas com a média e o desvio padrão das distribuições.

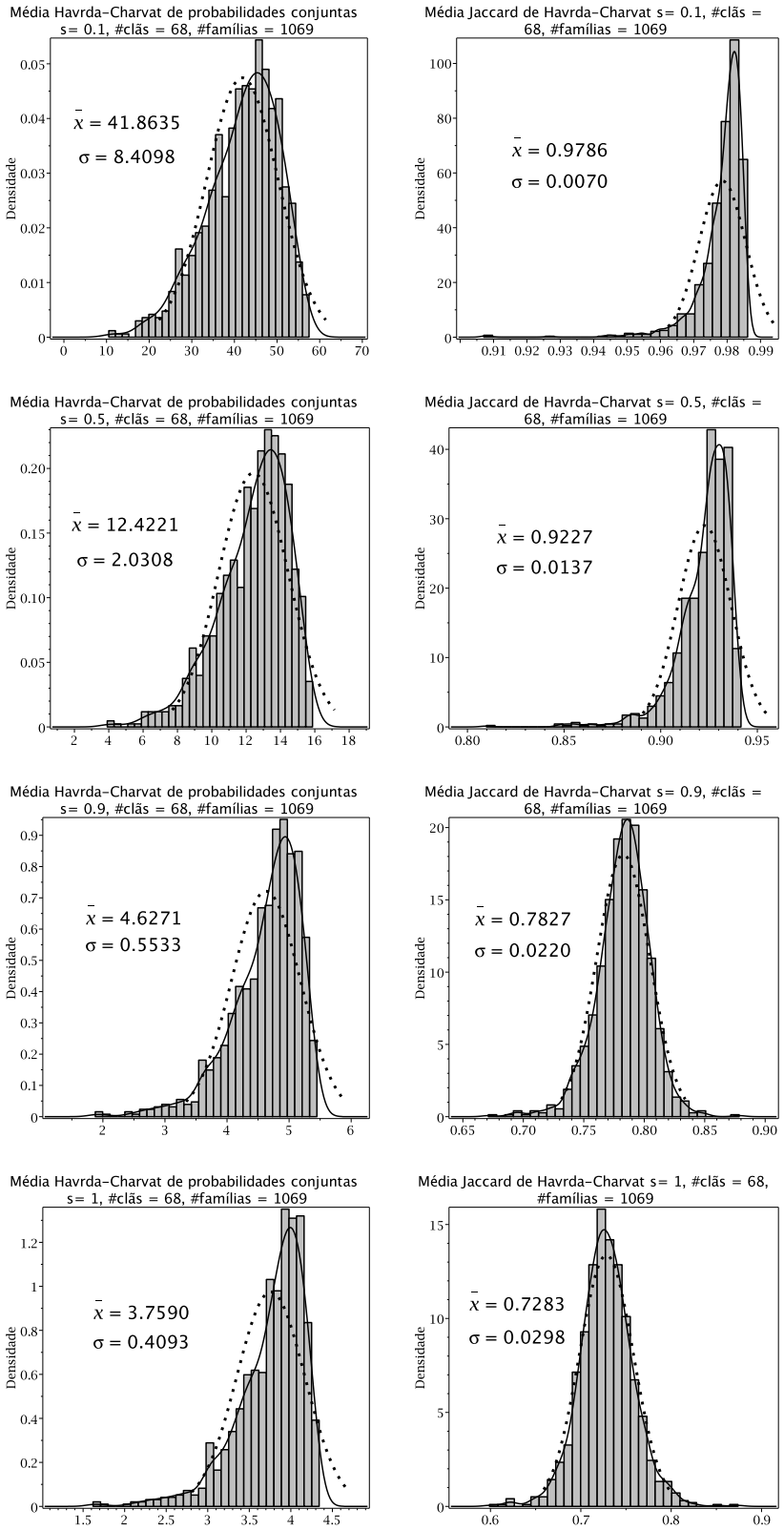


Figura 5.4: Histogramas de densidade dos valores das médias de entropia Havrda-Charvat de probabilidade conjunta (esquerda) e Jaccard (direita) das famílias, para blocos representativos (100×200). Linhas contínuas indicam a curva ajustada ao histograma enquanto as linhas pontilhadas representam a curva gaussiana construída com a média e o desvio padrão da distribuição.

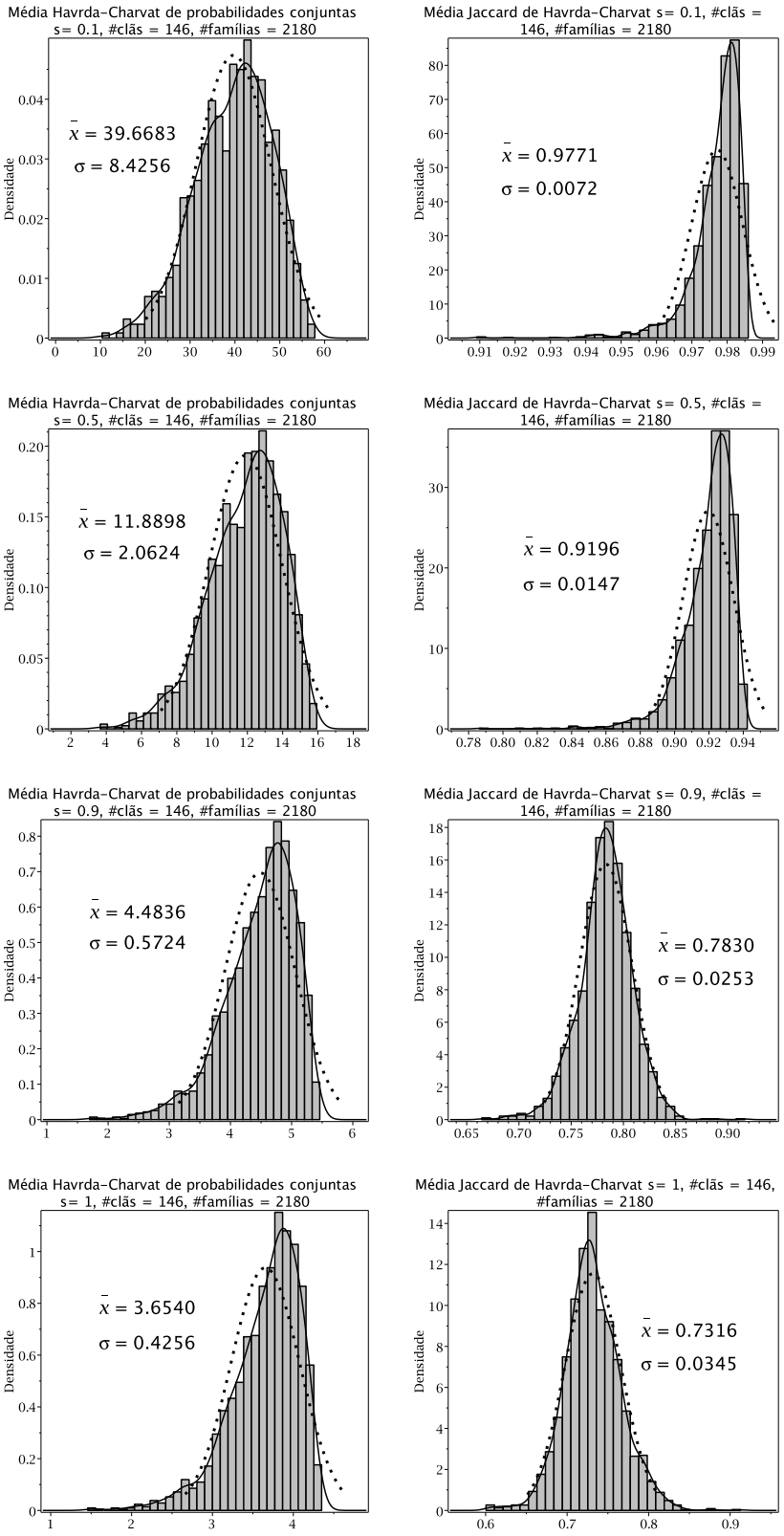


Figura 5.5: Histogramas de densidade dos valores das médias de entropia Havrda-Charvat de probabilidade conjunta (esquerda) e Jaccard (direita) das famílias, para blocos representativos (100×100). Linhas contínuas indicam a curva ajustada ao histograma enquanto as linhas pontilhadas representam a curva gaussiana construída com a média e o desvio padrão da distribuição.

É interessante notar o comportamento dos histogramas em relação aos valores do parâmetro s (Figuras 5.4 e 5.5). Enquanto os histogramas da distribuição das médias da entropia de Havrda-Charvat sofrem pequenas alterações em sua forma, os histogramas da distribuição das médias da entropia de Jaccard se aproximam cada vez mais de uma distribuição gaussiana, chegando muito próximo no valor de $s = 1$, correspondente ao limite Shannon. Sem a restrição de famílias pertencentes a clãs, apenas com as famílias que contêm blocos (100×100), a curva ajustada do histograma é aproximadamente gaussiana [36].

Capítulo 6

Sistemas Operacionais, Sistemas Computacionais, Linguagens de Programação e Estruturas de Dados

Neste capítulo apresentamos comparações entre as performances computacionais entre o sistema de computação algébrica Maple e a linguagem de programação Perl para os cálculos de probabilidades simples e conjuntas, das potências destas probabilidades e das entropias, para o caso com blocos representativos de famílias 100×200. Cálculos utilizando diferentes tipos de estruturas de dados e em diferentes sistemas operacionais também são apresentados.

Selecionamos uma família ao acaso, PF06850, a fim de obter uma ideia do tempo real e de CPU necessários para calcular as probabilidades e suas potências para o conjunto de 1069 famílias. A Tabela 6.1 abaixo apresenta os tempos obtidos para os cálculos de probabilidades com o Maple versão 18. Os tempos de cálculo de probabilidades conjuntas são muito maiores do que os de probabilidade simples. Estes resultados não devem variar de forma significativa entre as famílias, de forma que o tempo total necessário para o cálculo de todas as probabilidades pode ser estimado multiplicando estes valores por 1069, o número total de famílias para o caso com blocos representativos 100×200.

Tabela 6.1: Tempo de CPU e tempo real associados à família de domínios de proteínas PF06850 para o cálculo de probabilidades de ocorrência simples e conjunta.

Maple, versão 18.0	t_{CPU} (s)	t_R (s)
Probabilidades Simples	0.527	0.530
Probabilidades Conjuntas	5073.049	4650.697

Para o cálculo de probabilidade simples de uma família, temos um total de 20 valores (número de aminoácidos) para cada coluna do bloco, neste caso 200. Desta forma, cada família tem um total de $200 \cdot 20 = 4 \cdot 10^3$ valores de probabilidades simples. Já para o cálculo de probabilidades conjuntas, temos um total de 400 possíveis pares de aminoácidos, enquanto que os pares de colunas são determinados por uma combinação de 200 colunas tomadas duas a duas, resultando em 19900 pares. Então, as probabilidades conjuntas são um total de $\frac{200\cdot(200-1)}{2} \cdot 20 \cdot 20 = 7.96 \cdot 10^6$ valores para apenas uma família. O cálculo das probabilidades conjuntas é o mais extenuante dentre todos os outros cálculos realizados. Uma vez calculadas, estas probabilidades são agrupadas em conjuntos de 400 valores cada, referentes aos pares de colunas correspondentes. Da mesma forma, as probabilidades simples são agrupadas em conjuntos de 20 valores para cada coluna.

Tabela 6.2: Tempo de CPU e tempo real para o cálculo de 19 valores de potência s das probabilidades de ocorrência simples associadas a família de domínios de proteína PF06850.

Maple, versão 18.0 $\left(p_j(a)\right)^s$		
s	t_{CPU} (s)	t_R (s)
0.1	0.263	0.358
0.2	0.137	0.145
0.3	0.268	0.277
0.4	0.139	0.153
0.5	0.240	0.219
0.6	0.144	0.157
0.7	0.276	0.254
0.8	0.144	0.157
0.9	0.264	0.235
1.0	0.088	0.095
2.0	0.153	0.095
3.0	0.128	0.131
4.0	0.148	0.141
5.0	0.096	0.144
6.0	0.148	0.167
7.0	0.148	0.155
8.0	0.181	0.094
9.0	0.104	0.092
10.0	0.104	0.100
Total	3.173	3.164

Tabela 6.3: Tempo de CPU e tempo real para o cálculo de 19 valores de potência s das probabilidades de ocorrência conjuntas associadas a família de domínios de proteína PF06850.

Maple, versão 18.0 $\big(p_{jk}(a,b)\big)^s$		
s	t_{CPU} (s)	t_R (s)
0.1	390.432	206.646
0.2	382.887	202.282
0.3	401.269	210.791
0.4	416.168	216.993
0.5	427.572	221.541
0.6	430.604	223.227
0.7	421.904	218.484
0.8	434.888	224.267
0.9	431.948	223.023
1.0	442.933	224.731
2.0	176.212	147.455
3.0	234.100	174.853
4.0	289.184	181.552
5.0	327.740	178.117
6.0	334.800	194.691
7.0	349.064	195.258
8.0	361.304	195.437
9.0	386.217	197.150
10.0	397.276	197.868
Total	7036.502	3834.366

Após os cálculos de todos os valores de probabilidades simples e conjuntas, procedemos com os cálculos de probabilidades $(p_j(a))^s$ e $(p_{jk}(a,b))^s$, utilizando 19 valores do parâmetro s. Estes valores são posteriormente utilizados para os cálculos das entropias de Havrda-Charvat de acordo com as equações (5.3) e (5.4). Nas Tabelas 6.2 e 6.3, apresentamos os tempos no Maple para os cálculos de potências de probabilidades simples e conjunta, respectivamente. A última linha destas tabelas contém a soma dos tempos dos 19 cálculos de potência. Como as probabilidades conjuntas contêm arquivos com mais valores e em quantidades muito maiores do que as probabilidades simples, era de se esperar que os tempos para os cálculos de potência também fossem maiores. Porém, a diferença já não é tão grande quanto para o cálculo das probabilidades. Estes valores são então armazenados em $19 \cdot 200$

$= 3.8 \cdot 10^3$ arquivos, no caso das probabilidades simples, e em $19 \cdot 19900 = 3.781 \cdot 10^5$ arquivos, no caso das probabilidades conjuntas.

As Tabelas 6.4 e 6.5 apresentam os tempos no Maple para os cálculos de entropias de Havrda-Charvat, $H_j(s)$ e $H_{jk}(s)$, respectivamente. Estes tempos correspondem aos cálculos utilizando as potências s previamente calculadas. Como esperado, os tempos de cálculos das entropias advindas de probabilidades conjuntas são maiores do que as de probabilidades simples. Desta vez, são gerados arquivos para cada um dos valores do parâmetro s, de forma que os arquivos de entropia $H_j(s)$ contêm 200 valores e os de entropia $H_{jk}(s)$ contêm 19900 valores. A última linha das tabelas apresentam a soma dos tempos para o cálculo dos 19 valores de entropia.

Tabela 6.4: Tempo de CPU e tempo real para o cálculo das medidas de entropia $H_j(s)$ para a família de domínios de proteína PF06850.

Maple, versão 18.0 $H_j(s)$		
s	t_{CPU} (s)	t_R (s)
0.1	0.148	0.261
0.2	0.084	0.153
0.3	0.120	0.189
0.4	0.124	0.198
0.5	0.160	0.299
0.6	0.092	0.139
0.7	0.159	0.199
0.8	0.137	0.175
0.9	0.120	0.166
1.0	0.192	0.339
2.0	0.144	0.099
3.0	0.147	0.105
4.0	0.084	0.101
5.0	0.136	0.070
6.0	0.096	0.119
7.0	0.115	0.078
8.0	0.120	0.109
9.0	0.133	0.080
10.0	0.132	0.133
Total	2.443	3.012

Tabela 6.5: Tempo de CPU e tempo real para o cálculo das medidas de entropia $H_{jk}(s)$ para a família de domínios de proteína PF06850.

Maple, versão 18.0 $H_{jk}(s)$		
s	t_{CPU} (s)	t_R (s)
0.1	156.332	133.242
0.2	160.797	136.706
0.3	169.024	140.960
0.4	176.824	147.853
0.5	184.120	150.163
0.6	190.304	154.058
0.7	196.633	157.750
0.8	205.940	164.101
0.9	215.559	169.549
1.0	253.648	204.634
2.0	141.148	184.030
3.0	158.536	167.173
4.0	173.136	181.282
5.0	197.680	238.723
6.0	215.000	111.476
7.0	145.257	115.221
8.0	156.848	122.957
9.0	157.300	126.233
10.0	166.399	135.080
Total	3420.485	2941.791

O tempo total de CPU e real para o cálculo completo das entropias Havrda-Charvat $H_j(s)$ para os 19 valores do parâmetro s são obtidos ao somarmos os valores nas Tabelas 6.1, 6.2 e 6.4, e para as entropias $H_{jk}(s)$ somamos os valores das Tabelas 6.1, 6.3 e 6.5. Estes resultados são apresentados na tabela 6.6 a seguir. Como dito anteriormente, consideramos como certo que os tempos para calcular o conteúdo das medidas de entropia de cada família não diferirá muito, de forma que os resultados das terceira e quinta linhas são obtidos ao multiplicarmos os valores da segunda e quarta linha por 1069, o número total de famílias do espaço amostral para o caso dos blocos representativos 100×200.

Os resultados obtidos na Tabela 6.6 sugerem a inadequação do sistema de computacão algébrica Maple para analisar o conteúdo da medida de entropia de uma amostra do banco de dados de proteínas. Todos os cálculos foram feitos utilizando o

sistema operacional Linux (Ubuntu MATE 15.10). Como alternativa os códigos para os cálculos foram reescritos utilizando a linguagem de programação Perl. A escolha por esta linguagem se deve ao fato de ser amplamente utilizada por pesquisadores na área de Bioinformática, como por exemplo, na administração dos dados do Projeto Genoma Humano. Perl é uma linguagem de alto nível de fácil aprendizado, extremamente portátil e capaz de manipular um grande volume de dados de forma eficiente.

Tabela 6.6: Tempo total de CPU e tempo real total para o cálculo das medidas da família de domínios de proteína PF06850 e aproximações para o total geral de todo o espaço amostral.

Maple, versão 18.0	$H_j(s)$ – 19 valores de s	$H_{jk}(s)$ – 19 valores de s
Tempo total de CPU (família PF06850)	$0.527 + 3.173 + 2.443$ $= 6.143$ s	$5073.049 + 7036.502+$ $3420.485 = 15530.036$ s
Tempo geral total de CPU (1069 famílias)	6566.867 s $= 1.824$ h	$16,601,608.484$ s $= 192.148$ dias
Tempo real total (família PF06850)	$0.530 + 3.164 + 3.012$ $= 6.706$ s	$4650.697 + 3834.366+$ $2941.791 = 11426.854$ s
Tempo geral total real (1069 famílias)	7168.714 s $= 1.991$ h	$12,215,306.926$ s $= 141.381$ dias

Quatro configurações diferentes utilizando o sistema de computação algébrica Maple (M), a linguagem de programação Perl (P), os sistemas operacionais Linux (L) e Mac OSX (O), e as estruturas de dados Array I (A_I), Array II (A_{II}) e Hash (H), para realizar a tarefa de avaliação de bancos de dados de proteínas com medidas de entropias são listadas a seguir:

1. MLA_I : Maple, Linux, Array I
2. POH : Perl, OSX, Hash
3. PLA_{II} : Perl, Linux, Array II
4. POA_{II} : Perl, OSX, Array II

A estrutura Array I corresponde às matrizes utilizadas para o armazenamento e tratamento dos valores no Maple, e é muito eficiente nos cálculos que requerem conhecimentos de métodos matemáticos. Já a estrutura Array II é uma forma de adaptar a forma como os valores são tratados no Maple. A linguagem Perl contém a estrutura de dados array, que pode ser considerada como um vetor unidimensional, de forma que para termos uma estrutura similar a uma matriz, fazemos um “array

de array", onde cada elemento de um vetor linha seria um vetor coluna, por exemplo. A estrutura hash, muito utilizada em Perl, tem como vantagem ser muito rápida na busca de um elemento, porém quando se faz necessário uma análise da sequência ordenada de um grande volume de dados, o tempo necessário se torna impraticável [37, 38].

A Tabela 6.7 mostra a comparação dos tempos de CPU e dos tempos reais dos 19 valores de parâmetro s das potências de probabilidades conjuntas $(p_{jk}(a,b))^s$ para a família de domínios de proteínas PF06850 nas quatro configurações apresentadas acima. Deve ser enfatizado que estamos comparando os tempos de cálculo da potência s com os valores de probabilidade previamente calculados e armazenados em outros arquivos.

Ao checarmos os valores dos tempos de cálculo na Tabela 6.7, notamos que estes são genericamente ordenados como:

$$t_{MLA_I} > t_{POA_{II}} > t_{PLA_{II}} > t_{POH}$$

Os tempos obtidos com a estrutura hash foram os menores, mas isto se deve ao fato de ter sido feita a inserção sem ordenação, sendo então este o resultado já esperado [37]. Porém, o tempo necessário para se fazer uma inserção ordenada ou para se fazer uma busca para realizar os cálculos da entropia de Havrda-Charvat colocam a estrutura em desvantagem e, portanto, para este fim sua utilização deve ser evitada.

Uma observação deve ser feita em relação ao teste. Foram utilizadas duas máquinas diferentes para os cálculos: um notebook com processador Intel® CORE™ i7 de segunda geração, memória RAM de 4 GB DDR2 e 500 GB de HD, com Linux Ubuntu MATE 15.10; um notebook com processador Intel® CORE™ i7 de segunda geração, memória RAM de 8 GB DDR2 e 1 TB de HD, com OSX. Os computadores Mac têm como grande vantagem o fato do seu projeto de hardware ser desenvolvido conjuntamente com o sistema operacional, de forma que a máquina seja o mais eficiente possível e a execução de tarefas seja mais fluida. Desta forma, o resultado obtido de $t_{POA_{II}} > t_{PLA_{II}}$ não era o esperado. Um possível motivo pode ser algum tipo de tarefa rodando em segundo plano no Mac. Cabe então mais observação sobre o possível motivo: na máquina com Linux, os scripts de Perl com os cálculos foram executados direto do terminal sem qualquer outro programa aberto, enquanto que no Mac foi executado pela IDE Eclipse por outro experimentador, que foi instruído a não ter qualquer outro programa aberto, mas não é sabido se ele o fez como fora solicitado.

Apesar disso, acreditamos que com uma máquina com um sistema operacional Linux mais puro e com configurações de hardware melhores, podemos ter melhores resultados. Os códigos nos scripts também podem ser otimizados com alternativas

aos laços de iteração.

As conclusões comentadas acima são resumidas nas Tabelas 6.8 e 6.9 para o cálculo dos tempos de CPU e real dos 19 valores de potência s de probabilidades conjuntas $(p_{jk}(a,b))^s$ com as configurações POA_{II} e PLA_{II}, respectivamente, e nas Tabelas 6.10 e 6.11 para os valores correspondentes das medidas de entropia de Havrda-Charvat com as configurações POA_{II} e PLA_{II}, respectivamente. Os cálculos nestas tabelas foram realizados com seis famílias pertencentes a três clãs distintos. O tempo necessário para o cálculo das probabilidades conjuntas em si não foi levado em consideração.

Como uma última observação, devemos levar em consideração a restrição $0 < s \leq 1$ no trabalho com as medidas de entropia de Jaccard. Na Tabela 6.12 a seguir, apresentamos os resultados obtidos com a configuração PLA_{II} dos tempos de CPU e real para o cálculo da entropia de Havrda-Charvat com o conjunto de valores do parâmetro $s = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$.

Os tempos correspondentes para os cálculos de probabilidades conjuntas e de suas potências s foram somados para relatar os resultados do tempo total estimado para o cálculo das entropias de Havrda-Charvat de todas as 1069 famílias do espaço amostral no caso dos blocos representativos 100×200. A Tabela 6.13 apresenta os tempos para a configuração POA_{II} e a Tabela 6.14 apresenta os tempos para a configuração PLA_{II}. Acreditamos que os resultados obtidos são bem acessíveis.

Tabela 6.7: Uma comparação entre os tempos de cálculo (CPU e real) para 19 valores de potência de probabilidades conjuntas da família PF06850, utilizando as quatro configurações MLA_I, POA_{II}, PLA_{II}, POH.

s	$\mathbf{t_{CPU}}$ (s)				$\mathbf{t_R}$ (s)			
	MLA_I	POA_{II}	PLA_{II}	POH	MLA_I	POA_{II}	PLA_{II}	POH
0.1	390.432	33.478	**41.633**	24.849	206.646	88.527	83.139	26.862
0.2	382.887	31.711	26.483	25.093	202.282	80.624	53.384	26.904
0.3	401.269	31.726	25.286	24.751	210.791	80.519	50.689	27.305
0.4	416.168	31.448	26.138	26.345	216.993	79.032	51.409	27.687
0.5	427.572	32.860	**25.822**	26.726	221.541	93.048	51.334	27.528
0.6	430.604	33.444	27.013	25.021	223.227	102.317	52.889	25.408
0.7	421.904	31.053	25.814	25.011	218.484	79.255	51.466	26.414
0.8	434.888	31.526	26.725	25.183	224.267	80.469	53.388	25.668
0.9	431.948	31.482	26.895	25.409	223.023	80.002	53.579	25.536
1.0	442.933	32.056	25.990	25.096	224.731	80.917	51.687	25.640
2.0	176.212	32.638	27.089	26.012	147.454	80.751	54.384	26.960
3.0	234.100	31.892	25.766	24.498	174.853	85.853	51.843	24.717
4.0	284.184	31.662	26.515	25.251	181.552	91.837	52.636	25.718
5.0	327.740	32.295	27.516	24.925	178.117	87.486	54.908	25.814
6.0	334.800	32.674	28.126	25.440	194.691	86.569	54.611	25.847
7.0	349.064	31.674	23.908	26.389	195.258	86.215	49.262	27.745
8.0	361.304	33.105	26.020	25.106	195.437	116.601	51.889	26.735
9.0	386.217	31.881	26.208	24.783	197.150	81.372	53.114	25.155
10.0	397.276	32.269	26.125	24.979	197.868	87.963	52.541	26.504
Total	7036.502	611.374	515.072	480.867	3834.365	1649.357	1028.655	500.147
Total (1069) famílias	7,522,020.640 =87.067 dias	653,588.806 =7.565 dias	550,611.968 =6.373 dias	514,046.813 =5.950 dias	4,098,936.190 =47.441 dias	1,763,162.630 =20.407 dias	1,099,632.200 =12.727 dias	534,157.143 =6.188 dias

Tabela 6.8: Tempo de CPU e tempo real para o cálculo de 19 valores de potência s das probabilidades de ocorrência conjuntas de seis famílias pertencentes a três clãs com a configuração POA_{II}.

s	CL0028				CL0023				CL0257			
	PF06850		PF00135		PF00005		PF13481		PF02388		PF09924	
	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)
0.1	33.478	88.527	46.747	130.014	41.475	169.679	44.187	171.636	44.716	242.376	46.177	288.259
0.2	31.711	80.624	42.157	90.687	36.893	78.893	38.567	88.754	37.088	88.091	40.709	148.371
0.3	31.726	80.519	39.957	82.240	36.929	97.270	38.800	81.613	37.350	88.664	38.986	119.781
0.4	31.448	79.032	41.737	87.934	36.252	80.428	38.500	78.757	35.592	80.337	36.773	87.217
0.5	32.860	93.048	41.130	89.997	38.203	93.595	37.618	80.179	35.117	78.476	35.534	82.528
0.6	33.944	102.317	41.417	120.420	37.143	81.289	38.387	81.667	45.900	501.222	35.830	83.439
0.7	31.053	79.255	40.422	78.531	36.386	84.421	43.216	562.659	43.452	259.861	35.261	72.605
0.8	31.526	80.469	41.556	118.519	36.955	79.543	41.862	148.028	35.047	78.469	35.718	85.142
0.9	31.482	80.002	41.386	81.747	36.811	81.025	35.518	78.547	35.540	74.570	40.534	204.673
1.0	32.056	80.917	40.724	80.958	37.234	80.587	36.610	87.848	36.353	78.767	39.095	125.292
2.0	32.638	80.751	40.701	79.667	38.452	79.336	36.916	111.199	36.658	81.415	40.415	182.552
3.0	31.892	85.853	40.822	79.293	38.223	80.688	37.356	84.312	36.658	81.415	39.774	157.090
4.0	31.662	91.837	41.208	79.825	37.936	98.905	37.000	86.906	36.012	77.741	38.794	140.857
5.0	32.295	87.486	41.059	82.191	38.293	83.289	36.245	80.873	35.308	77.225	39.927	147.368
6.0	32.674	86.569	41.215	86.311	37.601	82.375	36.395	97.307	35.748	76.573	39.327	154.829
7.0	31.674	86.215	41.290	88.714	38.341	80.991	36.724	95.148	36.194	75.341	39.826	153.204
8.0	33.105	116.601	40.984	83.157	37.756	81.073	40.524	105.466	36.716	82.921	41.056	175.943
9.0	31.881	81.372	41.469	113.561	38.748	82.499	37.874	94.981	40.341	121.311	39.847	144.595
10.0	32.269	87.963	41.466	89.366	37.807	81.043	37.134	88.005	40.053	136.243	40.333	171.395
Total	611.374	1649.357	787.447	1743.132	717.438	1676.929	729.433	2303.885	719.843	2381.018	743.916	2725.140

Tabela 6.9: Tempo de CPU e tempo real para o cálculo de 19 valores de potência s das probabilidades de ocorrência conjuntas de seis famílias pertencentes a três clãs com a configuração PLA_{II}.

s	CL0028				CL0023				CL0257			
	PF06850		PF00135		PF00005		PF13481		PF02388		PF09924	
	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)
0.1	41.633	83.139	43.135	83.901	44.003	84.655	26.613	52.365	27.452	53.350	43.132	83.249
0.2	26.483	53.384	26.373	50.442	26.610	52.904	26.599	51.822	43.199	83.497	25.762	50.203
0.3	25.286	50.689	26.644	50.992	29.965	53.250	24.946	48.846	25.498	50.303	25.904	50.737
0.4	26.138	51.409	25.255	48.576	26.696	52.482	27.769	53.837	25.753	51.013	25.684	50.371
0.5	25.822	51.837	26.041	50.492	26.648	52.171	25.026	48.927	25.727	49.887	25.513	48.793
0.6	27.013	52.889	25.778	50.324	24.912	49.340	26.062	51.312	26.066	51.417	25.435	49.686
0.7	25.814	51.466	25.496	49.954	25.284	50.268	23.698	48.134	25.723	50.760	25.822	51.122
0.8	26.725	53.388	26.254	50.865	25.456	50.870	25.816	51.196	26.883	52.511	29.837	57.367
0.9	26.895	53.579	23.740	47.296	26.237	51.725	28.441	54.743	25.237	50.532	27.951	54.289
1.0	25.990	51.687	27.225	54.384	26.014	51.969	27.148	52.701	23.672	48.547	26.314	51.932
2.0	27.089	54.384	26.237	50.335	27.756	54.761	26.424	52.503	24.811	49.859	25.846	51.125
3.0	25.766	51.843	27.342	52.508	25.135	50.954	27.753	53.650	25.072	49.142	25.727	50.990
4.0	26.515	52.636	25.716	50.531	25.106	50.449	24.871	50.279	25.674	51.536	26.897	52.227
5.0	27.516	54.908	26.002	50.390	24.666	49.463	23.973	47.554	25.675	51.032	27.810	53.820
6.0	28.126	54.611	27.441	53.032	26.014	52.209	25.676	50.426	26.235	51.375	26.180	51.346
7.0	23.908	49.262	26.812	51.556	24.696	49.783	25.519	50.741	25.863	50.995	25.593	50.611
8.0	26.020	51.889	25.431	49.135	26.757	52.518	26.369	49.668	26.401	52.682	25.084	50.596
9.0	26.208	53.114	25.856	50.090	25.803	51.253	24.628	47.880	25.379	51.014	27.049	52.891
10.0	26.125	52.541	24.963	47.402	24.369	48.065	42.270	83.281	24.878	49.926	24.563	48.616
Total	515.072	1028.655	511.741	992.205	509.127	1009.089	509.601	999.865	505.198	999.378	516.103	1009.971

Tabela 6.10: Tempo de CPU e tempo real para o cálculo das medidas de entropia de Havrda-Charvat para probabilidades conjuntas de seis famílias pertencentes a três clãs com a configuração POA_{II}.

s	CL0028				CL0023				CL0257			
	PF06850		PF00135		PF00005		PF13481		PF02388		PF09924	
	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)
0.1	19.451	26.196	22.620	56.287	25.411	121.508	18.968	32.706	23.380	80.826	22.502	77.826
0.2	19.245	24.901	22.851	69.992	23.362	65.508	18.810	26.835	23.366	87.098	23.330	60.684
0.3	19.582	26.001	23.963	60.960	22.598	72.363	18.602	26.028	20.012	43.734	21.929	48.947
0.4	20.194	31.418	22.211	54.562	23.448	63.369	19.176	30.482	21.218	67.817	22.182	45.480
0.5	20.162	31.653	23.163	67.391	22.153	55.173	19.281	34.495	21.037	55.237	23.200	62.384
0.6	20.941	34.979	23.961	62.158	22.342	57.081	20.794	44.558	21.087	54.900	22.477	62.421
0.7	20.805	34.057	23.679	82.669	19.587	35.750	19.982	41.314	20.375	32.699	22.398	66.438
0.8	20.900	34.146	22.787	60.924	19.081	34.116	18.890	28.054	20.167	32.685	23.056	61.376
0.9	20.909	34.568	22.808	54.890	19.030	33.258	18.894	29.712	19.422	26.934	23.543	65.099
1.0	21.353	35.024	22.860	51.308	19.789	32.218	19.869	37.922	21.076	35.774	22.528	52.209
2.0	20.505	32.920	21.085	46.921	19.672	33.778	19.083	62.081	22.530	82.336	21.783	51.656
3.0	23.923	58.898	22.020	47.298	18.788	31.926	19.097	35.399	24.210	83.371	21.636	69.905
4.0	24.622	68.692	21.954	50.909	18.556	26.512	19.208	32.014	24.300	112.613	21.458	49.931
5.0	24.221	60.107	22.131	58.975	19.424	33.185	18.231	25.898	24.199	95.807	22.011	55.109
6.0	24.475	64.843	22.911	72.004	20.582	38.006	18.330	25.959	22.741	61.516	21.857	67.976
7.0	24.593	70.336	22.779	51.309	20.564	38.390	19.583	30.719	23.222	88.997	23.018	85.054
8.0	25.613	83.423	20.058	34.861	20.460	35.589	19.977	37.130	23.825	110.933	24.474	90.322
9.0	24.139	73.873	21.418	44.709	19.274	32.129	18.871	29.089	24.207	106.471	23.349	74.724
10.0	25.785	101.370	22.183	46.878	23.625	87.442	22.652	64.895	23.779	86.225	21.802	57.444
Total	421.418	927.405	427.442	1075.005	397.746	927.301	368.298	675.290	424.153	1345.973	428.533	1204.985

Tabela 6.11: Tempo de CPU e tempo real para o cálculo das medidas de entropia de Havrda-Charvat para probabilidades conjuntas de seis famílias pertencentes a três clãs com a configuração PLA_{II}.

s	CL0028				CL0023				CL0257			
	PF06850		PF00135		PF00005		PF13481		PF02388		PF09924	
	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)
0.1	21.219	26.895	18.698	23.442	19.383	24.675	21.351	26.966	19.059	24.560	21.451	26.829
0.2	22.604	28.014	21.109	26.536	20.336	25.665	19.555	24.562	19.370	24.681	22.504	27.855
0.3	19.056	24.130	20.692	25.663	21.934	27.466	21.084	26.236	19.317	24.934	19.681	24.732
0.4	21.552	27.442	19.981	25.021	18.794	24.142	21.065	26.485	20.878	26.291	21.947	26.979
0.5	31.952	88.406	21.161	26.374	19.123	24.309	22.534	27.851	19.729	25.011	21.888	27.168
0.6	20.195	25.362	21.006	26.371	21.426	30.303	20.640	26.050	19.424	24.279	20.750	26.175
0.7	19.168	23.924	20.989	26.482	21.440	27.129	19.370	24.287	19.274	24.779	21.545	26.830
0.8	21.982	27.362	19.760	24.681	20.662	25.948	21.324	26.648	21.685	26.964	21.644	27.160
0.9	21.356	27.235	21.251	26.550	19.518	24.880	21.276	26.736	21.283	26.981	19.217	23.910
1.0	20.206	25.608	20.914	26.173	20.726	25.877	21.430	26.625	20.102	25.198	19.810	25.348
2.0	19.435	24.882	20.660	25.675	20.798	26.326	20.301	25.481	19.588	24.837	20.446	25.755
3.0	20.549	25.646	19.988	25.479	20.629	25.516	19.914	25.230	20.438	25.648	19.393	24.815
4.0	19.528	24.693	20.649	25.610	19.428	24.637	20.505	26.001	20.868	26.009	20.060	25.643
5.0	19.698	24.824	21.076	26.468	19.865	24.753	19.120	24.267	19.466	24.423	21.760	27.244
6.0	20.809	26.319	20.352	25.914	19.507	24.362	20.110	25.587	20.708	25.961	21.240	26.665
7.0	20.287	25.951	21.073	26.459	19.482	24.861	18.272	23.535	19.015	24.460	20.646	25.964
8.0	21.427	26.885	19.494	24.421	20.107	25.228	20.645	26.095	21.039	26.222	20.014	25.155
9.0	21.623	27.335	21.554	26.517	19.239	24.529	20.629	25.929	21.035	26.450	21.086	26.526
10.0	20.815	26.379	21.127	26.630	19.927	25.340	19.781	25.063	21.438	27.019	17.286	22.390
Total	403.461	557.294	391.524	490.466	382.324	485.946	388.906	489.634	383.716	484.707	392.368	493.143

Tabela 6.12: Tempo de CPU e tempo real para o cálculo das medidas de entropia de Havrda-Charvat para probabilidades conjuntas de seis famílias pertencentes a três clãs com parâmetros $0 < s \leq 1$ e com a configuração PLA_{II}.

s	CL0028				CL0023				CL0257			
	PF06850		PF00135		PF00005		PF13481		PF02388		PF09924	
	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)	t_{CPU} (s)	t_R (s)
0.1	21.219	26.895	18.698	23.442	19.383	24.675	21.351	26.966	19.059	24.560	21.451	26.829
0.2	22.604	28.014	21.109	26.536	20.336	25.665	19.555	24.562	19.370	24.681	22.504	27.855
0.3	19.056	24.130	20.692	25.663	21.934	27.466	21.084	26.236	19.317	24.934	19.681	24.732
0.4	21.552	27.442	19.981	25.021	18.794	24.142	21.065	26.485	20.878	26.291	21.947	26.979
0.5	31.952	88.406	21.161	26.374	19.123	24.309	22.534	27.851	19.729	25.011	21.888	27.168
0.6	20.195	25.362	21.006	26.371	21.426	30.303	20.640	26.050	19.424	24.279	20.750	26.175
0.7	19.168	23.924	20.989	26.482	21.440	27.129	19.370	24.287	19.274	24.779	21.545	26.830
0.8	21.982	27.362	19.760	24.681	20.662	25.948	21.324	26.648	21.685	26.964	21.644	27.160
0.9	21.356	27.235	21.251	26.550	19.518	24.880	21.276	26.736	21.283	26.981	19.217	23.910
1.0	20.206	25.608	20.914	26.173	20.726	25.877	21.430	26.625	20.102	25.198	19.810	25.348
Total	219.290	324.380	205.561	257.293	203.342	260.394	209.629	262.446	200.121	253.678	210.437	262.986
Total Geral	1284.343	1895.873	1199.098	1727.300	1131.878	1693.572	1066.375	1511.927	1124.465	1681.087	1211.808	1772.106

Tabela 6.13: Tempo total de CPU e tempo real total para o cálculo das medidas de entropia de Havrda-Charvat da família de domínios de proteína PF06850 e aproximações para o total geral de todo o espaço amostral com a configuração POA_{II}.

POA_{II}	$H_j(s)$ – 19 valores de s	$H_{jk}(s)$ – 19 valores de s
Tempo total de CPU (família PF06850)	$0.358 + 2.562 + 0.292$ $= 3.212$ s	$550.129 + 611.374$ $+421.418 = 1,582.921$ s
Tempo geral total de CPU (1069 famílias)	$3,433.628$ s $= 0.954$ h	$1,692,142.549$ s $= 19.585$ dias
Tempo real total (família PF06850)	$1.062 + 9.300 + 0.332$ $= 10.694$ s	$593.848 + 1,649.357+$ $927.405 = 3.170.61$ s
Tempo geral total real (1069 famílias)	$11,431.886$ s $= 3.175$ h	$3,389,382.090$ s $= 39.229$ dias

Tabela 6.14: Tempo total de CPU e tempo real total para o cálculo das medidas de entropia de Havrda-Charvat da família de domínios de proteína PF06850 e aproximações para o total geral de todo o espaço amostral com a configuração PLA_{II}.

PLA_{II}	$H_j(s)$ – 19 valores de s	$H_{jk}(s)$ – 19 valores de s
Tempo total de CPU (família PF06850)	$0.291 + 1.801 + 0.640$ $= 2.732$ s	$787.254 + 515.072$ $+403.461 = 1,705.787$ s
Tempo geral total de CPU (1069 famílias)	$2,920.508$ s $= 0.811$ h	$1,823,486.303$ s $= 21.105$ dias
Tempo real total (família PF06850)	$0.642 + 7.261 + 1.291$ $= 9.194$ s	$1,068.026 + 1,028.655$ $+557.294 = 2,603.975$ s
Tempo geral total real (1069 famílias)	$9,828.386$ s $= 2.730$ h	$2,783,649.275$ s $= 32.218$ dias

Capítulo 7

Análise Estatística – ANOVA

Os cálculos de entropia das colunas individualmente ou dos pares de colunas de uma família resultam em uma quantificação da incerteza em relação à distribuição de aminoácidos em uma coluna ou em um par de colunas. Em posse desses valores, podemos dar início aos testes estatísticos para verificar a classificação das famílias de domínios de proteínas em clãs. Para tanto foi adotada a técnica *one-way ANOVA* (*Analysis of Variance*) que compara um valor experimental, determinado através da comparação entre as variabilidades das distribuições de entropias dos clãs (*intra-clãs*) com as variabilidades das distribuições de entropias entre clãs (*inter-clãs*) [39], com um valor teórico determinado a partir da distribuição **F** de *Fisher-Snedecor* [40, 41].

A distribuição F corresponde a um modelo ideal, um caso especial composto a partir de dois conjuntos de distribuições, um com μ elementos e o outro com ν. Cada um dos conjuntos é utilizado na construção de uma distribuição qui-quadrado (χ^2):

$$X = \chi^2_\mu = X_1^2 + X_2^2 + \ldots + X_\mu^2$$
$$Y = \chi^2_\nu = Y_1^2 + Y_2^2 + \ldots + Y_\nu^2$$

sendo que as seguintes condições são previamente necessárias:

1. As μ variáveis aleatórias, e da mesma forma as ν variáveis aleatórias, são independentes entre si.

2. Todas as μ e ν distribuições são normais.

3. Os valores esperados das variáveis aleatórias são todos identicamente iguais a 0. Além disso, todas as variáveis aleatórias possuem a mesma variância σ^2.

Sendo todos os elementos dos dois conjuntos de variáveis aleatórias, $X_1, X_2, \ldots, X_\mu$ e $Y_1, Y_2, \ldots, Y_\nu$, também mutuamente independentes entre si, po-

demos definir a variável aleatória T, que tem uma distribuição F, como a razão:

$$T = \frac{X/\mu}{Y/\nu} \tag{7.1}$$

Na seção a seguir apresentamos como é feita a formulação da distribuição F.

7.1 Distribuição F de Fisher-Snedecor

Dado um conjunto de variáveis aleatórias X_j normalmente distribuídas, onde $j = 1, 2, \ldots, \mu$, suas funções densidade de probabilidade (**pdf** — *probability density function*) são definidas da seguinte forma:

$$N_{\bar{x}_j,\sigma_j}(x_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_j-\bar{x}_j)^2}{2\sigma_j^2}} \quad , x_j \in (-\infty, +\infty)\,, \tag{7.2}$$

onde $\bar{x}_j$ é o valor esperado da variável aleatória X_j e σ_j seu desvio padrão. Para o caso em que todas possuem a mesma média igual a 0, e a mesma variância σ^2, temos:

$$N_{0,\sigma}(x_j) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x_j^2}{2\sigma^2}}\,. \tag{7.3}$$

Seja um conjunto de variáveis aleatórias Z_j, funções das variáveis aleatórias X_j, tal que $Z_j = X_j^2$, temos então que suas funções de distribuição acumulada (**cdf** — *cumulative distribution function*) são descritas da seguinte forma:

$$\begin{aligned}\mathrm{cdf}(Z_j = X_j^2) &= P(Z_j \leq z_j) = P(X_j^2 \leq z_j) = P(-\sqrt{z_j} \leq X_j \leq \sqrt{z_j}) \\ &= \int_{-\sqrt{z_j}}^{\sqrt{z_j}} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{x_j^2}{2\sigma^2}}\,\mathrm{d}x_j\end{aligned}$$

Como a função normal de média 0 é uma função par, temos que:

$$\mathrm{cdf}(Z_j = X_j^2) = \frac{2}{\sqrt{2\pi\sigma^2}} \int_0^{\sqrt{z_j}} e^{-\frac{x_j^2}{2\sigma^2}}\,\mathrm{d}x_j \tag{7.4}$$

Sua pdf é então:

$$\begin{aligned}f_{Z_j}(z_j) &= \frac{\mathrm{d}}{\mathrm{d}z_j}\,\mathrm{cdf}(Z_j = X_j^2) \\ &= \frac{\mathrm{d}}{\mathrm{d}z_j}\,\frac{2}{\sqrt{2\pi\sigma^2}} \int_0^{\sqrt{z_j}} e^{-\frac{x_j^2}{2\sigma^2}}\,\mathrm{d}x_j \\ &= \frac{2}{\sqrt{2\pi\sigma^2}}\,\frac{\mathrm{d}}{\mathrm{d}\sqrt{z_j}}\,\frac{\mathrm{d}\sqrt{z_j}}{\mathrm{d}z_j} \int_0^{\sqrt{z_j}} e^{-\frac{x_j^2}{2\sigma^2}}\,\mathrm{d}x_j\end{aligned}$$

$$f_{Z_j}(z_j) = \frac{2}{\sqrt{2\pi\sigma^2}} \frac{1}{2\sqrt{z_j}} e^{-\frac{(\sqrt{z_j})^2}{2\sigma^2}}$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} z_j^{-1/2} e^{-\frac{z_j}{2\sigma^2}} \quad , z_j \in (0, +\infty) \tag{7.5}$$

Desta forma, a pdf de cada variável aleatória Z_j corresponde a uma distribuição qui-quadrado com **um** grau de liberdade. Da mesma forma, também é idêntica à *distribuição gama*:

$$f_{\theta,\beta}(w) = \frac{1}{\Gamma(\beta)} \theta^\beta w^{\beta-1} e^{-\theta w} \qquad \theta > 0,\ \beta > 0 \tag{7.6}$$

quando os parâmetros θ e β assumem os valores $1/2\sigma^2$ e $1/2$, respectivamente:

$$f_{\frac{1}{2\sigma^2},\frac{1}{2}}(z_j) = \frac{1}{\sqrt{2\pi\sigma^2}} z_j^{-1/2} e^{-\frac{z_j}{2\sigma^2}} \tag{7.7}$$

Onde $\Gamma(.)$ é a *função gama de Euler* [42]:

$$\Gamma(t) = \int_0^\infty w^{t-1} e^{-w} \mathrm{d}w \tag{7.8}$$

A função gama apresenta a seguinte propriedade:

$$\Gamma(t) = (t-1)\Gamma(t-1) \quad \forall t > 0 \tag{7.9}$$

E para todo número N inteiro não negativo temos:

$$\Gamma(N+1) = N! \tag{7.10}$$

Uma distribuição qui-quadrado com dois graus de liberdade, χ^2_2, é obtida através da soma de duas variáveis aleatórias independentes e identicamente distribuídas X^2_j (eq. (7.5), (7.7)). Assim, para uma variável aleatória $Z = Z_1 + Z_2 = X^2_1 + X^2_2$, temos:

$$\mathrm{cdf}(Z = Z_1 + Z_2) = P(Z_1 + Z_2 \leq z) = \int_0^\infty \int_0^{z-u} f_{Z_1}(u) f_{Z_2}(v) \mathrm{d}v\, \mathrm{d}u \tag{7.11}$$

E sua pdf é obtida através da equação

$$f_Z(z) = \frac{\mathrm{d}}{\mathrm{d}z} \int_0^\infty \int_0^{z-u} f_{Z_1}(u) f_{Z_2}(v) \mathrm{d}v\, \mathrm{d}u$$
$$= \int_0^z f_{Z_1}(u) f_{Z_2}(z-u) \mathrm{d}u \tag{7.12}$$

que é a convolução das pdfs das variáveis aleatórias Z_1 e Z_2: $f_{Z_1}(z) * f_{Z_2}(z)$. Resol-

vendo a equação (7.12), temos:

$$f_Z(z) = \int_0^z \frac{1}{\sqrt{2\pi\sigma^2}}\, u^{-1/2} e^{-\frac{u}{2\sigma^2}}\, \frac{1}{\sqrt{2\pi\sigma^2}} (z-u)^{-1/2} e^{-\frac{(z-u)}{2\sigma^2}}\, \mathrm{d}u$$

$$= \frac{1}{2\pi\sigma^2}\, e^{-\frac{z}{2\sigma^2}} \int_0^z (z-u)^{-1/2} u^{-1/2} \mathrm{d}u$$

Fazendo a mudança de variável $u = zt$, ficamos com:

$$f_Z(z) = \frac{1}{2\pi\sigma^2}\, e^{-\frac{z}{2\sigma^2}} \int_0^1 (z - zt)^{-1/2} (zt)^{-1/2}\, z\, \mathrm{d}t$$

$$= \frac{1}{2\pi\sigma^2}\, e^{-\frac{z}{2\sigma^2}} \int_0^1 (1-t)^{-1/2} t^{-1/2}\, \mathrm{d}t$$

$$= \frac{1}{2\pi\sigma^2}\, e^{-\frac{z}{2\sigma^2}}\, \mathrm{B}\left(\frac{1}{2}, \frac{1}{2}\right)$$

$$= \frac{1}{2\pi\sigma^2}\, e^{-\frac{z}{2\sigma^2}}\, (\sqrt{\pi})^2$$

$$f_Z(z) = f_{\frac{1}{2\sigma^2},1}(z) = \frac{1}{2\sigma^2}\, e^{-\frac{z}{2\sigma^2}} \tag{7.13}$$

A pdf da variável aleatória Z é uma distribuição gama com parâmetros $1/2\sigma^2$ e 1. B(.) é a *função Beta* [42]:

$$\mathrm{B}(u,v) = \int_0^1 t^{u-1}(1-t)^{v-1}\, \mathrm{d}t = \int_0^\infty \frac{t^{u-1}}{(1+t)^{u+v}}\, \mathrm{d}t \tag{7.14}$$

que pode ser escrita como uma relação entre funções gama da seguinte forma:

$$\mathrm{B}(u,v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)} \tag{7.15}$$

É fácil comprovar que uma distribuição qui-quadrado com três graus de liberdade é obtida através da convolução $f_{Z_1}(z) * f_{Z_2}(z) * f_{Z_3}(z)$, cujo resultado é uma distribuição gama com parâmetros $1/2\sigma^2$ e $3/2$:

$$f_{\frac{1}{2\sigma^2},\frac{3}{2}}(z) = \frac{2}{\sqrt{\pi}} \left(\frac{1}{2\sigma^2}\right)^{3/2} z^{1/2}\, e^{-\frac{z}{2\sigma^2}} \tag{7.16}$$

Uma vez que a distribuição gama tem a seguinte propriedade:

$$f_{\theta,\gamma_1}(z) * f_{\theta,\gamma_2}(z) = f_{\theta,\gamma_1+\gamma_2}(z) \tag{7.17}$$

podemos generalizar para o caso da distribuição qui-quadrado com μ graus de

liberdade:

$$f_{\frac{1}{2\sigma^2},\frac{\mu}{2}}(z) = \frac{1}{\Gamma\left(\frac{\mu}{2}\right)}\left(\frac{1}{2\sigma^2}\right)^{\frac{\mu}{2}} z^{\frac{\mu}{2}-1}\, e^{-\frac{z}{2\sigma^2}} \tag{7.18}$$

Para uma variável aleatória W, definida como a razão entre as variáveis aleatórias X e Y, respectivamente iguais a $X_1^2 + X_2^2 + \ldots + X_\mu^2$ e $Y_1^2 + Y_2^2 + \ldots + Y_\nu^2$, com $x \in [0, \infty)$ e $y \in [0, \infty)$ temos:

$$\text{cdf}\left(W = \frac{X}{Y}\right) = F_W(w) = P(W \leq w) = P\left(\frac{X}{Y} \leq w\right) = P(X \leq wY)$$

Como todas as variáveis aleatórias $X_1, X_2, \ldots, X_\mu$ e $Y_1, Y_2, \ldots, Y_\nu$ são independentes entre si, temos:

$$F_W(w) = \int_0^\infty \int_0^{wy} f_X(x) f_Y(y)\, \mathrm{d}x\, \mathrm{d}y \tag{7.19}$$

e

$$\begin{aligned} f_W(w) = \frac{\mathrm{d}}{\mathrm{d}w} F_W(w) &= \frac{\mathrm{d}}{\mathrm{d}(wy)} \frac{\mathrm{d}(wy)}{\mathrm{d}w} \int_0^\infty \int_0^{wy} f_X(x) f_Y(y)\, \mathrm{d}x\, \mathrm{d}y \\ &= \int_0^\infty f_X(wy) f_Y(y)\, y\, \mathrm{d}y \end{aligned} \tag{7.20}$$

Como X e Y são variáveis aleatórias com distribuições chi-quadrado com μ e ν graus de liberdade respectivamente, temos:

$$f_W(w) = \int_0^\infty \frac{1}{\Gamma\left(\frac{\mu}{2}\right)}\left(\frac{1}{2\sigma^2}\right)^{\frac{\mu}{2}} (wy)^{\frac{\mu}{2}-1}\, e^{-\frac{wy}{2\sigma^2}} \frac{1}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{1}{2\sigma^2}\right)^{\frac{\nu}{2}} y^{\frac{\nu}{2}-1}\, e^{-\frac{y}{2\sigma^2}}\, y\, \mathrm{d}y$$

$$= \frac{w^{\frac{\mu}{2}-1}}{(2\sigma^2)^{\frac{\mu+\nu}{2}}\Gamma(\frac{\mu}{2})\Gamma(\frac{\nu}{2})} \int_0^\infty y^{\frac{\mu+\nu}{2}-1} e^{-\frac{(w+1)y}{2\sigma^2}} \mathrm{d}y$$

Através de uma mudança de variável $\tau = \frac{(w+1)y}{2\sigma^2}$, podemos escrever:

$$f_W(w) = \frac{w^{\frac{\mu}{2}-1}}{(2\sigma^2)^{\frac{\mu+\nu}{2}}\Gamma(\frac{\mu}{2})\Gamma(\frac{\nu}{2})} \int_0^\infty \left(\frac{2\sigma^2\tau}{w+1}\right)^{\frac{\mu+\nu}{2}-1} e^{-\tau} \frac{2\sigma^2}{w+1}\, \mathrm{d}\tau$$

$$= \frac{w^{\frac{\mu}{2}-1}}{(2\sigma^2)^{\frac{\mu+\nu}{2}}\Gamma(\frac{\mu}{2})\Gamma(\frac{\nu}{2})} \left(\frac{2\sigma^2}{w+1}\right)^{\frac{\mu+\nu}{2}} \int_0^\infty \tau^{\frac{\mu+\nu}{2}-1} e^{-\tau}\, \mathrm{d}\tau$$

e utilizando a definição da função gama (eq. (7.8)) temos:

$$f_W(w) = \frac{\Gamma(\frac{\mu+\nu}{2})}{\Gamma(\frac{\mu}{2})\Gamma(\frac{\nu}{2})} \cdot \frac{w^{\frac{\mu}{2}-1}}{(w+1)^{\frac{\mu+\nu}{2}}} \tag{7.21}$$

A variável aleatória T (eq. (7.1)) pode ser definida em função da variável aleatória

W como:

$$T = \frac{\nu}{\mu} W$$

Assim, temos:

$$\text{cdf}\left(T = \frac{\nu}{\mu}W\right) = F_T(t) = P(T \leq t) = P\left(\frac{\nu}{\mu}W \leq t\right) = P\left(W \leq \frac{\mu}{\nu}t\right)$$

e

$$\begin{aligned} f_T(t) = \frac{\mathrm{d}}{\mathrm{d}t}F_T(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\int_0^{\frac{\mu}{\nu}t} f_W(w)\mathrm{d}w \\ &= \frac{\mathrm{d}}{\mathrm{d}\left(\frac{\mu}{\nu}t\right)}\frac{\mathrm{d}\left(\frac{\mu}{\nu}t\right)}{\mathrm{d}t}\int_0^{\frac{\mu}{\nu}t} f_W(w)\mathrm{d}w \\ &= \frac{\mu}{\nu} f_W\left(\frac{\mu}{\nu}t\right) \end{aligned}$$

A pdf da distribuição F pode finalmente ser escrita como:

$$f(\mu,\nu;t) = f_T(t) = \frac{\Gamma(\frac{\mu+\nu}{2})}{\Gamma(\frac{\mu}{2})\Gamma(\frac{\nu}{2})} \cdot \mu^{\frac{\mu}{2}} \cdot \nu^{\frac{\nu}{2}} \cdot \frac{t^{\frac{\mu}{2}-1}}{(\mu\, t + \nu)^{\frac{\mu+\nu}{2}}} \tag{7.22}$$

7.2 Análise de Variância (ANOVA) e o Teste F

A análise de variância (ANOVA), desenvolvida por Ronald A. Fisher na década de 1920 [43], é um tipo de teste de hipóteses, utilizado para fazer inferências estatísticas a partir de dados experimentais, ou seja, auxilia na conclusão de determinadas características de uma distribuição de probabilidades. A ANOVA é um teste estatístico que consiste na comparação entre um conjunto de amostras de dois ou mais grupos de dados com um modelo ideal.

Em um teste de hipóteses analizamos uma afirmação feita a respeito da relação entre os conjuntos de dados. A *hipótese nula*, H_0, considerada como verdadeira *a priori*, diz que não existe uma diferença significativa entre as amostras. Já a *hipótese alternativa*, H_a, afirma que uma ou mais amostras dos grupos, não necessariamente todas, são estatisticamente diferentes. Rejeitar a hipótese nula H_0 não significa aceitar indubitavelmente a hipótese alternativa H_a [44].

Como dito anteriormente, a comparação entre um valor intra-clã (como os valores de entropia das famílias variam em relação a média de entropia do clã) com um valor inter-clã (como as médias de entropias de cada clã variam em relação a média total) corresponde ao valor experimental, uma vez que é obtido de amostras do banco de dados. Já o valor teórico, denominado de coeficiente de correlação F de Fisher, é determinado a partir da distribuição F de Fisher-Snedecor, e depende de três parâmetros adimensionais: os parâmetros referentes aos graus de liberdade das distribuições χ^2, μ e ν, e o parâmetro α, denominado de nível de significância

do teste. α é usualmente arbitrado com os valores 0.1, 0.05 ou 0.01. Este valor é a probabilidade de rejeitarmos a hipótese nula sendo ela verdadeira [44], o que é conhecido como *erro do tipo I*. Um *erro do tipo II* ocorre quando o método adotado é incapaz de rejeitar uma hipótese nula quando ela é falsa.

O coeficiente de correlação $F_{\mu\,\nu\,\alpha}$ é obtido, de acordo com o parâmetro α escolhido, a partir da cdf da distribuição F de Fisher-Snedecor, $g(\mu, \nu; F_{\mu\,\nu\,\alpha})$, da seguinte forma:

$$g(\mu, \nu; F_{\mu\,\nu\,\alpha}) = \int_0^{F_{\mu\,\nu\,\alpha}} f(\mu, \nu; t)\,\mathrm{d}t \tag{7.23}$$

$$= \frac{\Gamma(\frac{\mu+\nu}{2})}{\Gamma(\frac{\mu}{2})\Gamma(\frac{\nu}{2})} \cdot \mu^{\frac{\mu}{2}} \cdot \nu^{\frac{\nu}{2}} \cdot \int_0^{F_{\mu\,\nu\,\alpha}} \frac{t^{\frac{\mu}{2}-1}}{(\mu\,t+\nu)^{\frac{\mu+\nu}{2}}}\,\mathrm{d}t \tag{7.24}$$

$$= \frac{\Gamma(\frac{\mu+\nu}{2})}{\Gamma(\frac{\mu}{2})\Gamma(\frac{\nu}{2})}\, 2\,\mu^{\frac{\mu-2}{2}} \left(\frac{F_{\mu\,\nu\,\alpha}}{\nu}\right)^{\frac{\mu}{2}} {}_2F_1\left(\frac{\mu+\nu}{2}, \frac{\mu}{2}; 1+\frac{\mu}{2}; -\frac{\mu}{\nu}F_{\mu\,\nu\,\alpha}\right) \tag{7.25}$$

Onde

$${}_2F_1\,(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-tz)^a}\mathrm{d}t$$

é a *função hipergeométrica de Gauss* [42].

O cálculo da integral da pdf até o valor do coeficiente $F_{\mu\,\nu\,\alpha}$ corresponde a uma certa porcentagem de sua área total, determinada a partir do parâmetro α da seguinte forma:

$$\int_0^{F_{\mu\,\nu\,\alpha}} f(\mu, \nu; t)\,\mathrm{d}t = 1 - \alpha \tag{7.26}$$

ou seja,

$$g(\mu, \nu; F_{\mu\,\nu\,\alpha}) = 1 - \alpha \tag{7.27}$$

uma vez que a cdf é uma função monótona não decrescente e

$$0 \leq g(\mu, \nu; F_{\mu\,\nu\,\alpha}) \leq 1$$

para todo $F_{\mu\,\nu\,\alpha}$. As Figuras 7.1 e 7.2 mostram graficamente a influência do parâmetro α na pdf e na cdf, respectivamente.

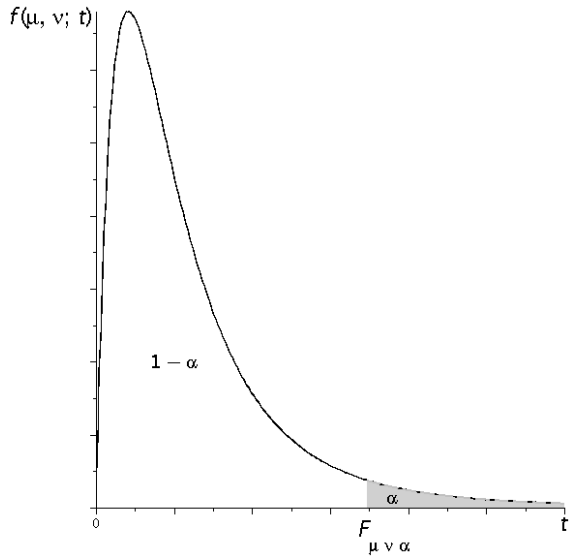


Figura 7.1: Gráfico da função densidade de probabilidade (pdf) da distribuição F de Fisher-Snedecor. A área em cinza corresponde a $\alpha\%$ da área limitada pela curva e o eixo horizontal, portanto, quanto maior for o valor de α, maior será a área em cinza.

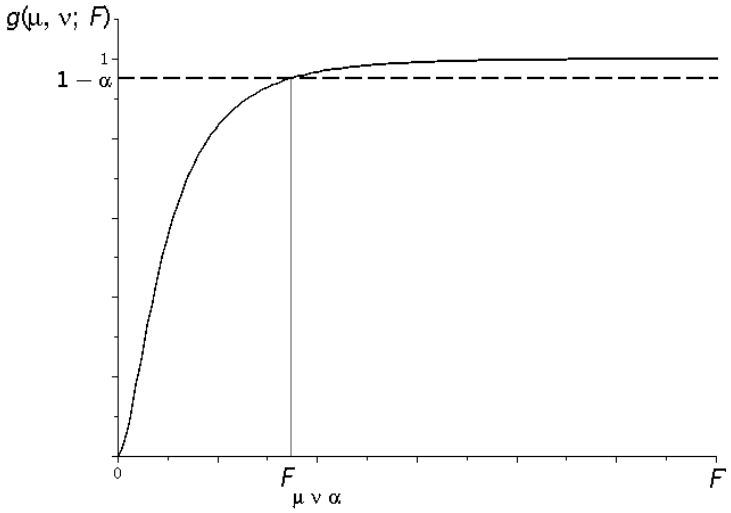


Figura 7.2: Gráfico da função de distribuição acumulada (cdf) da distribuição F de Fisher-Snedecor. A interseção da reta tracejada $(1-\alpha)$ com a curva $g(\mu, \nu; F)$ tem como abscissa o valor $F_{\mu\,\nu\,\alpha}$.

Para os valores dos parâmetros μ e ν dos graus de liberdade e para o nível de significância α desejado, resolvemos a equação (7.27) para $F_{\mu\,\nu\,\alpha}$. As Figuras 7.3 e 7.4 a seguir, apresentam, respectivamente, as pdfs e cdfs da distribuição F de cinco combinações diferentes dos parâmetros μ e ν. Os valores dos parâmetros correspondem a configurações do teste realizado na sessão seguinte.

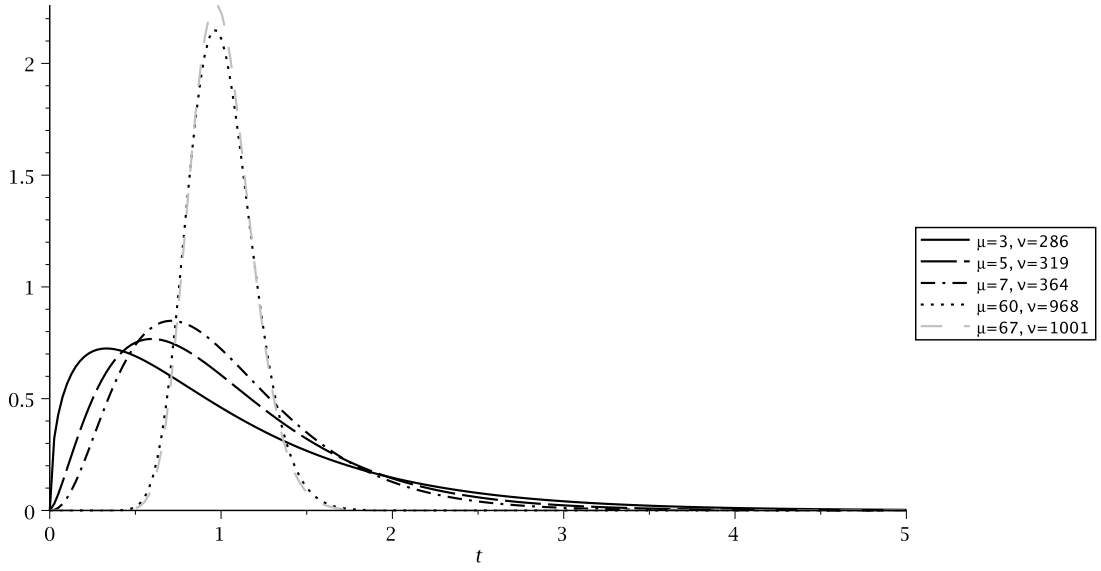


Figura 7.3: Exemplo de curvas de pdf com valores de parâmetro μ e ν utilizados nos testes.

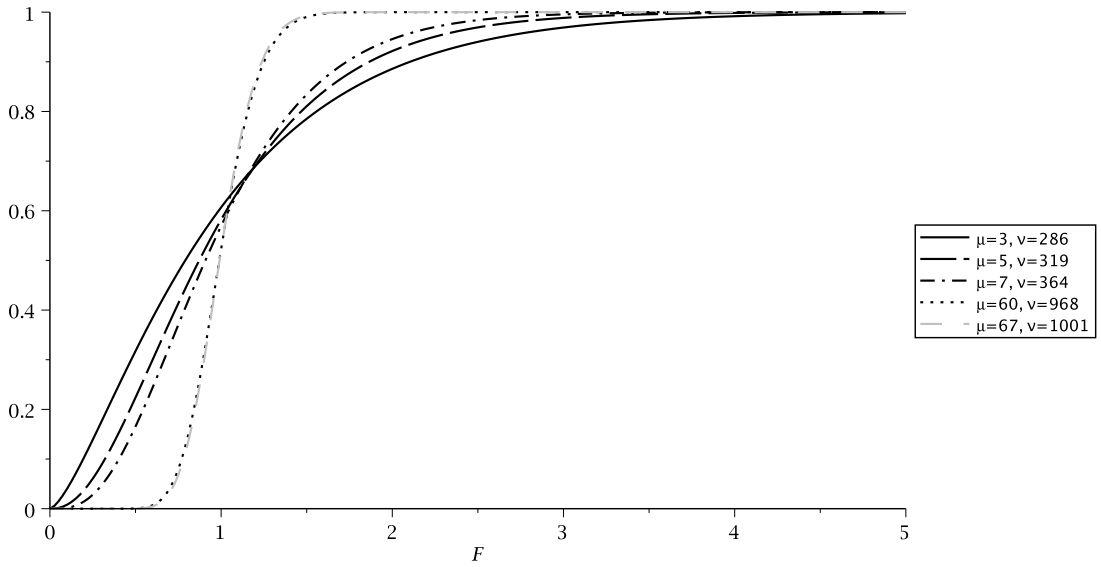


Figura 7.4: Exemplo de curvas de cdf com valores de parâmetro μ e ν utilizados nos testes.

Existem diversas tabelas com valores da distribuição F para uma grande variedade de combinações dos parâmetros μ e ν, e para alguns valores de α, porém, para certas combinações específicas, o cálculo para a determinação do valor $F_{\mu\,\nu\,\alpha}$ se faz necessário. Para o caso em que a distribuição χ^2 do denominador da variável aleatória T (eq.(7.1)) tem um grau de liberdade muito maior do que 1, podemos fazer a aproximação apresentada a seguir.

Seja a função

$$h(\mu,\nu;t) = \frac{t^{\frac{\mu}{2}-1}}{(\mu t+\nu)^{\frac{\mu+\nu}{2}}}\,, \tag{7.28}$$

que ao ser normalizada, resulta na equação (7.22).

Fazendo a substituição de variável $\theta = \frac{t}{F_{\mu\,\nu\,\alpha}}$ temos:

$$h(\mu,\nu;\theta) = \frac{(F_{\mu\,\nu\,\alpha}\theta)^{\frac{\mu}{2}-1}}{(\mu F_{\mu\,\nu\,\alpha}\theta+\nu)^{\frac{\mu+\nu}{2}}}$$

$$= \frac{(F_{\mu\,\nu\,\alpha})^{\frac{\mu}{2}-1}\theta^{\frac{\mu}{2}-1}}{\nu^{\frac{\mu+\nu}{2}}\left(\frac{\mu}{\nu}F_{\mu\,\nu\,\alpha}\theta+1\right)^{\frac{\mu+\nu}{2}}} \tag{7.29}$$

Para $\nu \gg 1$:

$$\lim_{\nu\gg1}\left(1+\frac{C}{\nu}\right)^{D\nu} = \left(\lim_{\nu\gg1}\left(1+\frac{C}{\nu}\right)^{\nu}\right)^{D} = \left(e^{C\nu}\right)^{D} = e^{CD\nu}$$

com

$$C = \mu\theta F_{\mu,\nu,\alpha}$$
$$D = \frac{1}{2}\left(1+\frac{\mu}{\nu}\right)$$

Assim,

$$h(\mu,\nu;\theta) \approx \frac{(F_{\mu\,\nu\,\alpha})^{\frac{\mu}{2}-1}\theta^{\frac{\mu}{2}-1}}{\nu^{\frac{\mu+\nu}{2}}e^{\frac{\mu F_{\mu\,\nu\,\alpha}}{2}\left(1+\frac{\mu}{\nu}\right)\theta}} \tag{7.30}$$

Para obtermos a pdf aproximada, normalizamos a função $h(\mu,\nu;\theta)$:

$$\bar{f}(\mu,\nu;\theta) \approx \frac{\left[\frac{\mu}{2}\left(1+\frac{\mu}{\nu}\right)\right]^{\frac{\mu}{2}}}{\Gamma(\frac{\mu}{2})}(F_{\mu\,\nu\,\alpha})^{\frac{\mu}{2}-1}\theta^{\frac{\mu}{2}-1}e^{-\frac{\mu}{2}F_{\mu\,\nu\,\alpha}\left(1+\frac{\mu}{\nu}\right)\theta} \tag{7.31}$$

A cdf aproximada é dada então por:

$$\bar{g}(\mu,\nu;F_{\mu\,\nu\,\alpha}) \approx \frac{\left[\frac{\mu}{2}F_{\mu\,\nu\,\alpha}\left(1+\frac{\mu}{\nu}\right)\right]^{\frac{\mu}{2}}}{\Gamma(\frac{\mu}{2})}\int_0^1 \theta^{\frac{\mu}{2}-1}e^{-\frac{\mu}{2}F_{\mu\,\nu\,\alpha}\left(1+\frac{\mu}{\nu}\right)\theta}\mathrm{d}\theta \tag{7.32}$$

A integral de (7.32) pode ser expressa da seguinte forma:

$$\int_0^1 v^{A-1}e^{-Bv}\mathrm{d}v = \frac{B^{-\frac{1}{2}(A+1)}e^{-\frac{1}{2}B}}{A}\,\mathrm{WhittakerM}\left(\frac{1}{2}(A-1),\frac{1}{2}A,B\right), \tag{7.33}$$

onde WhittakerM é a função *M de Whittaker*, uma solução especial da *equação de Whittaker* [42]. Substituindo em (7.32), temos a aproximação utilizada:

$$\bar{g}(\mu,\nu;F_{\mu\,\nu\,\alpha}) \approx \frac{\left[\frac{\mu}{2}F_{\mu\,\nu\,\alpha}\left(1+\frac{\mu}{\nu}\right)\right]^{\frac{1}{2}\left(\frac{\mu}{2}-1\right)} e^{-\frac{\mu F_{\mu\,\nu\,\alpha}}{4}\left(1+\frac{\mu}{\nu}\right)}}{\frac{\mu}{2}\,\Gamma\left(\frac{\mu}{2}\right)}$$
$$\cdot\,\mathrm{WhittakerM}\left(\frac{1}{2}\left(\frac{\mu}{2}-1\right),\frac{\mu}{4},\frac{\mu F_{\mu\,\nu\,\alpha}}{2}\left(1+\frac{\mu}{\nu}\right)\right) \qquad (7.34)$$
$$= 1-\alpha$$

Introduziremos a seguir o tratamento feito com os dados experimentais para a realização do teste F. Seja um conjunto de clãs, CL1, CL2, ..., CLN, com Φ_1, Φ_2, ..., Φ_N famílias, respectivamente. Após a restrição feita ao banco de dados de selecionarmos famílias pertencentes a clãs que contenham blocos de m linhas por n colunas, o número total de famílias são então reduzidos a φ_1, φ_2, ..., φ_N. Destes clãs com blocos representativos $(m \times n)$, apenas os que contêm um mínimo de cinco famílias são utilizados no teste.

Para uma coluna j, a média da entropia de probabilidade simples das famílias de um clã l é dada por:

$$\langle (HC)_j(\varphi_l)\rangle = \frac{1}{\varphi_l}\sum_{p=1}^{\varphi_l}(HC)_j^p(\varphi_l)\,, \qquad (7.35)$$

onde $p = 1, 2, \ldots, \varphi_l$ são as famílias que compõem o l-ésimo clã. A j-ésima coluna pertence ao intervalo $j = 1, 2, \ldots, n$ e o l-ésimo clã ao intervalo $l = 1, 2, \ldots, N$. Aqui a equação é escrita genericamente com a entropia de Havrda-Charvat, mas pode ser alterada para as outras medidas de entropia sem perda de generalidade. De forma similar, temos a média de um clã da entropia de um par de colunas jk:

$$\langle (HC)_{jk}(\varphi_l)\rangle = \frac{1}{\varphi_l}\sum_{p=1}^{\varphi_l}(HC)_{jk}^p(\varphi_l)\,, \qquad (7.36)$$

com $j = 1, 2, \ldots, (n-1)$ e $k = (j+1), (j+2), \ldots, n$. As variabilidades por coluna e por par de colunas são, respectivamente:

$$D(HC)_j^p(\varphi_l) = (HC)_j^p(\varphi_l) - \langle (HC)_j(\varphi_l)\rangle \qquad (7.37)$$

e

$$D(HC)_{jk}^p(\varphi_l) = (HC)_{jk}^p(\varphi_l) - \langle (HC)_{jk}(\varphi_l)\rangle \qquad (7.38)$$

A partir das equações (7.35) e (7.37), obtemos que:

$$\sum_{p=1}^{\varphi_l} D(HC)_j^p(\varphi_l) = 0\,, \qquad (7.39)$$

e, de forma similar, a partir das equações (7.36) e (7.38), obtemos:

$$\sum_{p=1}^{\varphi_l} D(HC)^p_{jk}(\varphi_l) = 0\,, \tag{7.40}$$

ambas com $(\varphi_l - 1)$ termos independentes.

Os desvios padrões por coluna, $\sigma_{j_{\varphi_l}}$, e por par de colunas, $\sigma_{jk_{\varphi_l}}$, associados às distribuições das entropias das famílias do l-ésimo clã, podem ser obtidos em função das variabilidades através das equações:

$$(\varphi_l - 1)\sigma^2_{j_{\varphi_l}} = \sum_{p=1}^{\varphi_l} \left(D(HC)^p_j(\varphi_l)\right)^2 \tag{7.41}$$

e

$$(\varphi_l - 1)\sigma^2_{jk_{\varphi_l}} = \sum_{p=1}^{\varphi_l} \left(D(HC)^p_{jk}(\varphi_l)\right)^2\,. \tag{7.42}$$

Assim, o l-ésimo clã com média de entropia por coluna $\langle (HC)_j(\Phi_l)\rangle$ e desvio padrão $\sigma_{j_{\Phi_l}}$, é representado pela média e desvio $\langle (HC)_j(\varphi_l)\rangle$ e $\sigma_{j_{\varphi_l}}$, respectivamente (Figura 7.5). O mesmo ocorre com a média e o desvio de entropia de pares de colunas $\langle (HC)_{jk}(\varphi_l)\rangle$ e $\sigma_{jk_{\varphi_l}}$.

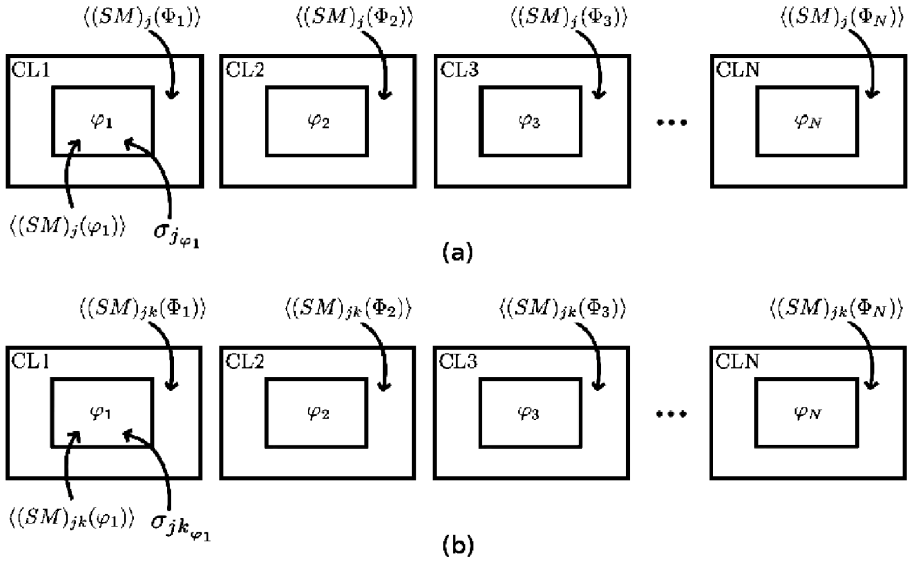


Figura 7.5: Clãs e amostras com restrições de blocos $(m \times n)$ de aminoácidos. (a) Entropia de distribuição de aminoácidos em uma coluna; (b) Entropia de distribuição de aminoácidos em um par de colunas.

A média total (todas as famílias de todos os clãs) para uma coluna j é, por sua vez:

$$\langle (HC)_j\rangle = \frac{1}{\sum_{l=1}^{N} \varphi_l} \sum_{l=1}^{N} \sum_{p=1}^{\varphi_l} (HC)^p_j(\varphi_l)\,. \tag{7.43}$$

E a média total para um par de colunas é:

$$\langle (HC)_{jk} \rangle = \frac{1}{\sum_{l=1}^{N} \varphi_l} \sum_{l=1}^{N} \sum_{p=1}^{\varphi_l} (HC)_{jk}^{p}(\varphi_l)\,. \tag{7.44}$$

As variabilidades por coluna e por par de colunas sobre a média total são expressas, respectivamente, como:

$$\Delta(HC)_{j}^{p}(\varphi_l) = (HC)_{j}^{p}(\varphi_l) - \langle (HC)_j \rangle \tag{7.45}$$

$$\Delta(HC)_{jk}^{p}(\varphi_l) = (HC)_{jk}^{p}(\varphi_l) - \langle (HC)_{jk} \rangle \tag{7.46}$$

Das equações (7.43) e (7.45), e das equações (7.44) e (7.46), obtemos, respectivamente:

$$\sum_{l=1}^{N} \sum_{p=1}^{\varphi_l} \Delta(HC)_{j}^{p}(\varphi_l) = 0 \tag{7.47}$$

e

$$\sum_{l=1}^{N} \sum_{p=1}^{\varphi_l} \Delta(HC)_{jk}^{p}(\varphi_l) = 0\,, \tag{7.48}$$

ambas com $\left(\sum_{l=1}^{N} \varphi_l - 1 \right)$ termos independentes. E, pelas equações (7.35) e (7.43), e (7.36) e (7.44), temos, respectivamente:

$$\sum_{l=1}^{N} \varphi_l \left(\langle (HC)_j(\varphi_l) \rangle - \langle (HC)_j \rangle \right) = 0 \tag{7.49}$$

e

$$\sum_{l=1}^{N} \varphi_l \left(\langle (HC)_{jk}(\varphi_l) \rangle - \langle (HC)_{jk} \rangle \right) = 0\,, \tag{7.50}$$

ambas com $(N-1)$ termos independentes.

Os desvios padrões das médias totais de entropias de probabilidade simples (σ_j) e de probabilidade conjunta (σ_{jk}) podem ser obtidos em função das variabilidades sobre a média total, respectivamente, através das equações:

$$\left(\sum_{l=1}^{N} \varphi_l - 1 \right) \sigma_j^2 = \sum_{l=1}^{N} \sum_{p=1}^{\varphi_l} \left(\Delta(HC)_{j}^{p}(\varphi_l) \right)^2 \tag{7.51}$$

e

$$\left(\sum_{l=1}^{N} \varphi_l - 1 \right) \sigma_{jk}^2 = \sum_{l=1}^{N} \sum_{p=1}^{\varphi_l} \left(\Delta(HC)_{jk}^{p}(\varphi_l) \right)^2 \tag{7.52}$$

A variabilidade entre as médias dos clãs em relação à média total, SSG (*Sum*

of Squares of samples between Groups — soma dos quadrados das amostras entre grupos), é definida como:

$$SSG \equiv \sum_{l=1}^{N} \varphi_l \left(\langle (HC)_j(\varphi_l) \rangle - \langle (HC)_j \rangle \right)^2 , \tag{7.53}$$

e com ela temos a seguinte relação:

$$\underbrace{\left(\sum_{l=1}^{N} \varphi_l - 1 \right) \sigma_j^2}_{\text{SST}} = \underbrace{\sum_{l=1}^{N} (\varphi_l - 1) \sigma_{j_{\varphi_l}}^2}_{\text{SSE}} + \underbrace{\sum_{l=1}^{N} \varphi_l \left(\langle (HC)_j(\varphi_l) \rangle - \langle (HC)_j \rangle \right)^2}_{\text{SSG}} , \tag{7.54}$$

onde SST (*Sum of Squares of Total samples* — soma dos quadrados de todas as amostras) é a variabilidade entre as entropias de cada família em relação à média total (7.43), e SSE (*Sum of Squares of samples within Groups* — soma dos quadrados das amostras pertencentes aos grupos) é a variabilidade entre as entropias de cada família em relação à média dos clãs (7.35) aos quais elas pertencem. Checando o número de termos independentes, temos:

$$\sum_{l=1}^{N} \varphi_l - 1 = \sum_{l=1}^{N} (\varphi_l - 1) + N - 1 = \sum_{l=1}^{N} \varphi_l - N + N - 1 \, .$$

Equivalentemente às equações (7.53) e (7.54), temos para o caso de entropias de pares de colunas, respectivamente, as equações:

$$SSG \equiv \sum_{l=1}^{N} \varphi_l \left(\langle (HC)_{jk}(\varphi_l) \rangle - \langle (HC)_{jk} \rangle \right)^2 , \tag{7.55}$$

e

$$\underbrace{\left(\sum_{l=1}^{N} \varphi_l - 1 \right) \sigma_{jk}^2}_{\text{SST}} = \underbrace{\sum_{l=1}^{N} (\varphi_l - 1) \sigma_{jk_{\varphi_l}}^2}_{\text{SSE}} + \underbrace{\sum_{l=1}^{N} \varphi_l \left(\langle (HC)_{jk}(\varphi_l) \rangle - \langle (HC)_{jk} \rangle \right)^2}_{\text{SSG}} . \tag{7.56}$$

Com as variabilidades devidamente apresentadas, podemos agora introduzir os coeficientes de correlação de Fisher para a entropia de probabilidade simples e para a entropia de probabilidade conjunta:

$$F_j = \frac{\frac{\text{SSG}}{N-1}}{\frac{\text{SSE}}{\sum_{l=1}^{N} \varphi_l - N}} = \left(\frac{\sum_{l=1}^{N} \varphi_l - N}{N-1} \right) \cdot \left(\frac{\left(\sum_{l=1}^{N} \varphi_l - 1 \right) \sigma_j^2}{\sum_{l=1}^{N} (\varphi_l - 1) \, \sigma_{j_{\varphi_l}}^2} - 1 \right) , \, j = 1, 2, \ldots, n \tag{7.57}$$

e

$$F_{jk} = \frac{\frac{\mathrm{SSG}}{N-1}}{\frac{\mathrm{SSE}}{\sum_{l=1}^{N}\varphi_l - N}} = \left(\frac{\sum_{l=1}^{N}\varphi_l - N}{N-1}\right) \cdot \left(\frac{\left(\sum_{l=1}^{N}\varphi_l - 1\right)\sigma_{jk}^2}{\sum_{l=1}^{N}\left(\varphi_l - 1\right)\sigma_{jk_{\varphi_l}}^2} - 1 \right) \quad , \quad \substack{j=1,2,\ldots,n-1 \\ k=(j+1),(j+2),\ldots,n} \quad , \tag{7.58}$$

com n testes ANOVA para o cálculo com cada coluna individualmente e $\frac{n(n-1)}{2}$ testes ANOVA para o cálculo com pares de colunas, por valor de parâmetro.

Os parâmetros μ e ν utilizados no cálculo do coeficiente F teórico são, respectivamente, os graus de liberdade do numerador e do denominador do coeficiente experimental, ou seja, temos:

$$\mu = N - 1 \tag{7.59}$$

e

$$\nu = \left(\sum_{l=1}^{N}\varphi_l - N\right) \tag{7.60}$$

7.3 Teste de Hipóteses

Após apresentarmos como são calculados os valores de F teórico e experimental para a realização do teste ANOVA, passamos para as hipóteses em si que serão analizadas. O teste de hipóteses tem como objetivo verificar se é falsa a denominada *Hipótese Nula*, H_0, cuja veracidade é considerada *a priori*. Caso estatisticamente não haja como comprová-la, a *Hipótese Alternativa*, H_a, é considerada.

As hipóteses nula e alternativa para o caso com entropias advindas de probabilidade simples são, respectivamente:

- $H_0 : \langle (HC)_j(\Phi_1)\rangle = \langle (HC)_j(\Phi_2)\rangle = \ldots = \langle (HC)_j(\Phi_N)\rangle \Rightarrow$ invalidação do conceito de clã.
- $H_a : \langle (HC)_j(\Phi_1)\rangle \neq \langle (HC)_j(\Phi_2)\rangle \neq \ldots \neq \langle (HC)_j(\Phi_N)\rangle$ (não sendo todos necessariamente diferentes) $\Rightarrow$ existência de clãs.

e, similarmente para as entropias de probabilidade conjunta, temos:

- $H_0 : \langle (HC)_{jk}(\Phi_1)\rangle = \langle (HC)_{jk}(\Phi_2)\rangle = \ldots = \langle (HC)_{jk}(\Phi_N)\rangle \Rightarrow$ invalidação do conceito de clã.
- $H_a : \langle (HC)_{jk}(\Phi_1)\rangle \neq \langle (HC)_{jk}(\Phi_2)\rangle \neq \ldots \neq \langle (HC)_{jk}(\Phi_N)\rangle$ (não sendo todos necessariamente diferentes) $\Rightarrow$ existência de clãs.

Para que a classificação atual das famílias de domínios de proteínas em clãs seja reconhecida como estatisticamente relevante, o valor experimental deve ser maior do que o valor teórico: $F_j > F_{\mu\,\nu\,\alpha}$ ou $F_{jk} > F_{\mu\,\nu\,\alpha}$.

Algumas suposições em relação aos dados a serem analizados são feitas para a realização do teste [41]. Para cada coluna (ou par de colunas no caso de probabilidades conjuntas) dos blocos representativos $(m \times n)$ temos:

1. As amostras (os valores de entropia das famílias) das N populações (clãs) são independentes.

2. As amostras são obtidas de N diferentes distribuições normais.

3. As N distribuições têm a mesma variância σ^2.

Confiando na robustez da estatística ANOVA as suposições 2 e 3 podem ser relaxadas. Se as dispersões das entropias de cada clã forem aproximadamente iguais, podemos considerar que a terceira suposição não é violada [39].

Foram executados diversos testes com diferentes números de clãs com a utilização de blocos representativos (100×200) [45], e consequentemente de famílias, indicados na Tabela 7.1. A ordem de inclusão dos clãs está indicada na Tabela 7.2. Já para o caso com blocos representativos (100×100), os cálculos de F teórico, assim como o número de clãs e famílias são apresentados na Tabela 7.3. A ordem de inclusão neste caso segue a mesma utilizada para os blocos (100×200) até alcançar os 68 clãs em comum. A partir daí são incluídos os clãs exclusivos para blocos (100×100).

Alguns dos resultados do teste de Fisher da distribuição das entropias Havrda-Charvat para diferentes números de clãs são mostrados nas figuras a seguir. Todos os testes foram feitos com um nível de significância $\alpha = 0.01$. Para as entropias de probabilidade simples (Figuras 7.6, 7.7, 7.8), temos um gráfico bidimensional, onde no eixo horizontal temos as 200 colunas e no eixo vertical os valores dos coeficientes de correlação de Fisher. A reta horizontal que corta o gráfico corresponde ao valor teórico, $F_{\mu\nu\alpha}$, obtido da distribuição de Fisher-Snedecor referente ao número de clãs, famílias e do nível de significância utilizados. Para as entropias de probabilidade conjunta (Figuras 7.9, 7.10, 7.11), o teste de Fisher é representado em um gráfico tridimensional, cujo eixo vertical também indica os valores dos coeficientes de correlação, enquanto no plano horizontal temos os pares de colunas, com um dos eixos (j) adotando valores de 1 a 199, e o outro (k) de 2 a 200, respeitando sempre a ordenação jk (com $k \geq j + 1$). O plano que corta o gráfico corresponde, novamente, ao valor teórico, $F_{\mu\nu\alpha}$. Da mesma forma como é feita a representação dos testes por pares de colunas para a entropia Havrda-Charvat de probabilidade conjunta, temos nas Figuras 7.12, 7.13 e 7.14, os resultados para a entropia de Jaccard associada a entropia de Havrda-Charvat.

Tabela 7.1: Número de clãs em experimentos sucessivos para o caso 100×200, o número de famílias e o valor de F teórico correspondente.

nº de clãs	nº de famílias	$F_{\mu\,\nu\,\alpha}$
4	290	3.85
6	325	3.07
8	372	2.69
13	412	2.23
19	471	1.97
21	490	1.92
22	500	1.89
23	509	1.87
24	557	1.84
26	584	1.81
29	605	1.76
30	639	1.74
31	658	1.73
33	688	1.70
36	712	1.67
38	726	1.65
48	884	1.57
56	953	1.52
59	980	1.50
61	1029	1.50
68	1069	1.47

Os pontos acima das retas ou dos planos (nos casos de probabilidade simples e conjunta, respectivamente) indicam a rejeição da Hipótese Nula. Fazemos então a contagem de todos os pontos acima do F teórico para todos os testes com diferentes valores de parâmetro e diferentes números de famílias. Os gráficos de contagem de pontos acima das retas e planos contra o número de famílias são apresentados nas Figuras 7.15, 7.17 e 7.19 para os blocos representativos (100×200), e nas Figuras 7.16, 7.18 e 7.20 para os blocos (100×100).

Na Figura 7.21 temos a comparação entre os testes com os blocos representativos (100×200) e (100×100) para a entropia de Havrda-Charvat e de Jaccard. Como o número de pontos e o número de famílias são diferentes, no eixo vertical temos o percentual de pontos acima das retas ou planos e no eixo horizontal temos o número de clãs. E, como para os blocos (100×200) temos apenas 68 clãs, os resultados dos blocos (100×100) são apresentados apenas até este valor nesta comparação.

Tabela 7.2: Ordem de inclusão dos clãs nos testes com blocos representativos 100×200.

nº de clãs	Clãs adicionados
4	CL0020, CL0023, CL0028, CL0063
6	CL0123, CL0186
8	CL0192, CL0236
13	CL0246, CL0257, CL0260, CL0268, CL0295
19	CL0004, CL0013, CL0014, CL0015, CL0016, CL0027
21	CL0029, CL0030
22	CL0034
23	CL0035
24	CL0036
26	CL0039, CL0040
29	CL0044, CL0046, CL0052
30	CL0058
31	CL0059
33	CL0061, CL0062
36	CL0064, CL0088, CL0093
38	CL0103, CL0105
48	CL0108, CL0110, CL0111, CL0113, CL0118, CL0125, CL0126, CL0127, CL0128, CL0137
56	CL0142, CL0144, CL0149, CL0151, CL0158, CL0163, CL0177, CL0179
59	CL0181, CL0182, CL0184
61	CL0193, CL0219
68	CL0254, CL0264, CL0270, CL0286, CL0292, CL0316, CL0373

Tabela 7.3: Número de clãs em experimentos sucessivos para o caso 100×100, o número de famílias e o valor de F teórico correspondente.

nº de clãs	nº de famílias	$F_{\mu\nu\alpha}$
4	447	3.83
6	520	3.05
8	595	2.67
13	672	2.21
19	747	1.96
21	794	1.90
22	808	1.88
23	818	1.85
24	871	1.83
26	906	1.79
29	947	1.74
30	983	1.73
31	1008	1.72
33	1038	1.69
36	1067	1.66
38	1088	1.64
48	1326	1.56
56	1404	1.51
59	1444	1.50
61	1517	1.49
68	1564	1.46
77	1651	1.43
85	1709	1.41
91	1755	1.39
97	1815	1.38
105	1891	1.36
109	1933	1.36
117	1999	1.34
123	2043	1.33
131	2097	1.32
141	2151	1.31
146	2180	1.31

Tabela 7.4: Ordem de inclusão dos clãs nos testes com blocos representativos 100×100.

nº de clãs	Clãs adicionados
4	CL0020, CL0023, CL0028, CL0063
6	CL0123, CL0186
8	CL0192, CL0236
13	CL0246, CL0257, CL0260, CL0268, CL0295
19	CL0004, CL0013, CL0014, CL0015, CL0016, CL0027
21	CL0029, CL0030
22	CL0034
23	CL0035
24	CL0036
26	CL0039, CL0040
29	CL0044, CL0046, CL0052
30	CL0058
31	CL0059
33	CL0061, CL0062
36	CL0064, CL0088, CL0093
38	CL0103, CL0105
48	CL0108, CL0110, CL0111, CL0113, CL0118, CL0125, CL0126, CL0127, CL0128, CL0137
56	CL0142, CL0144, CL0149, CL0151, CL0158, CL0163, CL0177, CL0179
59	CL0181, CL0182, CL0184
61	CL0193, CL0219
68	CL0254, CL0264, CL0270, CL0286, CL0292, CL0316, CL0373
77	CL0021, CL0025, CL0026, CL0031, CL0032, CL0037, CL0042, CL0050, CL0051
85	CL0060, CL0065, CL0066, CL0072, CL0073, CL0085, CL0098, CL0104
91	CL0109, CL0112, CL0115, CL0116, CL0124, CL0129
97	CL0131, CL0135, CL0143, CL0145, CL0154, CL0159
105	CL0161, CL0169, CL0170, CL0172, CL0174, CL0178, CL0183, CL0198
109	CL0202, CL0209, CL0212, CL0220
117	CL0221, CL0222, CL0237, CL0255, CL0263, CL0265, CL0266, CL0280
123	CL0291, CL0307, CL0310, CL0315, CL0317, CL0318
131	CL0322, CL0328, CL0329, CL0330, CL0331, CL0334, CL0336, CL0339
141	CL0342, CL0344, CL0362, CL0366, CL0375, CL0381, CL0382, CL0401, CL0413, CL0442
146	CL0479, CL0523, CL0526, CL0533, CL0549

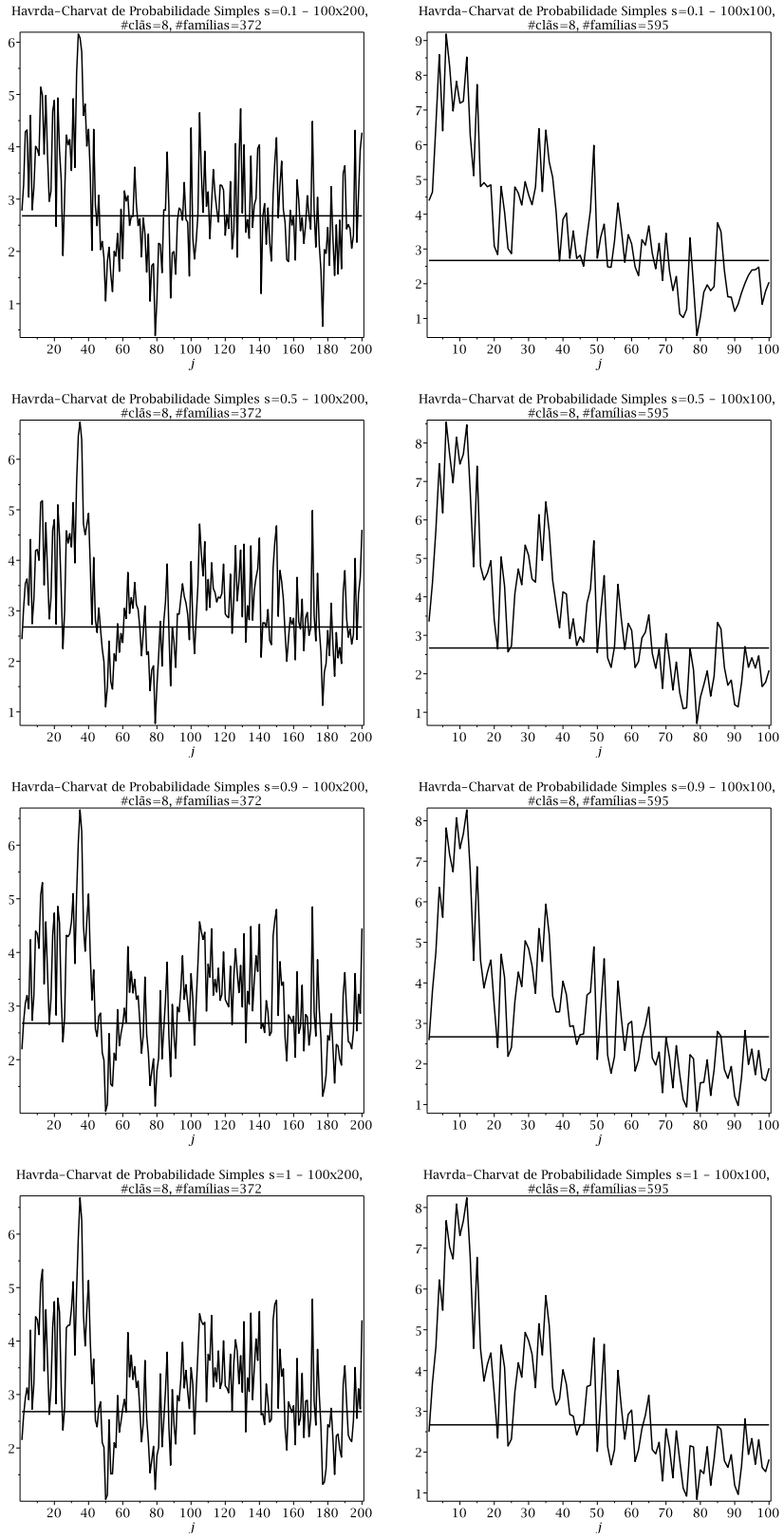


Figura 7.6: Variação dos valores experimentais de F contra as colunas correspondentes para um número fixo de clãs (8) para a entropia Havrda-Charvat de probabilidade simples. O valor teórico de F é dado pela altura da linha reta. Na esquerda, são apresentados os resultados com blocos representativos 100×200, e na direita 100×100. Da primeira a quarta linha temos s igual a 0.1, 0.5, 0.9 e 1.

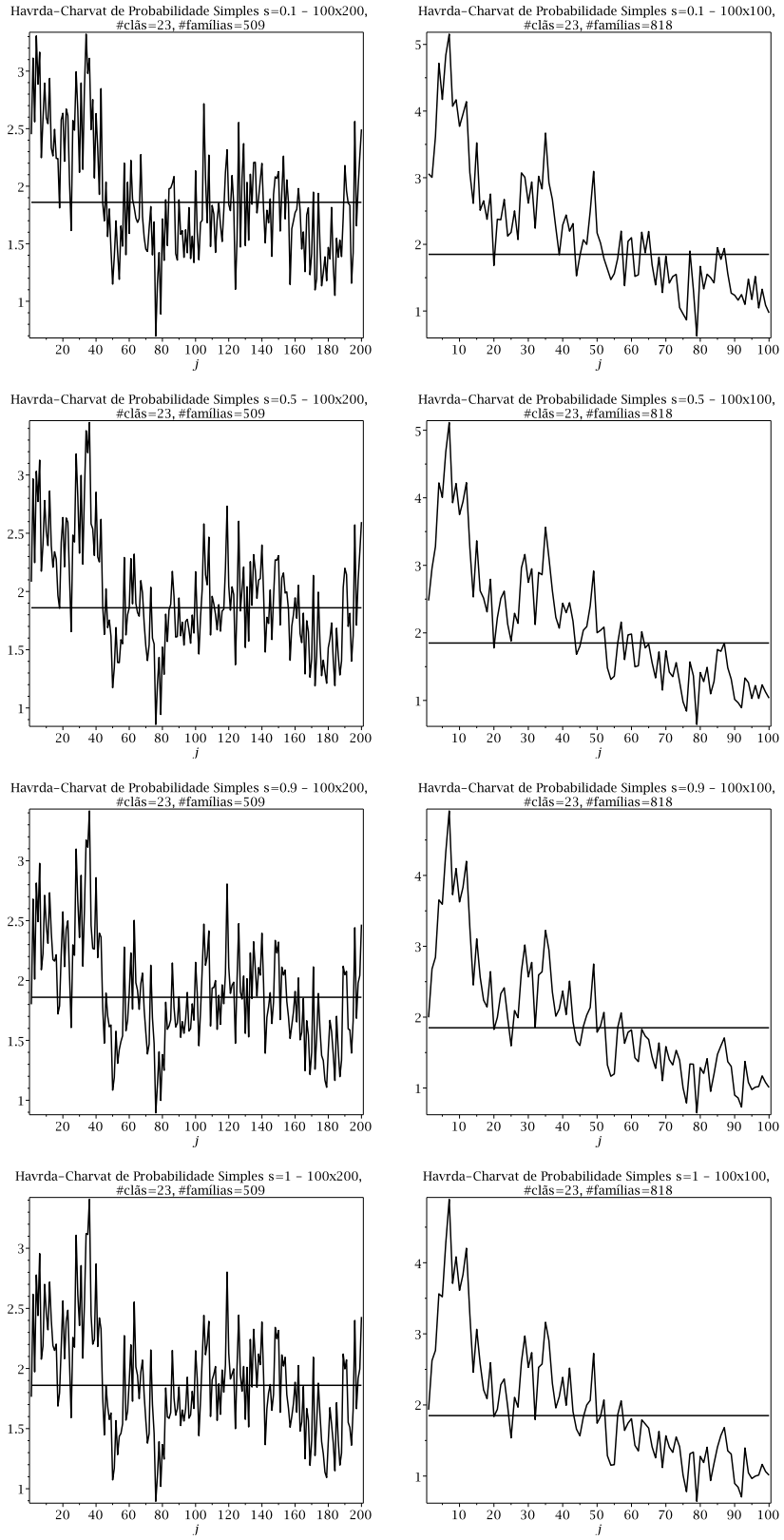


Figura 7.7: Variação dos valores experimentais de F contra as colunas correspondentes para um número fixo de clãs (23) para a entropia Havrda-Charvat de probabilidade simples. O valor teórico de F é dado pela altura da linha reta. Na esquerda, são apresentados os resultados com blocos representativos 100×200, e na direita 100×100. Da primeira a quarta linha temos s igual a 0.1, 0.5, 0.9 e 1.

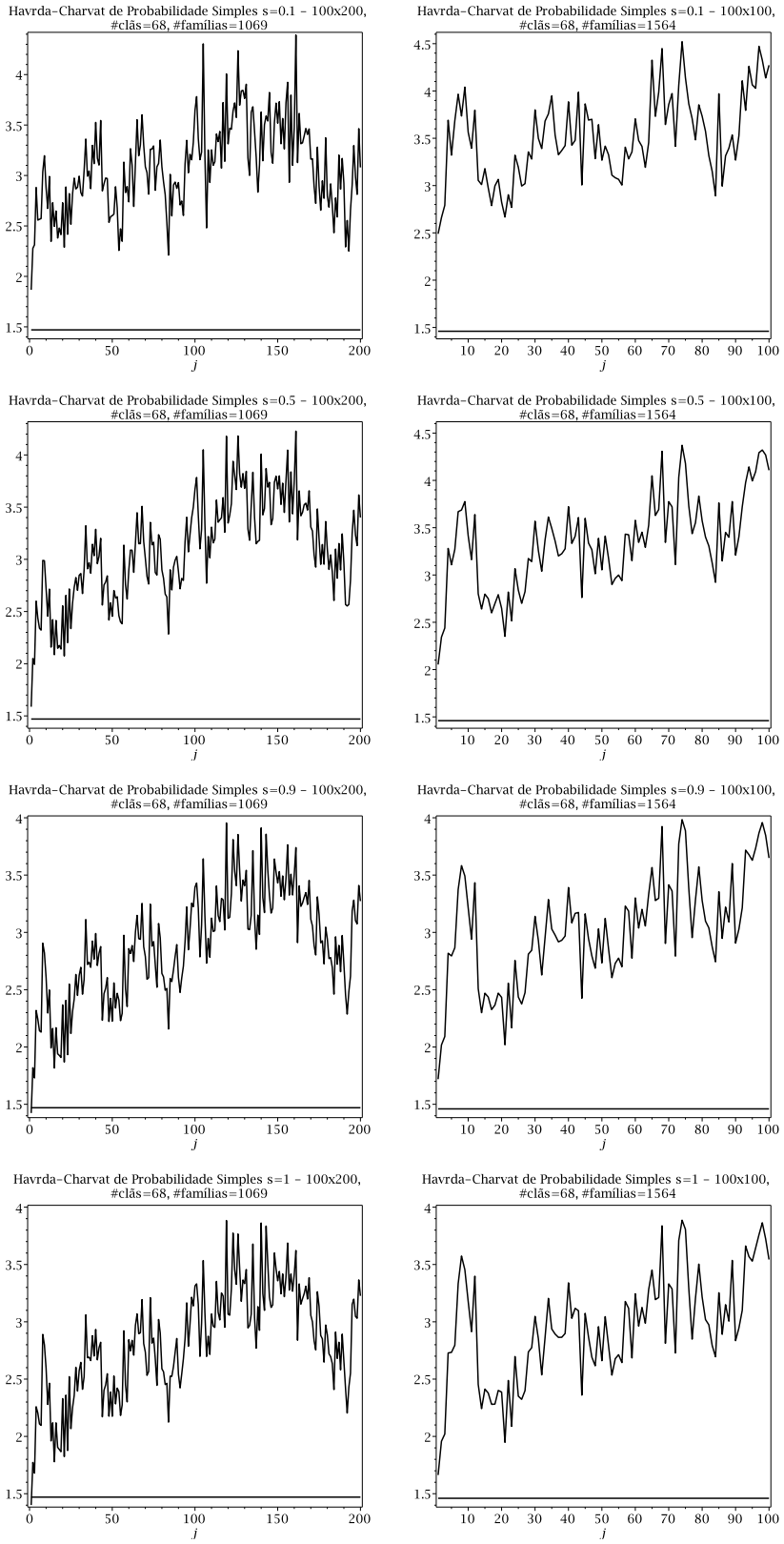


Figura 7.8: Variação dos valores experimentais de F contra as colunas correspondentes para um número fixo de clãs (68) para a entropia Havrda-Charvat de probabilidade simples. O valor teórico de F é dado pela altura da linha reta. Na esquerda, são apresentados os resultados com blocos representativos 100×200, e na direita 100×100. Da primeira a quarta linha temos s igual a 0.1, 0.5, 0.9 e 1.

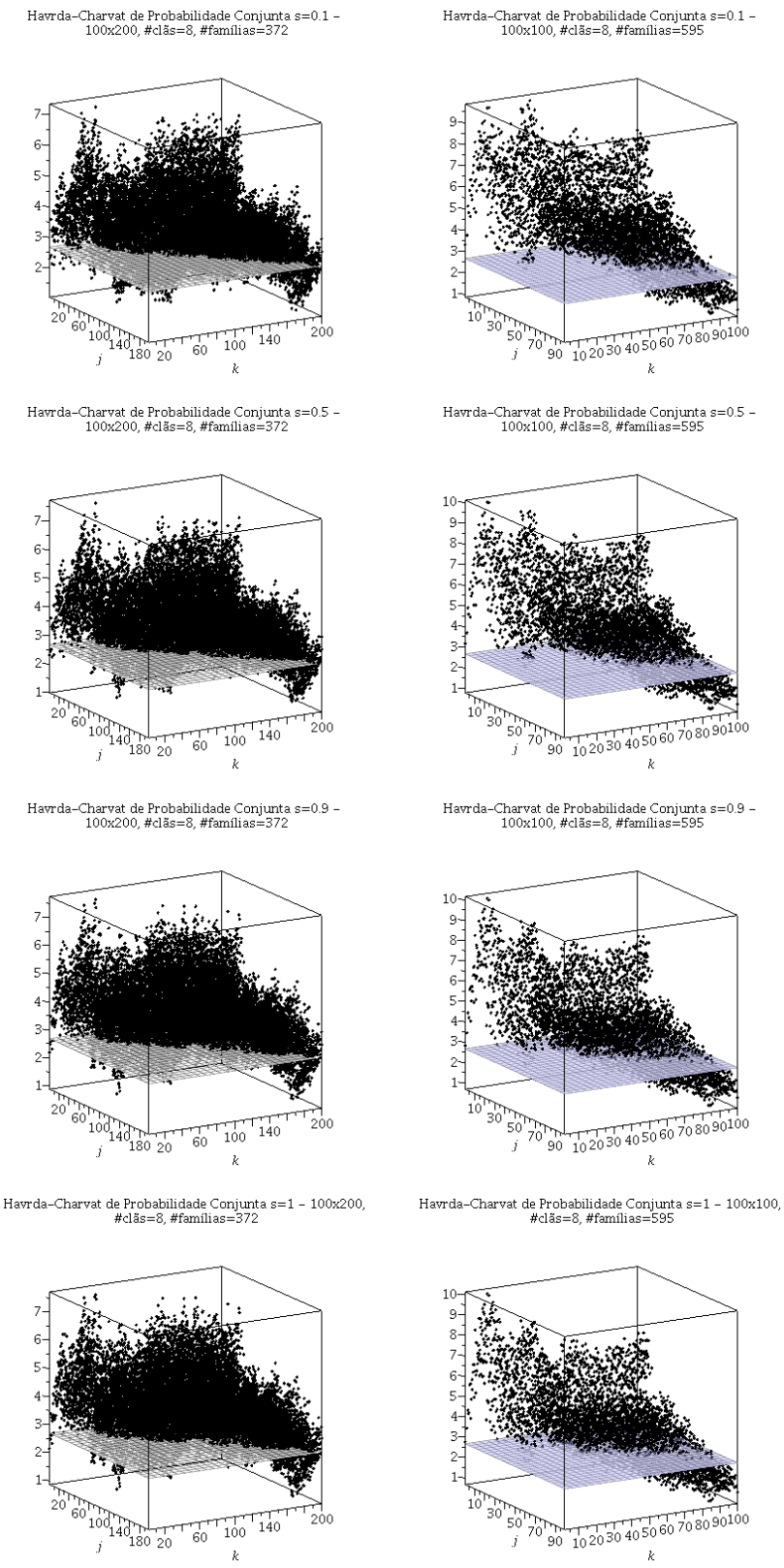


Figura 7.9: Variação dos valores experimentais de F contra as colunas correspondentes para um número fixo de clãs (8) para a entropia Havrda-Charvat de probabilidade conjunta. O valor teórico de F é dado pela altura do plano. Na esquerda, são apresentados os resultados com blocos representativos 100×200, e na direita 100×100. Da primeira a quarta linha temos s igual a 0.1, 0.5, 0.9 e 1.

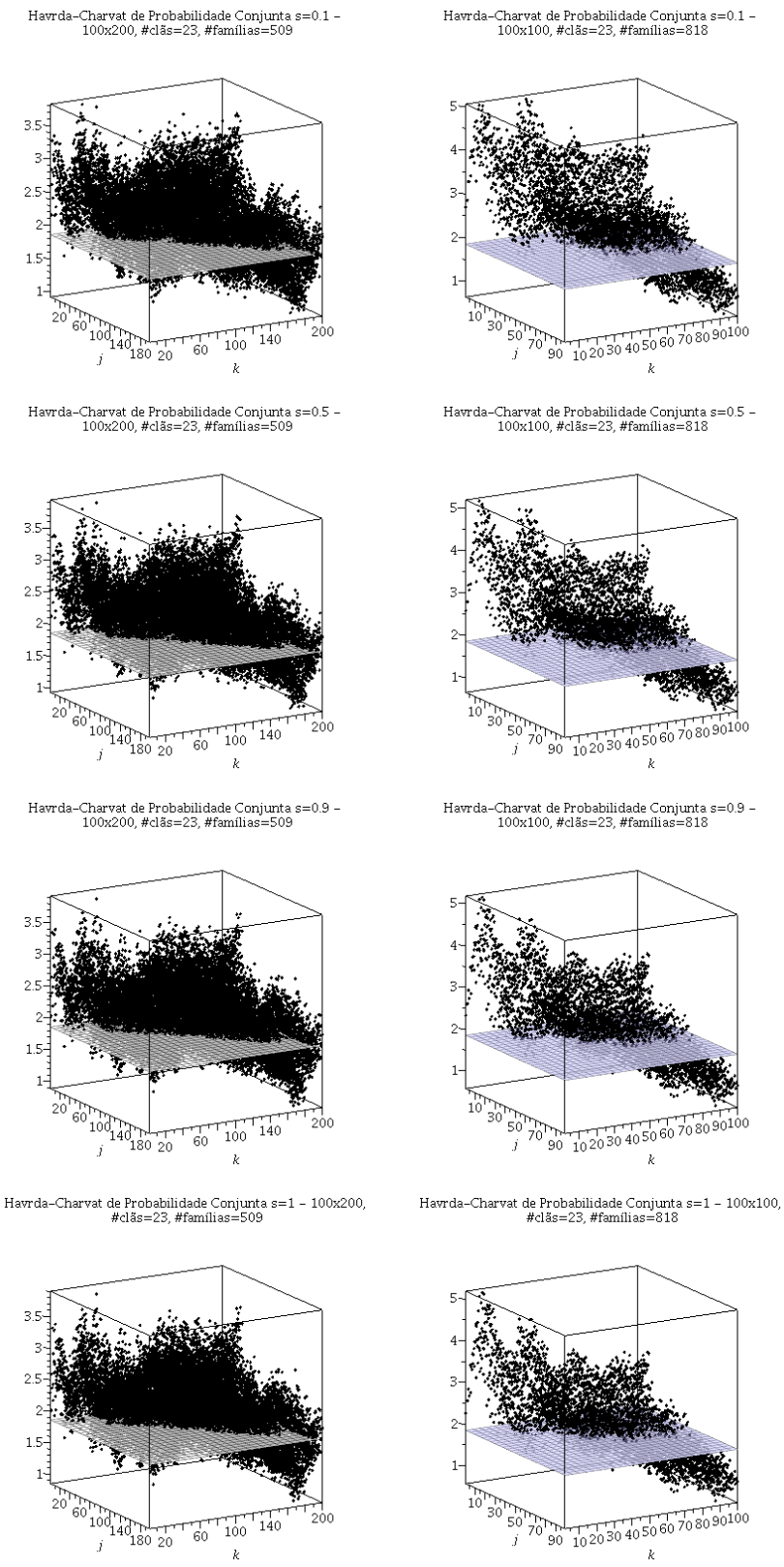


Figura 7.10: Variação dos valores experimentais de F contra as colunas correspondentes para um número fixo de clãs (23) para a entropia Havrda-Charvat de probabilidade conjunta. O valor teórico de F é dado pela altura do plano. Na esquerda, são apresentados os resultados com blocos representativos 100×200, e na direita 100×100. Da primeira a quarta linha temos s igual a 0.1, 0.5, 0.9 e 1.

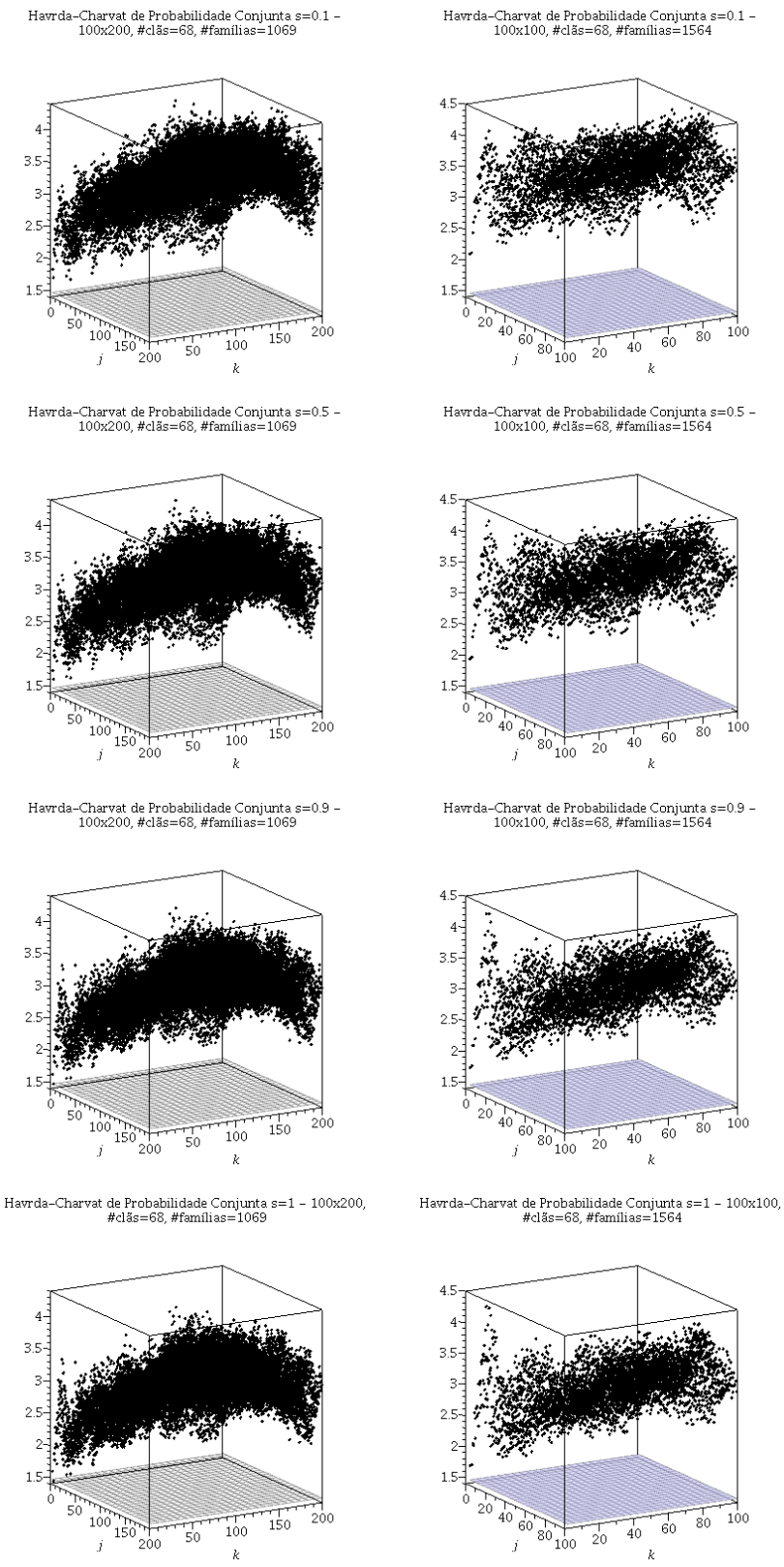


Figura 7.11: Variação dos valores experimentais de F contra as colunas correspondentes para um número fixo de clãs (68) para a entropia Havrda-Charvat de probabilidade conjunta. O valor teórico de F é dado pela altura do plano. Na esquerda, são apresentados os resultados com blocos representativos 100×200, e na direita 100×100. Da primeira a quarta linha temos s igual a 0.1, 0.5, 0.9 e 1.

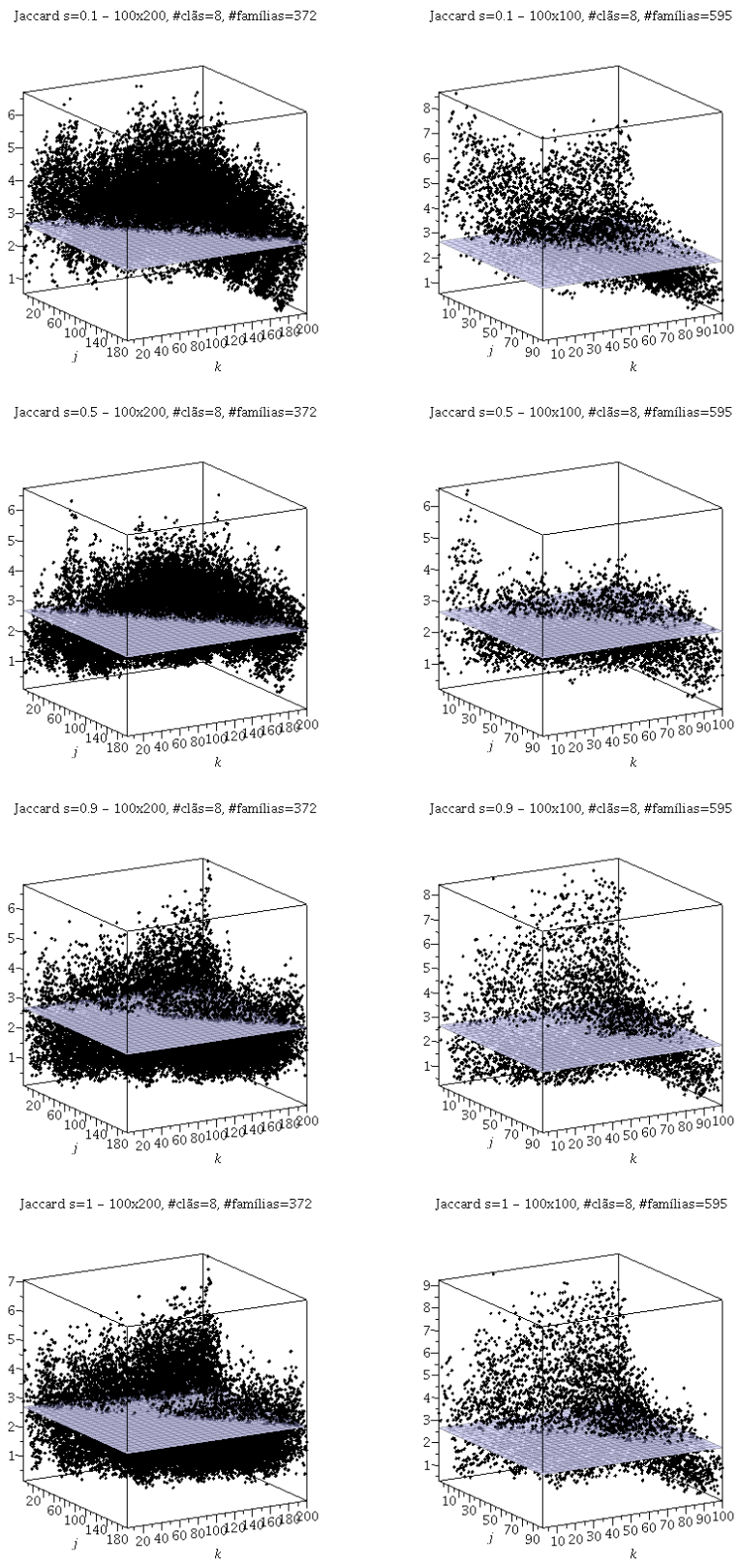


Figura 7.12: Variação dos valores experimentais de F contra as colunas correspondentes para um número fixo de clãs (8) para a entropia Jaccard associada a entropia de Havrda-Charvat. O valor teórico de F é dado pela altura do plano. Na esquerda, são apresentados os resultados com blocos representativos 100×200, e na direita 100×100. Da primeira a quarta linha temos s igual a 0.1, 0.5, 0.9 e 1.

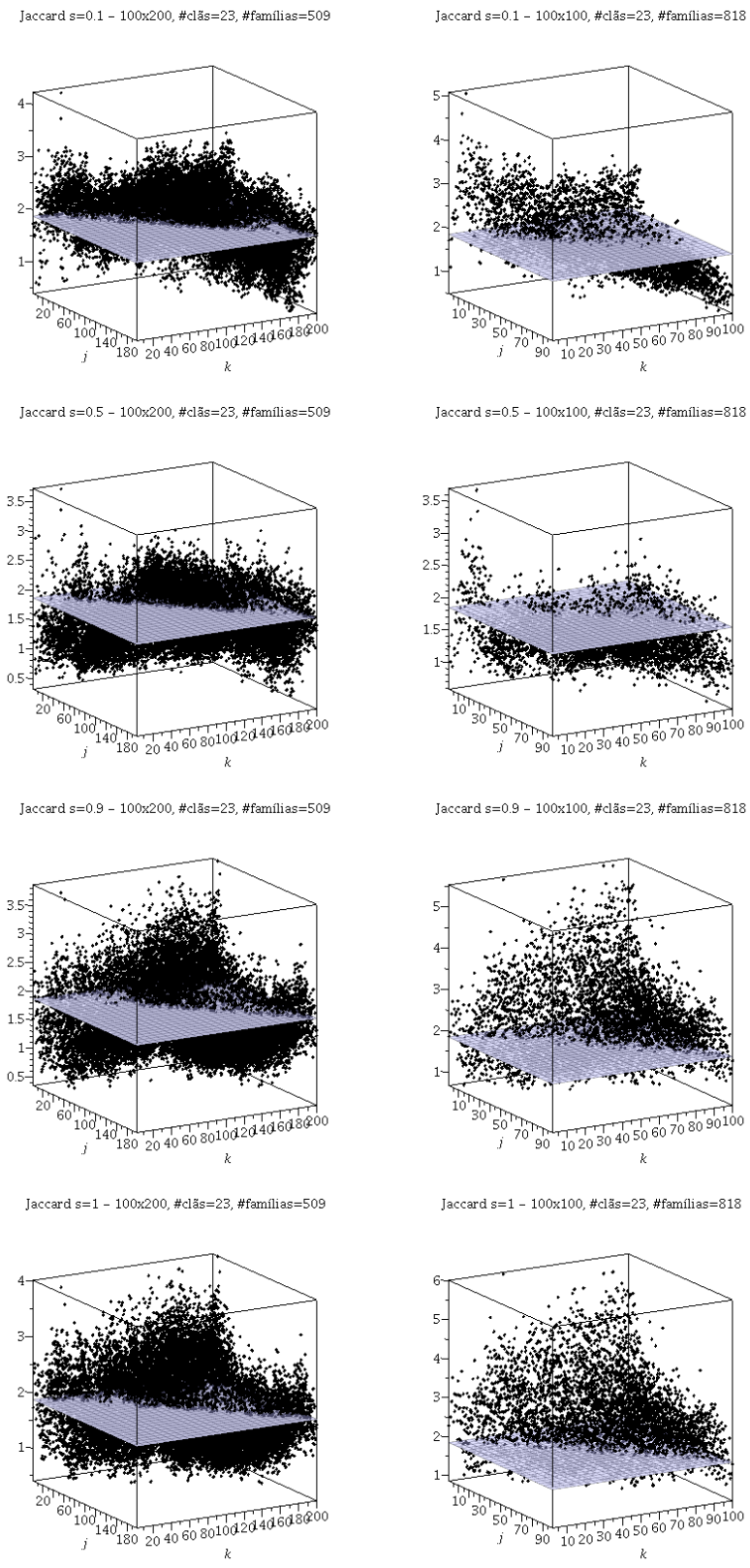


Figura 7.13: Variação dos valores experimentais de F contra as colunas correspondentes para um número fixo de clãs (23) para a entropia Jaccard associada a entropia de Havrda-Charvat. O valor teórico de F é dado pela altura do plano. Na esquerda, são apresentados os resultados com blocos representativos 100×200, e na direita 100×100. Da primeira a quarta linha temos s igual a 0.1, 0.5, 0.9 e 1.

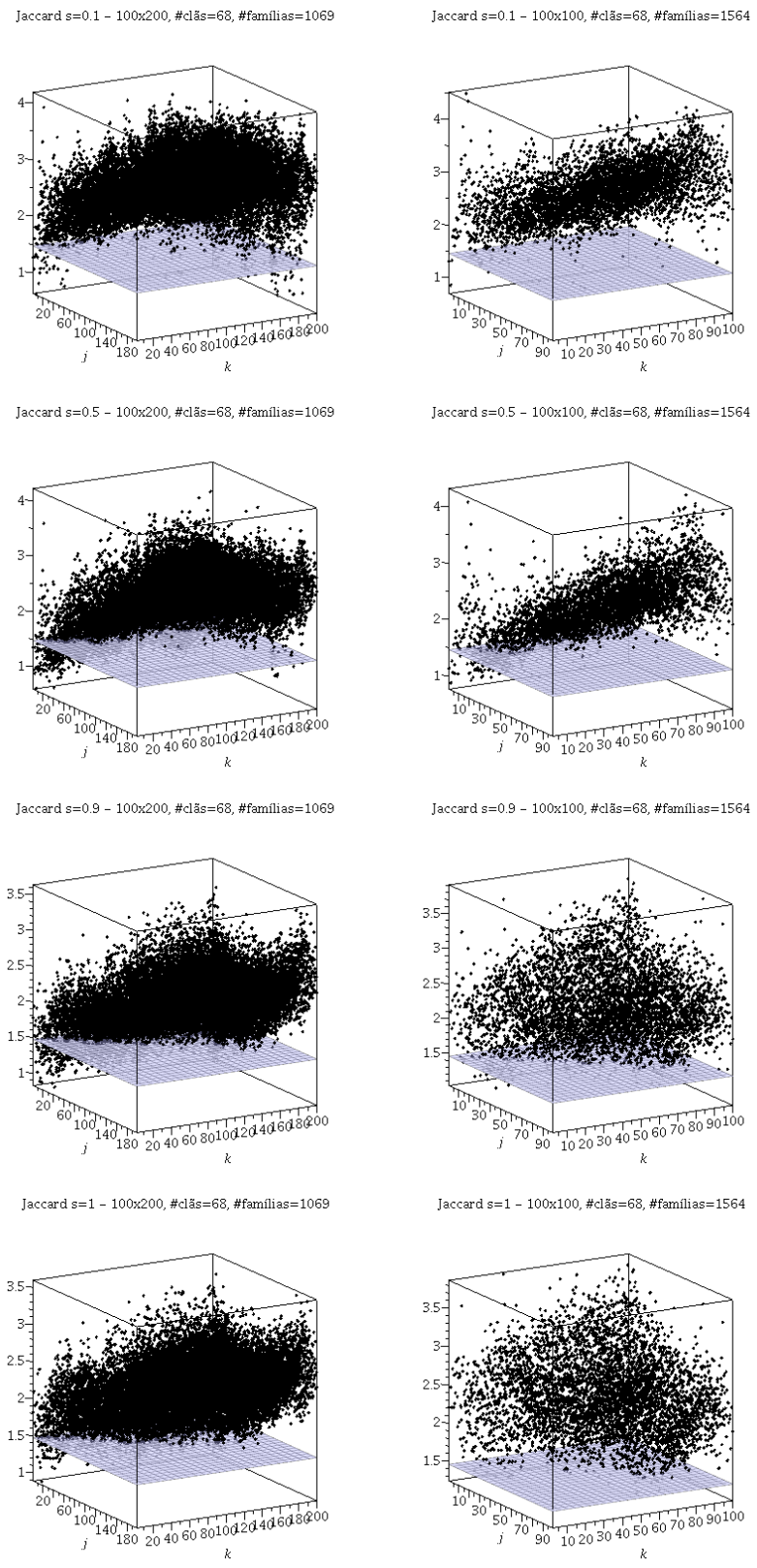
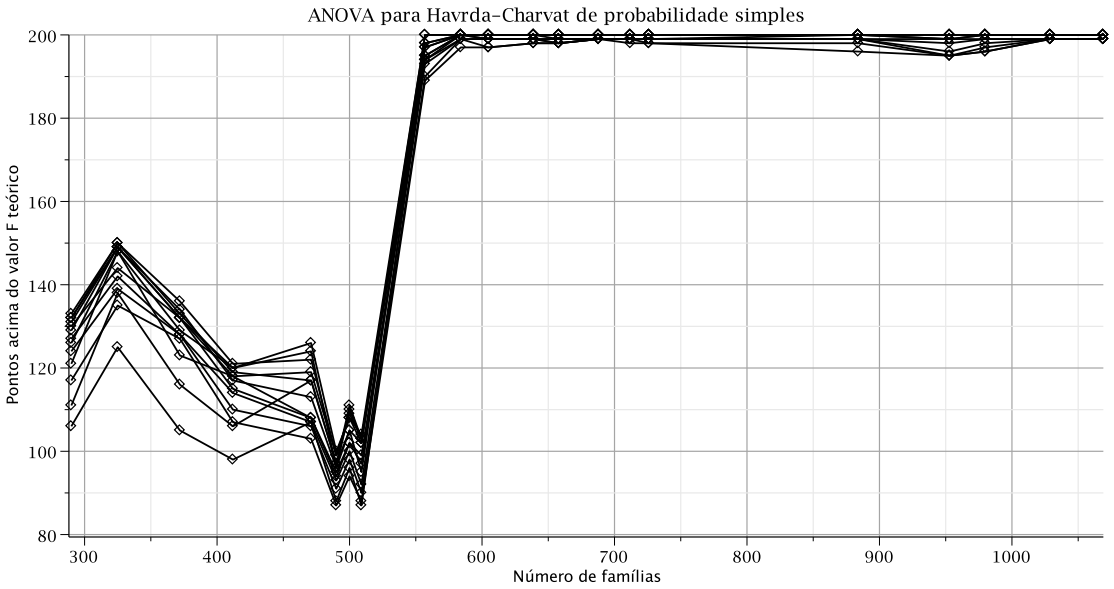
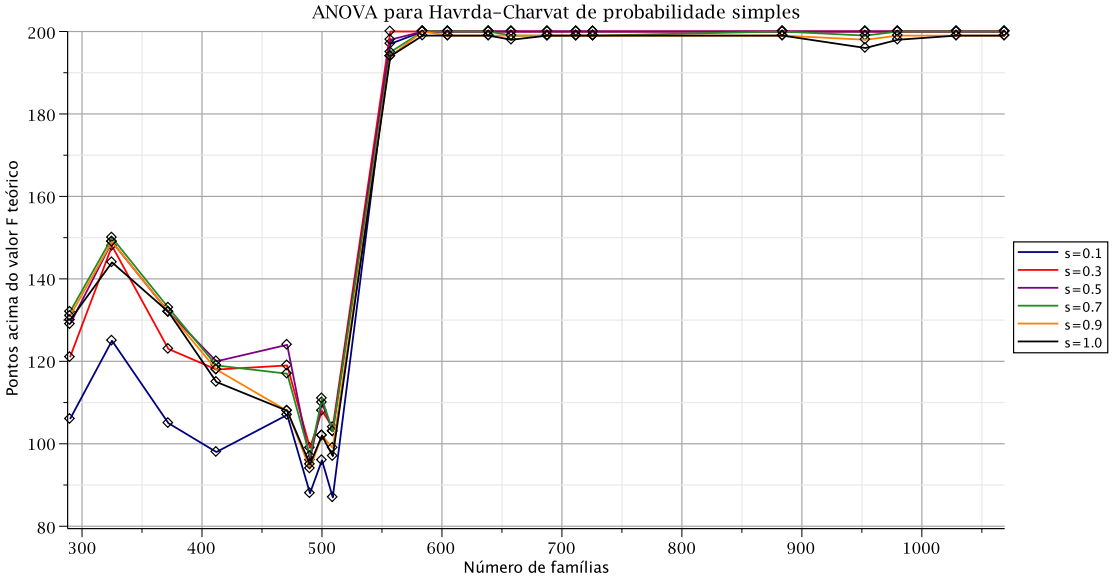


Figura 7.14: Variação dos valores experimentais de F contra as colunas correspondentes para um número fixo de clãs (68) para a entropia Jaccard associada a entropia de Havrda-Charvat. O valor teórico de F é dado pela altura do plano. Na esquerda, são apresentados os resultados com blocos representativos 100×200, e na direita 100×100. Da primeira a quarta linha temos s igual a 0.1, 0.5, 0.9 e 1.

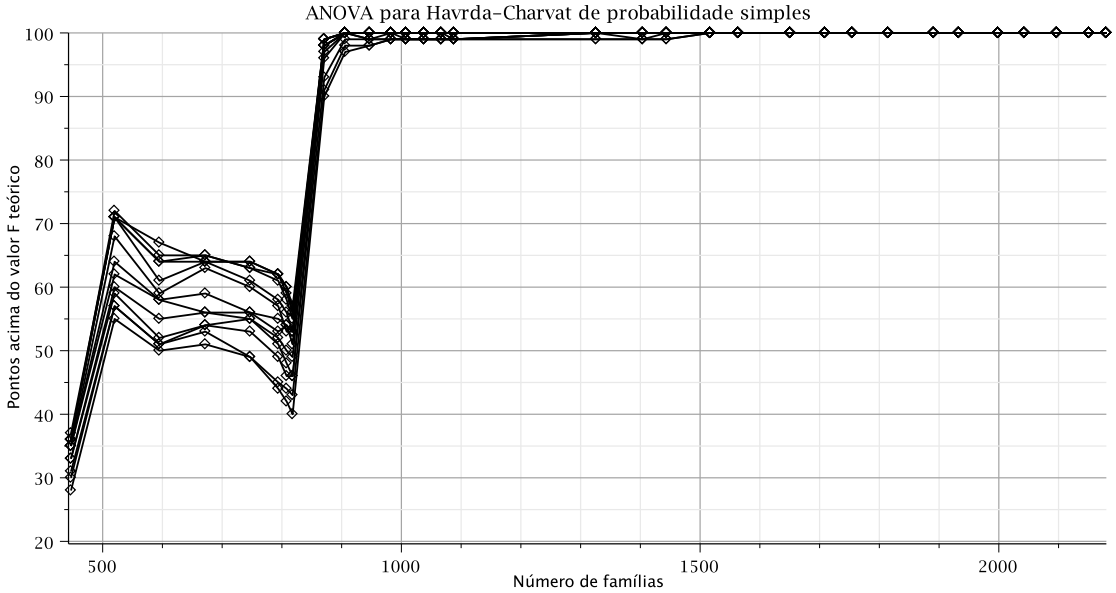


(a)

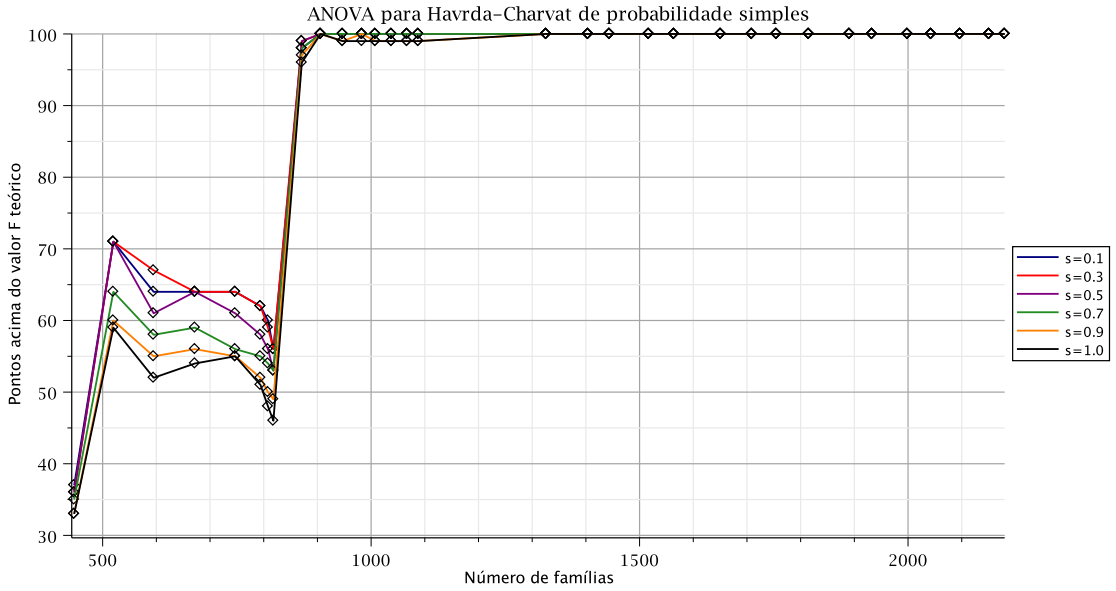


(b)

Figura 7.15: Número de valores de F experimental acima dos valores de F teórico ($F_j > F_{\mu\nu\alpha}$) para um número cumulativo de famílias (probabilidades $p_j(a)$) para a entropia Havrda-Charvat com blocos representativos (100×200). (a) Curvas correspondentes a 13 valores do parâmetro s: 0.1, 0.2, ..., 1.3. (b) Curvas com os valores de parâmetro $s = 0.1$, 0.3, 0.5, 0.7, 0.9 e 1.0.

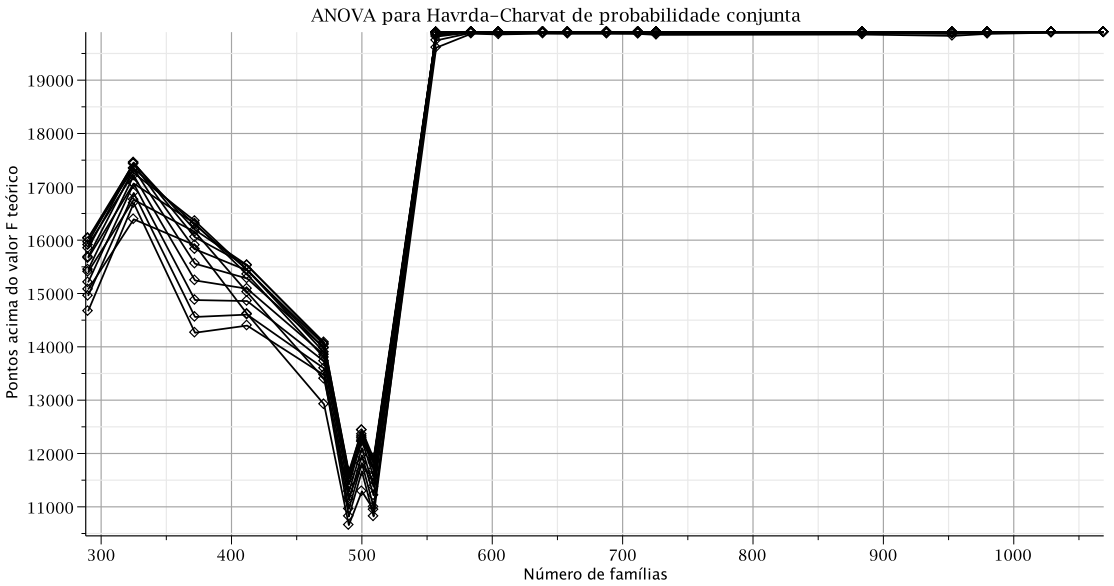


(a)

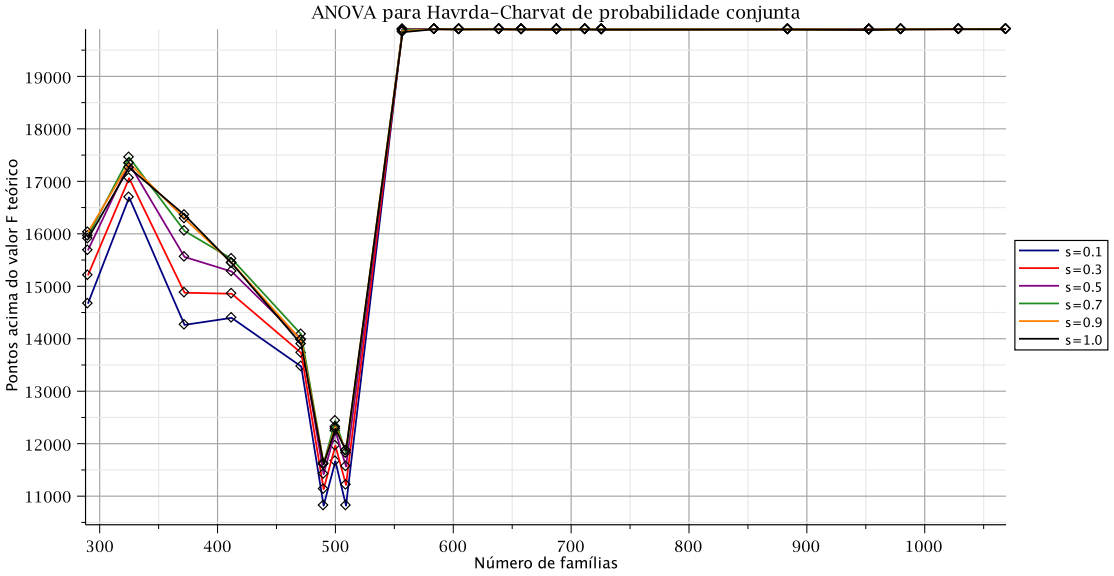


(b)

Figura 7.16: Número de valores de F experimental acima dos valores de F teórico ($F_j > F_{\mu\nu\alpha}$) para um número cumulativo de famílias (probabilidades $p_j(a)$) para a entropia Havrda-Charvat com blocos representativos (100×100). (a) Curvas correspondentes a 13 valores do parâmetro s: 0.1, 0.2, ..., 1.3. (b) Curvas com os valores de parâmetro $s = 0.1$, 0.3, 0.5, 0.7, 0.9 e 1.0.

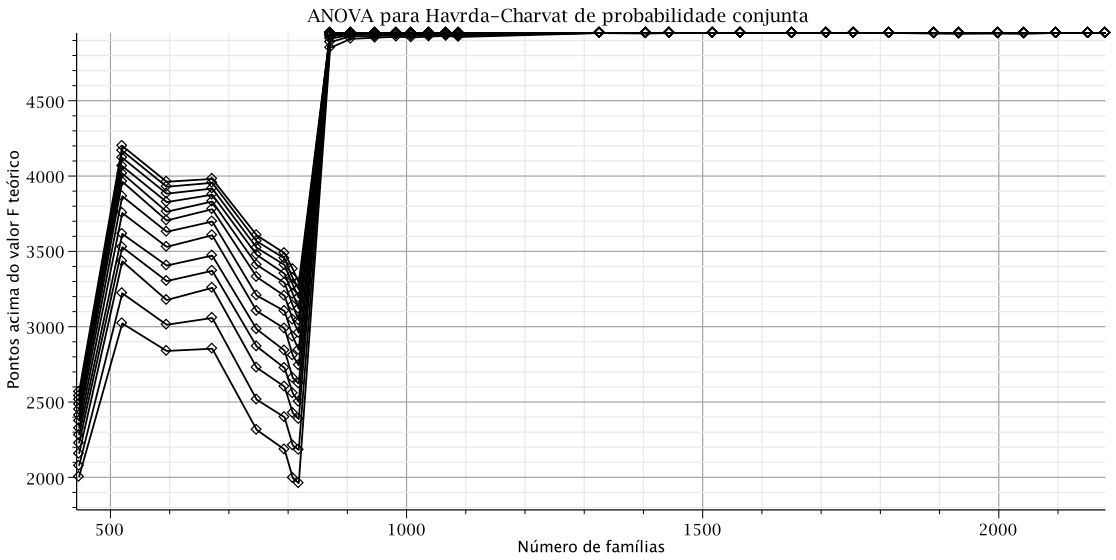


(a)

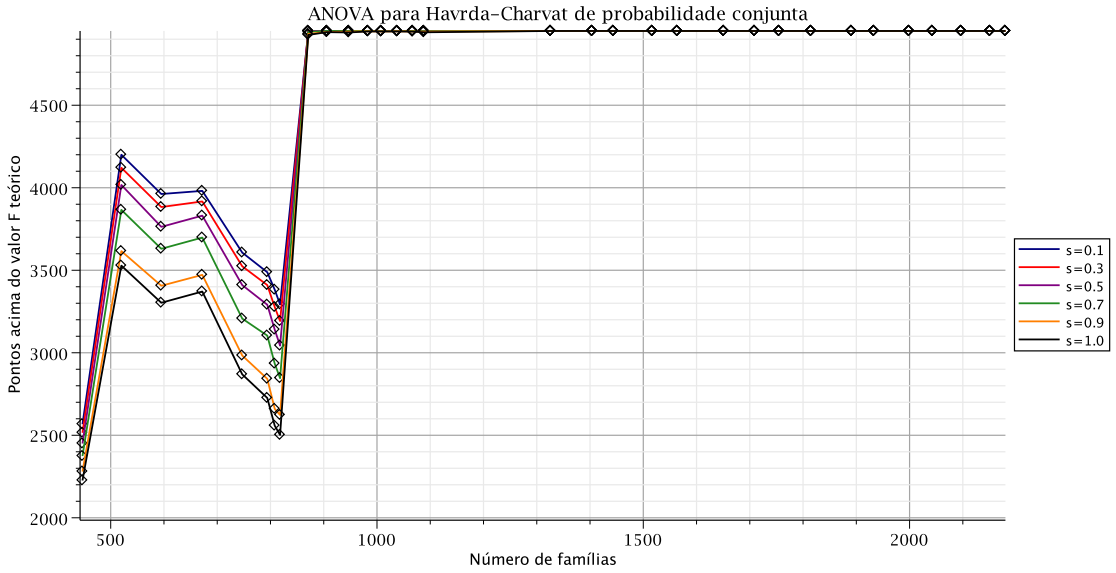


(b)

Figura 7.17: Número de valores de F experimental acima dos valores de F teórico ($F_{jk} > F_{\mu\,\nu\,\alpha}$) para um número cumulativo de famílias (probabilidades $p_{jk}(a, b)$) para a entropia Havrda-Charvat com blocos representativos (100×200). (a) Curvas correspondentes a 13 valores do parâmetro s: 0.1, 0.2, ..., 1.3. (b) Curvas com os valores de parâmetro $s = 0.1$, 0.3, 0.5, 0.7, 0.9 e 1.0.

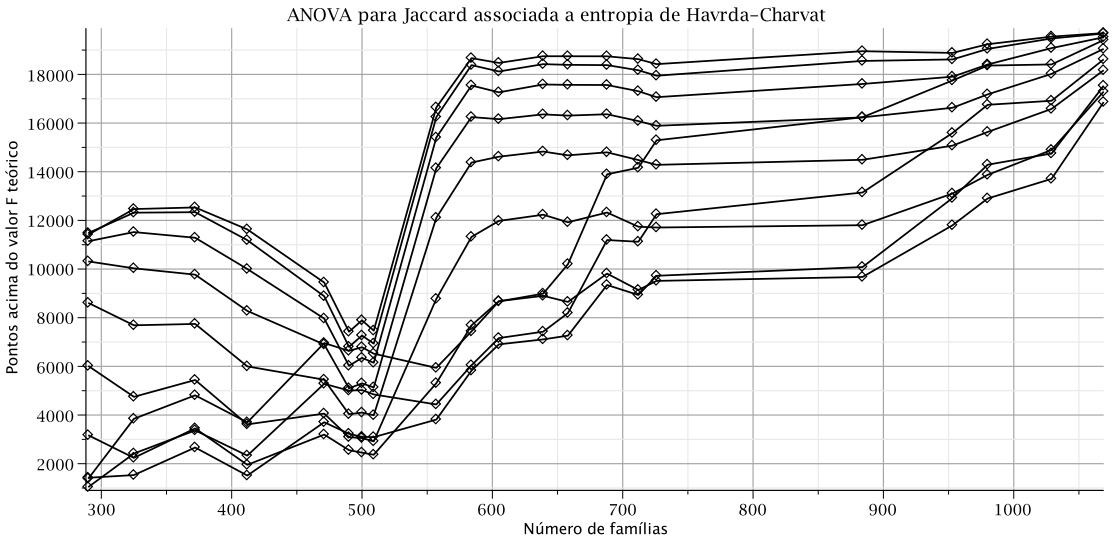


(a)

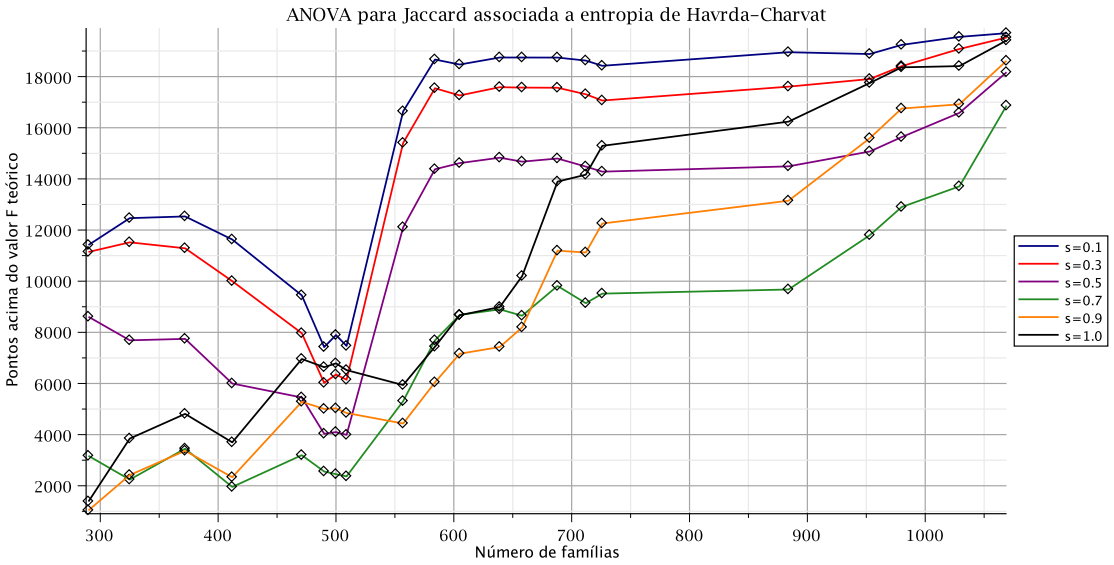


(b)

Figura 7.18: Número de valores de F experimental acima dos valores de F teórico ($F_{jk} > F_{\mu\,\nu\,\alpha}$) para um número cumulativo de famílias (probabilidades $p_{jk}(a, b)$) para a entropia Havrda-Charvat com blocos representativos (100×100). (a) Curvas correspondentes a 13 valores do parâmetro s: 0.1, 0.2, ..., 1.3. (b) Curvas com os valores de parâmetro $s = 0.1$, 0.3, 0.5, 0.7, 0.9 e 1.0.

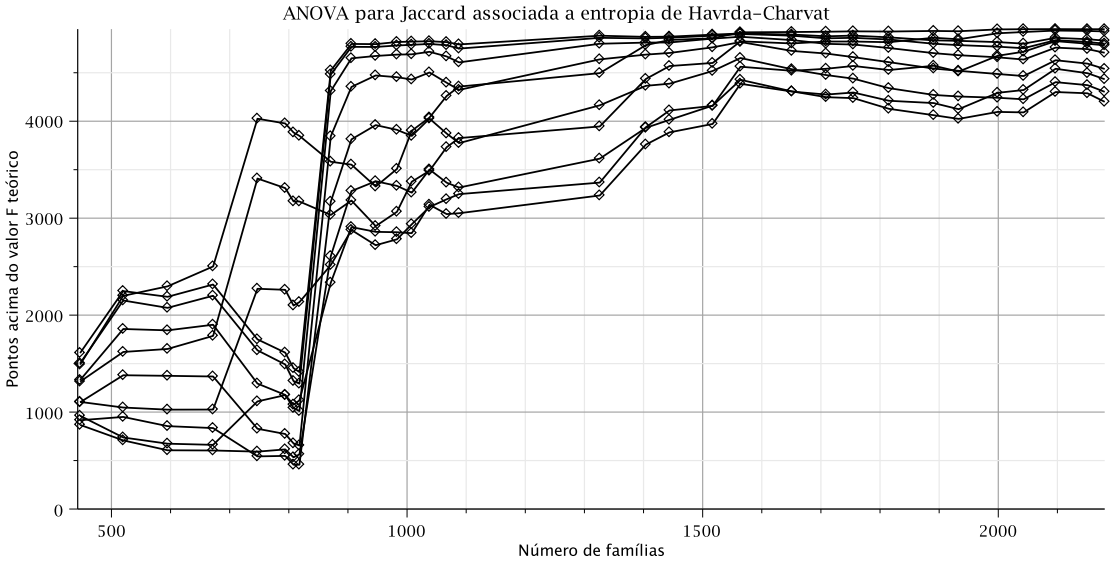


(a)

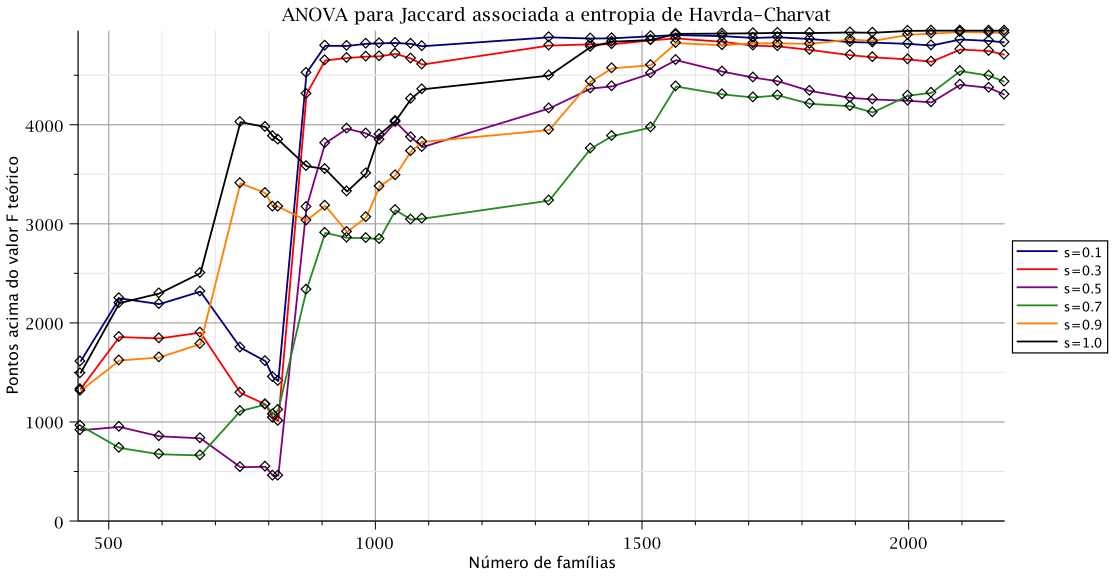


(b)

Figura 7.19: Número de valores de F experimental acima dos valores de F teórico ($F_{jk} > F_{\mu\,\nu\,\alpha}$) para um número cumulativo de famílias para a entropia Jaccard associada a entropia de Havrda-Charvat com blocos representativos (100×200). (a) Curvas correspondentes a 10 valores do parâmetro s: 0.1, 0.2, ..., 1.0. (b) Curvas com os valores de parâmetro $s = 0.1$, 0.3, 0.5, 0.7, 0.9 e 1.0.

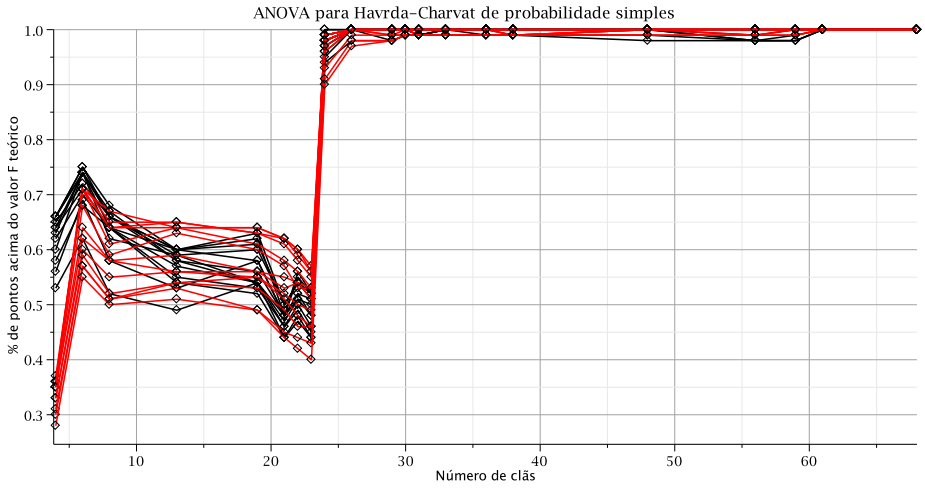


(a)

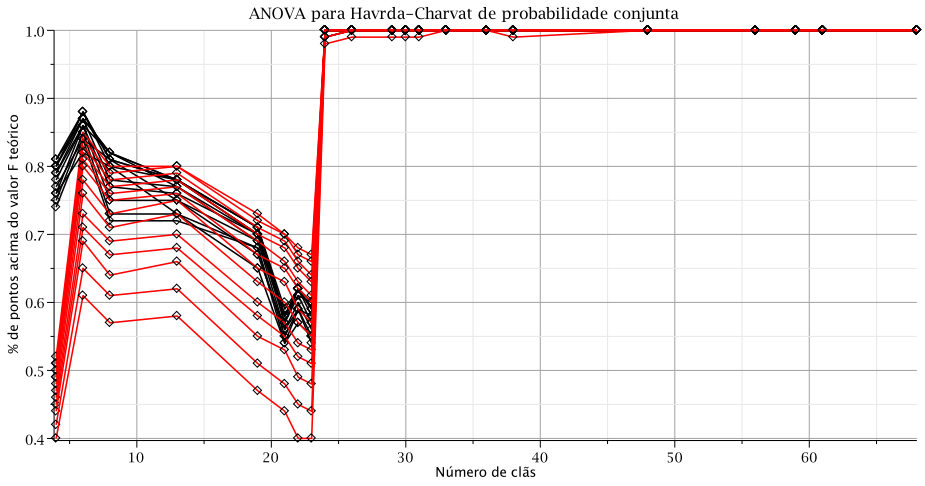


(b)

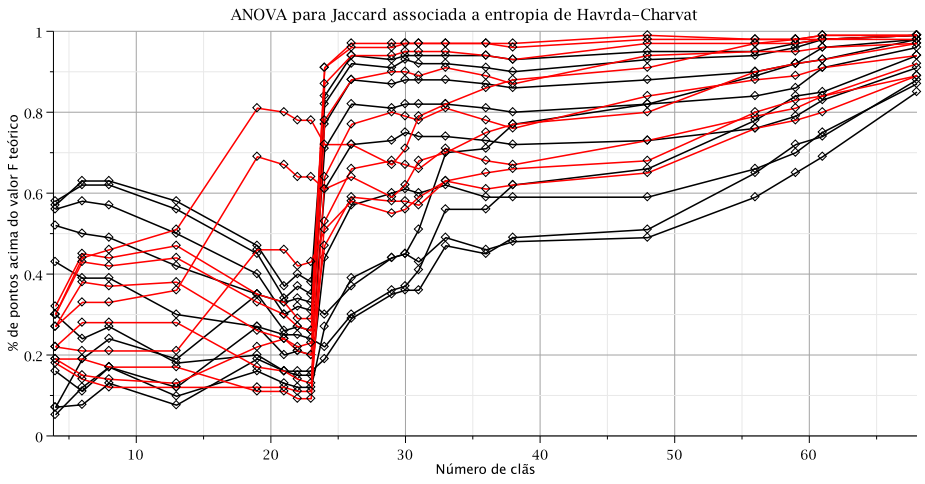
Figura 7.20: Número de valores de F experimental acima dos valores de F teórico ($F_{jk} > F_{\mu\nu\alpha}$) para um número cumulativo de famílias para a entropia Jaccard associada a entropia de Havrda-Charvat com blocos representativos (100×100). (a) Curvas correspondentes a 10 valores do parâmetro s: 0.1, 0.2, ..., 1.0. (b) Curvas com os valores de parâmetro $s = 0.1$, 0.3, 0.5, 0.7, 0.9 e 1.0.



(a)



(b)



(c)

Figura 7.21: Comparação entre os testes ANOVA com blocos representativos (100×200) e (100×100). Em preto temos as curvas correspondentes aos blocos (100×200) e em vermelho temos as curvas dos testes com (100×100). (a) Entropia de Havrda-Charvat de probabilidade simples. (b) Entropia de Havrda-Charvat de probabilidade conjunta. (c) Entropia de Jaccard associada a entropia de Havrda-Charvat.

É interessante notar algumas particularidades obtidas em cada um dos testes. Para a entropia de Havrda-Charvat de probabilidade simples, no caso com os blocos (100×200) os resultados melhoram conforme o valor do parâmetro aumenta até chegar a um máximo em torno de $s = 0.5$ (Figura 7.15b), quando então começa a decrescer. Já para os blocos (100×100), este máximo está por volta de $s = 0.3$ (Figura 7.16b) e passa a decrescer conforme o parâmetro s aumenta. Uma pequena observação: aqui e de agora em diante, quando dizemos "resultados melhores" em relação ao valor do parâmetro, estamos nos referindo a ter mais pontos acima das retas ou planos, ou seja, a um maior número de testes em que houve a rejeição da hipótese nula.

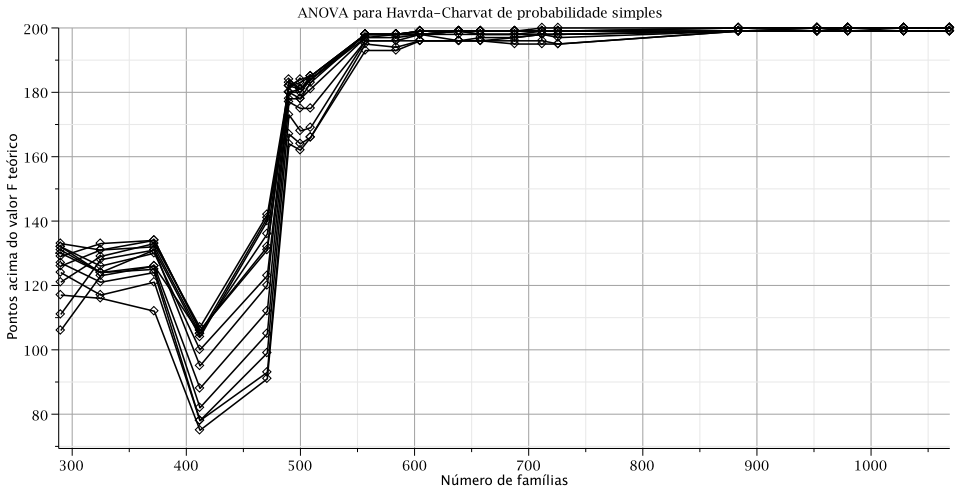
Os testes com probabilidades conjuntas se mostraram bem diferentes. Enquanto que para blocos (100×200) as curvas aumentam conforme o parâmetro s aumenta até $s = 1$ (com o máximo variando entre alguns valores), quando então passam a decrescer (Figura 7.17), para os blocos (100×100) elas decrescem invariavelmente conforme o parâmetro s aumenta (Figura 7.18). Já para a entropia de Jaccard, em ambos os casos os valores começam a decrescer conforme o valor do parâmetro s aumenta até um valor em torno de $s = 0.5$, passando a partir daí a crescer, ao mesmo tempo que assumem um formato diferente.

Uma tendência comum a todos os testes com a entropia de Havrda-Charvat é a ocorrência de uma queda nos resultados na região que contém 21, 22 e 23 clãs, quando então temos uma subida abrupta ao incluirmos o vigésimo quarto clã. Para Jaccard este comportamento é observado até $s = 0.5$, quando então este efeito é amenizado. Para verificar se este efeito é devido a características individuais dos clãs, uma ordenação alternativa foi testada para o caso com blocos representativos (100×200), conforme a Tabela 7.5. A quantidade de clãs e de famílias em cada teste foi mantida igual a ordenação original, sendo trocados clãs que contêm o mesmo número de famílias.

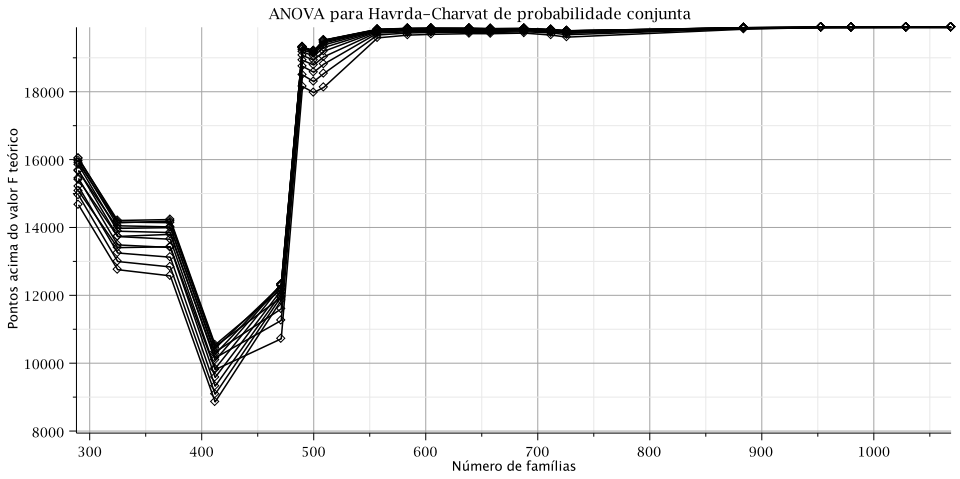
A Figura 7.22 apresenta os resultados dos testes com as entropias de Havrda-Charvat de probabilidade simples e conjunta e a de Jaccard associada para blocos (100×200), utilizando a ordenação da Tabela 7.5. As curvas apresentam uma alteração considerável na região mencionada anteriormente e nos valores com números menores de clãs. Porém, novamente ao incluirmos o vigésimo quarto clã, os resultados sobem em direção ao máximo. Estes resultados indicam que dentro deste grupo de clãs utilizados já temos pelo menos um que é significativamente diferente dos outros. Futuramente outros testes serão realizados alterando a ordem dos clãs sem respeitar o número de famílias para verificar se podemos deslocar este resultado para a direita ou para a esquerda.

Tabela 7.5: Ordem alternativa de inclusão dos clãs nos testes com blocos representativos 100×200.

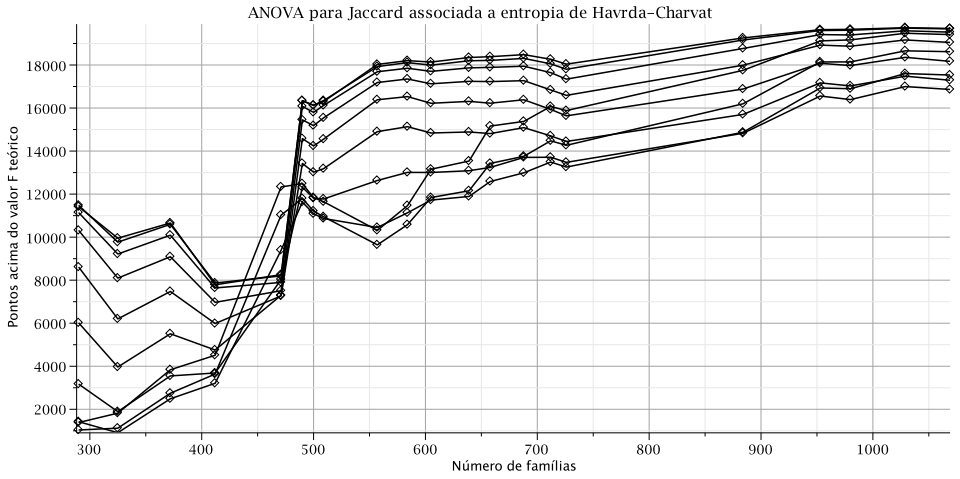
nº de clãs	Clãs adicionados
4	CL0020, CL0023, CL0028, CL0063
6	CL0013, CL0110
8	CL0126, CL0015
13	CL0128, CL0373, CL0016, CL0181, CL0103
19	CL0149, CL0292, CL0286, CL0236, CL0034, CL0163
21	CL0029, CL0254
22	CL0260
23	CL0151
24	CL0036
26	CL0113, CL0184
29	CL0105, CL0004, CL0295
30	CL0058
31	CL0059
33	CL0061, CL0111
36	CL0137, CL0027, CL0123
38	CL0052, CL0316
48	CL0108, CL0186, CL0219, CL0039, CL0270, CL0125, CL0192, CL0264, CL0144, CL0179
56	CL0158, CL0246, CL0040, CL0035, CL0142, CL0046, CL0177, CL0064
59	CL0014, CL0182, CL0088
61	CL0193, CL0062
68	CL0030, CL0257, CL0044, CL0268, CL0093, CL0118, CL0127



(a)



(b)



(c)

Figura 7.22: Testes ANOVA com uma ordenação alternativa para blocos representativos (100×200). (a) Entropia de Havrda-Charvat de probabilidade simples. (b) Entropia de Havrda-Charvat de probabilidade conjunta. (c) Entropia de Jaccard associada a entropia de Havrda-Charvat.

Como um último teste para verificar se não estamos confiando cegamente na robustez do teste ANOVA, introduzimos o conceito de "pseudo-clãs" para realizarmos comparações com os clãs utilizados previamente. Os pseudo-clãs nada mais são do que clãs fictícios, construídos a partir das famílias dos clãs originais. Os testes realizados foram apenas para blocos representativos (100×100). Foram feitas duas formulações diferentes para a construção dos pseudo-clãs, mostradas a seguir:

1. **Troca de famílias entre pares de clãs.** A proposta aqui apresentada consiste em separar os clãs em pares, sendo que um deles sempre tem um número de famílias maior do que o outro. Em seguida, são criados dois pseudo-clãs: um com o mesmo número de famílias do menor clã e outro com o mesmo número de famílias do maior. As famílias que constituem o menor pseudo-clã são provenientes do maior clã, enquanto o maior pseudo-clã recebe as famílias remanescentes do maior clã e mais todas as famílias do menor clã.

$$\underline{68 \text{ clãs}}$$

$$\begin{array}{cc} \textbf{CL1} & \textbf{CL2} \\ \mathrm{F}_1 & \mathrm{F}_2 \end{array}$$

$$\mathrm{F}_1 < \mathrm{F}_2$$

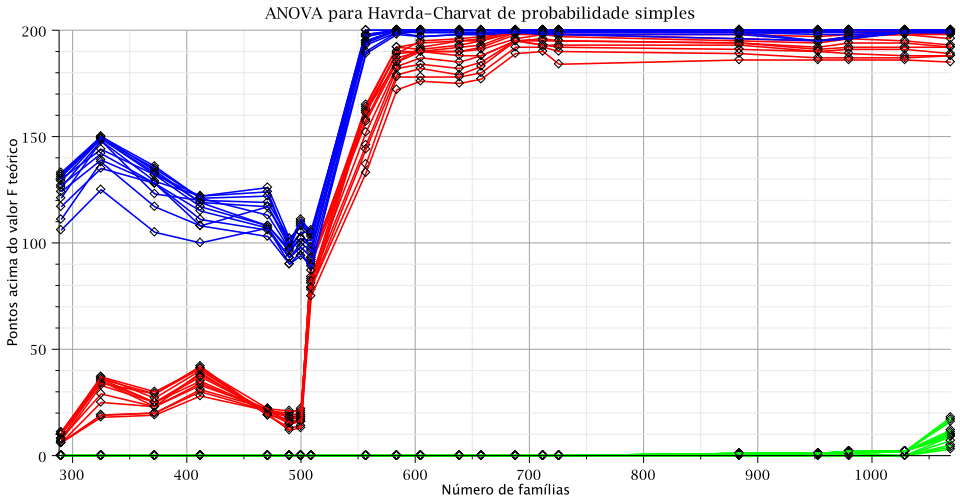
$$\begin{array}{cc} \textbf{PCL1} & \textbf{PCL2} \\ (\mathrm{F}_1 \in \mathrm{CL2}) & \begin{pmatrix} (\mathrm{F}_2 - \mathrm{F}_1) \in \mathrm{CL2} \\ + \\ \mathrm{F}_1 \in \mathrm{CL1} \end{pmatrix} \end{array}$$

2. **Sorteio aleatório de famílias sem reposição.** São criados pseudo-clãs com o mesmo número de famílias que os clãs originais. As famílias dos clãs são agrupadas e em seguida distribuidas aleatoriamente entre os pseudo-clãs.

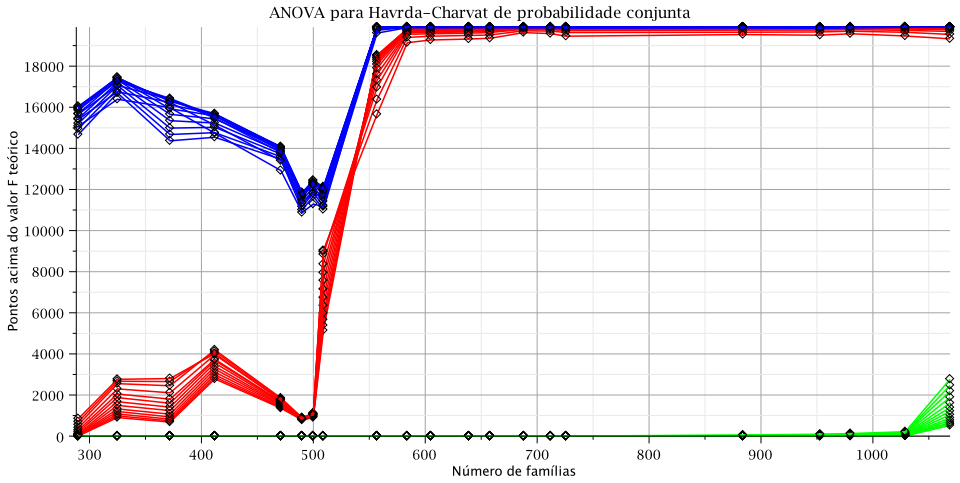
$$\bigcup_{j=1}^{1069} \mathrm{PF}_j$$

$$\begin{array}{rl} \mathrm{F}_1 & \longrightarrow \mathrm{PCL1} \\ \mathrm{F}_2 & \longrightarrow \mathrm{PCL2} \\ \vdots & \quad \vdots \\ \mathrm{F}_{68} & \longrightarrow \mathrm{PCL68} \end{array}$$

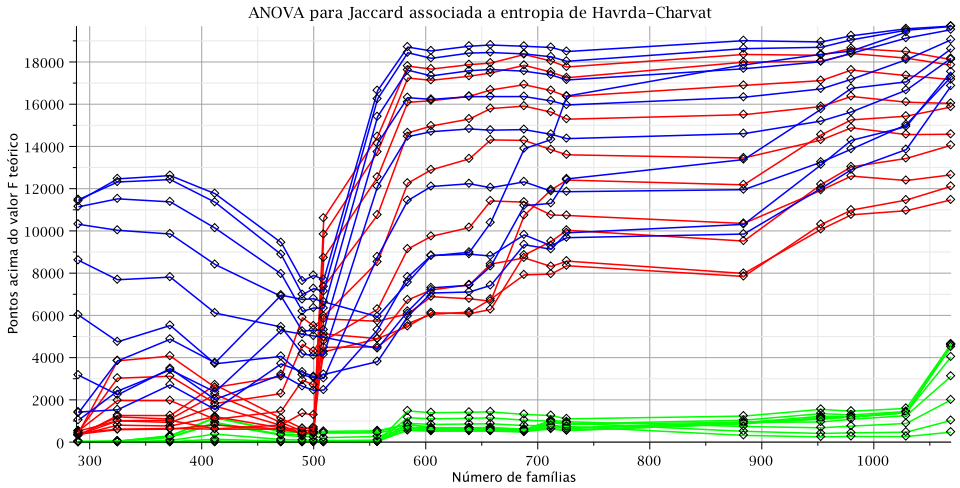
A Figura 7.23 a seguir apresenta os gráficos de contagem dos pontos acima dos valores de F teóricos para os clãs e para os pseudo-clãs.



(a)



(b)



(c)

Figura 7.23: Número de valores de F experimental acima dos valores de F teórico para um número cumulativo de famílias para clãs e pseudo-clãs. Clãs são apresentados em azul e os pseudo-clãs em vermelho (primeira formulação) e em verde (segunda formulação). (a) Entropia Havrda-Charvat de probabilidade simples, (b) entropia Havrda-Charvat de probabilidade conjunta e (c) entropia Jaccard associada a Havrda-Charvat.

Ao inspecionarmos os gráficos, fica claro que os testes para os clãs originais apresentam resultados melhores que os obtidos com os pseudo-clãs, principalmente para as entropias de probabilidade simples e conjunta de Havrda-Charvat, reforçando a rejeição da Hipótese Nula de não existência de clãs.

Capítulo 8

Conclusão

Os resultados obtidos com os blocos (100×200) e (100×100) representativos das famílias indicam que não podemos negar a existência de famílias agrupadas em clãs. A criação de pseudo-clãs e a realização dos testes de Fisher para fins de comparação com os testes originais, reforçam a rejeição da Hipótese Nula de não existência dos clãs, dando um aporte substancial ao resultado com os clãs.

A partir de um determinado número de famílias, a rejeição da Hipótese Nula aumenta consideravelmente. Isso pode ser um indicador da saturação do teste. Os dados utilizados para os testes devem ser tratados mais adequadamente, de forma a termos distribuições mais homogêneas. As possíveis alterações a serem feitas são:

- Aumentar o limite inferior de famílias com blocos (100×200). O limite atual de 5 famílias pode não ser suficiente para representar corretamente o clã.
- Complementando a alteração anterior, podemos também limitar o número máximo de famílias, de forma que todos os clãs utilizados sejam representados por um número igual de famílias.

Para a realização dos testes, foi confiado na robustez da estatística ANOVA, mas esta pode não ser suficiente para lidar com distribuições que não estejam muito próximas da distribuição normal e com diferentes variâncias. Se os dados são balanceados (número de elementos iguais em cada grupo), o teste F da ANOVA é mais robusto em relação a desigualdade das variâncias. Através de transformações não lineares os dados podem ser tratados para que os resultados do teste sejam mais confiáveis. Outras estatísticas mais robustas, como a de Brown-Forsythe [46], podem ser mais convenientes para aperfeiçoar os resultados obtidos.

Os valores do parâmetro s das medidas de entropia utilizadas influencia diretamente os resultados dos testes, principalmente para os testes feitos com até 23 clãs. Novos testes devem ser realizados para termos um melhor esclarecimento sobre estes

efeitos, de forma que possamos futuramente determinar qual a melhor medida de entropia e o melhor valor de parâmetro para descrever o banco de dados.

Além das sugestões apresentadas, os trabalhos futuros a serem realizados são:

- Fazer os testes para as outras medidas de entropia (Renyi, Landsberg-Vedral e Sharma-Mittal) e para a Jaccard associada a elas;
- Comparar com resultados obtidos com versões anteriores e posteriores à versão 27.0 do PFAM;
- Trabalhar com outros blocos representativos.

Referências Bibliográficas

[1] KIM, C., XIAO, X., CHEN, S., et al. "Artificial strain of human prions created in vitro", *Nature Communications*, v. 9, june (2018). Article number: 2166.

[2] MONDAINI, R. P., DE ALBUQUERQUE NETO, S. C. "The Protein Family Classification in Protein Databases via Entropy Measures". ArXiv: 1806.05172 [q-bio.BM], (2018).

[3] MONDAINI, R. P., DE ALBUQUERQUE NETO, S. C. "Optimal Control of a Coarse-Grained Model for Protein Dynamics". In: *BIOMAT 2014 Int. Symp.*, pp. 12–25. World Scientific Co. Pte. Ltd., (2015).

[4] MONDAINI, R. P., DE ALBUQUERQUE NETO, S. C. "The Pattern Recognition of Probability Distributions of Amino Acids in Protein Families". In: *Mathematical Biology and Biological Physics — BIOMAT 2016*, pp. 29–50. World Scientific Co. Pte. Ltd., (2017).

[5] MONDAINI, R. P. "A Survey of Geometric Techniques for Pattern Recognition of Probability of Occurrence of Amino Acids in Protein Families". In: *Mathematical Biology and Biological Physics — BIOMAT 2016*, pp. 304–326. World Scientific Co. Pte. Ltd., (2017).

[6] MONDAINI, R. P., DE ALBUQUERQUE NETO, S. C. "Pattern recognition of amino acids via a Poisson statistical approach". In: *Physical and Mathematical Aspects of Symmetries: Proceedings of the 31st International Colloquium in Group Theoretical Methods in Physics*, pp. 263–270. Springer International Publ., (2018).

[7] MONDAINI, R. P., DE ALBUQUERQUE NETO, S. C. "Stochastic Assessment of Protein Databases by Generalized Entropy Measures". In: *Trends in Biomathematics: Modeling, Optimization and Computational Methods — BIOMAT 2017*, pp. 103–119. Springer International Publ., (2018).

[8] ALBERTS, B., JOHNSON, A., LEWIS, J., et al. *Molecular Biology of The Cell*. Garland Science, (2008).

[9] NEUMAIER, A. "Molecular Modelling of Proteins and Mathematical Prediction of Protein Structures", *SIAM Review*, v. 39, pp. 407–460, (1997).

[10] HUANG, K. *Lectures on Statistical Physics and Protein Folding*. World Scientific, (2005).

[11] FERSHT, A. *Structure and Mechanism in Protein Folding*. W. H. Freeman and Company, (1999).

[12] VOET, D., VOET, J. G., PRATT, C. W. *Fundamentals of Biochemistry — Life at the Molecular Level*. John Wiley & Sons, Inc., (2013).

[13] EISENBERG, D., MARCOTTE, E. M., XENARIOS, I., et al. "Protein Function in the Post-Genomic Era", *Nature*, v. 1, n. 405, pp. 823–826, (2000).

[14] "Universal Protein Resource (UniProt)". Disponível em: `http://www.uniprot.org/`. Acesso em: 02/02/2018.

[15] "Protein Research Foundation, Japan (PRF)". Disponível em: `https://www.prf.or.jp/index-e.html`. Acesso em: 02/02/2018.

[16] "Protein Family Database (PFAM)". Disponível em: `http://pfam.xfam.org/`. Acesso em: 02/02/2018.

[17] PUNTA, M., COGGILL, P., EBERHARDT, R. Y., et al. "The Pfam Protein Families Database", *Nucleic Acids Research*, v. 40, n. D1, pp. D290–D301, (2012).

[18] FINN, R. D., MISTRY, J., SCHUSTER-BÖCKLER, B., et al. "Pfam: Clans, Web Tools and Services", *Nucleic Acids Research*, v. 34, n. D1, pp. D247–D251, (2006).

[19] FINN, R. D., TATE, J., MISTRY, J., et al. "The Pfam Protein Families Database", *Nucleic Acids Research*, v. 9, n. 3, pp. 1–8, (2007).

[20] 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).

[21] SAMMUT, S. J., FINN, R. D., BATEMAN, A. "Pfam 10 Years on: 10000 Families and Still Growing", *Briefings in Bioinformatics*, v. 9, n. 3, pp. 210–219, (2008).

[22] HAUSSLER, D., KROGH, A. "Protein Alignment and Clustering". (1992). Presented at the conference Neural Networks for Computing.

[23] HAUSSLER, D., KROGH, A., MIAN, I. S., et al. *Protein Modeling Using Hidden Markov Models: Analysis of Globins*. Technical Report UCSC-CRL-92-23, University of California at Santa Cruz, Computer Science Dept., Santa Cruz, CA 95064, (1992).

[24] SONNHAMMER, E. L. L., EDDY, S. R., DURBIN, R. "Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments", *PROTEINS: Structure, Function, and Genetics*, v. 28, n. 3, pp. 405–420, (1997).

[25] KOLMOGOROV, A. N. *Foundations of the Theory of Probability*. New York, Chelsea, (1950).

[26] BAYES, T. "An Essay Towards Solving a Problem in the Doctrine of Chances", *Philosophical Transactions of the Royal Society of London*, v. 1, n. 53, pp. 370–418, (1763).

[27] SHARMA, B. D., MITTAL, D. P. "New Non-Additive Measures of Entropy for Discrete Probability Distributions", *J. Math Sci*, v. 10, pp. 28–40, (1972).

[28] HAVRDA, J., CHARVAT, F. "Quantification Method of Classification Processes. Concept of Structural a-Entropy", *Kybernetika*, v. 3, n. 1, pp. 30–35, (1967).

[29] LANDSBERG, P. T., VEDRAL, V. "Distributions and Channel Capacities in Generalized Statistical Mechanics", *Phys. Lett. A*, v. 224, pp. 326–330, (1997).

[30] RÉNYI, A. "On Measures of Entropy and Information". In: *Proc. Fourth Berkely Symp. Math. Statist. and Probability*, v. 1, pp. 547–561. University of California Press Berkely, (1961).

[31] SHANNON, C. E. "The Mathematical Theory of Communication", *Bell Syst, Tech. Journ.*, v. 27, pp. 379–423; 623–656, (1948).

[32] KHINCHIN, A. I. *Mathematical Foundations of Information Theory*. Dover Publications, Inc., (1957).

[33] FEINSTEIN, A. *Foundations of Information Theory*. New York, McGraw-Hill, (1958).

[34] KULLBACK, S., LEIBLER, R. A. "On Information and Sufficiency", *Annals of Mathematical Statistics*, v. 22, n. 1, (1951).

[35] KULLBACK, S. *Information Theory and Statistics*. Dover Publications, Inc., (1968).

[36] 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 Int. Symp.*, pp. 209–218. World Scientific Co. Pte. Ltd., (2014).

[37] CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L., et al. *Introduction to Algorithms*. second ed. Cambridge, Massachusetts, The MIT Press, (2001).

[38] FOURMENT, M., GILLINGS, M. R. "A Comparison of Common Programming Languages used in Bioinformatics", *BMC Bioinformatics*, v. 9, pp. 82, (2008).

[39] MONDAINI, R. P., DE ALBUQUERQUE NETO, S. C. "Entropy Measures and the Statistical Analysis of Protein Family Classification". In: *BIOMAT 2015 Int. Symp.*, pp. 193–210. World Scientific Co. Pte. Ltd., (2016).

[40] FELLER, W. *An Introduction to Probability Theory and Its Applications*, v. 2. John Wiley & Sons, Inc., (1971).

[41] DEGROOT, M. H., SCHERVISH, M. J. *Probability and Statistics*. Addison-Wesley, (2012).

[42] ABRAMOWITZ, M., STEGUN, I. A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. 9th printing ed. New York, Dover, (1965).

[43] FISHER, R. A. *Statistical Methods for Research Workers*. twelfth ed. New York, Hafner Publishing Company Inc., (1954).

[44] TABOGA, M. *Lectures on Probability Theory and Mathematical Statistics*. CreateSpace Independent Publishing Platform, (2012).

[45] MONDAINI, R. P., DE ALBUQUERQUE NETO, S. C. "The ANOVA Statistics of Protein Databases via Entropy Measures". (2017). Presented at BIOMAT 2017 International Symposium on Mathematical and Computational Biology, Oct. 30 – Nov. 03, 2017 by S. C. de Albuquerque Neto — Institute of Numerical Mathematics, Russian Academy of Sciences, Russia.

[46] BROWN, M. B., FORSYTHE, A. B. "Robust Tests for the Equality of Variances", *Journal of the American Statistical Association*, v. 69, n. 346, pp. 364–367, (1974).