

17th International Symposium on Mathematical and Computational Biology

Oct. 30 - Nov. 03 2017

Moscow, Institute of Numerical Mathematics, Russia

THE ANOVA STATISTICS OF PROTEIN DATABASES VIA ENTROPY MEASURES

R. P. Mondaini

S. C. de Albuquerque Neto

Federal University of Rio de Janeiro, RJ, Brazil

Centre of Technology, COPPE

OCTOBER 30, 2017

Formation and evolution of Protein Families

Statistical Analysis of amino acids distribution.

The Sample space is organized by selecting blocks of amino acids of m rows (protein domains) and n columns (amino acids) as obtained from a protein database.

[illegible]

Figure: $(m \times n)$ block of amino acids — a representation of a protein family — an element of the Sample space.

In order to organize a block, we consider rows with n_l amino acids, $n_l = n_1, n_2, \dots, n_m$. All domains such that $n_l < n$ are deleted as well as $(n_l - n)$ amino acids on all other domains.

Pfam Database

Biological Almanac instead of Astronomical Almanac (Ephemerides).

Table: Pfam Database Evolution

Pfam DATABASE				
version	year	n ^o of families	n ^o of families class. into clans	"Clans"
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

The work with version 27.0 allows for comparison with data of previous versions and the continuous prevision of data for future versions.

Pfam Database — Version 27.0

N° of families: 14831

Adopted restrictions for One-way ANOVA Statistical Analysis:

Restrictions	n° of families class. into clans	n° of Clans
none	4563	515
100x200 blocks, one block per family	1441	267
Clans with 5 or more families	1069	68

Probability Distribution

We associate a vector p_j to each column of m rows:

$$p_j = \begin{pmatrix} p_j(A) \\ \vdots \\ p_j(Y) \end{pmatrix}, \text{ } n \text{ vectors of 20 components}$$

$$p_j(a) = \frac{n_j(a)}{m}$$

$n_j(a)$ is the number of occurrences of amino acid “ a ” in the j^{th} column.

$a = \text{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}$

$$\sum_a p_j(a) = 1, \text{ } j = 1, 2, \dots, n$$

Probability Distribution

We now consider the joint probability $p_{jk}(a, b)$ of finding the amino acid a in column j and the amino acid b in column k :

$$p_{jk}(a, b) = \frac{n_{jk}(a, b)}{m}$$

$$\sum_a \sum_b p_{jk}(a, b) = 1, \quad \begin{matrix} j = 1, 2, \dots, (n-1) \\ k = (j+1), (j+2), \dots, n \end{matrix}$$

For a block $(m \times n)$ we have:

$$p_{jk} = \begin{pmatrix} p_{jk}(A, A) & \dots & p_{jk}(A, Y) \\ \vdots & \ddots & \vdots \\ p_{jk}(Y, A) & \dots & p_{jk}(Y, Y) \end{pmatrix}, \quad \frac{n(n-1)}{2} \quad \begin{matrix} \text{square matrices of 400} \\ \text{elements each} \end{matrix}$$

Sharma-Mittal Set of Entropy Measures

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

$$\xrightarrow{r \rightarrow s} H_j(s) = -\frac{1}{1-s} \left(1 - \sum_a (p_j(a))^s \right) \quad \text{Havrda-Charvat Entropy}$$

$$\xrightarrow{s \rightarrow 1} S_j = -\sum_a p_j(a) \log p_j(a) \quad \text{Shannon Entropy}$$

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

$$\xrightarrow{r \rightarrow s} H_{jk}(s) = -\frac{1}{1-s} \left(1 - \sum_a \sum_b (p_{jk}(a, b))^s \right) \quad \text{Havrda-Charvat Entropy}$$

$$\xrightarrow{s \rightarrow 1} S_{jk} = -\sum_a \sum_b p_{jk}(a, b) \log p_{jk}(a, b) \quad \text{Shannon Entropy}$$

Mutual Information

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

$$\xrightarrow{r \rightarrow s} M_{jk}(s) = \frac{1}{1-s} \left(1 - \frac{\sum_a \sum_b (p_{jk}(a, b))^s}{\sum_a \sum_b (p_j(a)p_k(b))^s} \right)$$

$$\xrightarrow{s \rightarrow 1} M_{jk}(1) = S_j + S_k - S_{jk}$$

$$M_{jk}(r, s) \geq 0, \quad (SM)_{jk}(r, s) - M_{jk}(r, s) \geq 0$$

$$0 \leq (SM)_{jk}(r, s) - M_{jk}(r, s) \leq (SM)_{jk}(r, s)$$

$$0 \leq 1 - \frac{M_{jk}(r, s)}{(SM)_{jk}(r, s)} \leq 1$$

Jaccard Entropy Measure

$$J_{jk}(r, s) = 1 - \frac{M_{jk}(r, s)}{(SM)_{jk}(r, s)}$$
$$\xrightarrow{r \rightarrow s} J_{jk}(s) = 1 - \frac{M_{jk}(s)}{H_{jk}(s)}$$

The corresponding mean Jaccard measure is given by:

$$J(r, s) = \frac{2}{n(n-1)} \sum_j \sum_k J_{jk}(r, s)$$

The mean Sharma-Mittal for simple and joint probability are, respectively:

$$(SM)(r, s) = \frac{1}{n} \sum_j (SM)_j(r, s)$$
$$(SM)(r, s) = \frac{2}{n(n-1)} \sum_j \sum_k (SM)_{jk}(r, s)$$

Mean Jaccard X Mean Havrda-Charvat (Joint Probability)

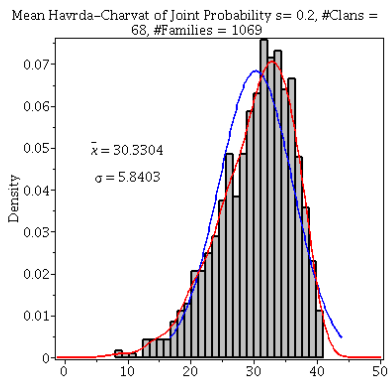
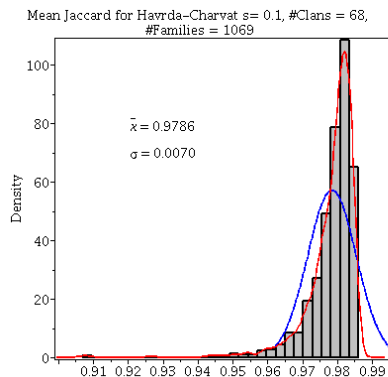


Figure: Histograms of Mean Jaccard (left side) and Mean Havrda-Charvat (right side) of 1069 families with $s = 0.1$.

Mean Jaccard X Mean Havrda-Charvat (Joint Probability)

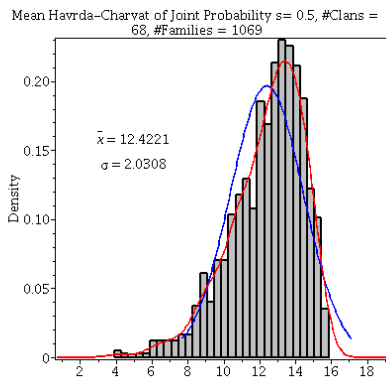
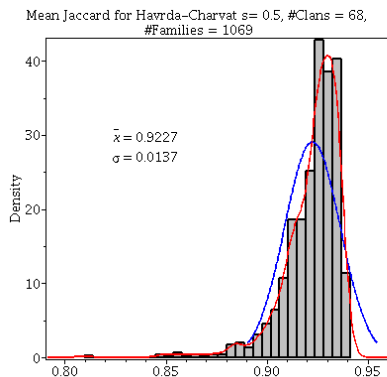


Figure: Histograms of Mean Jaccard (left side) and Mean Havrda-Charvat (right side) of 1069 families with $s = 0.5$.

Mean Jaccard X Mean Havrda-Charvat (Joint Probability)

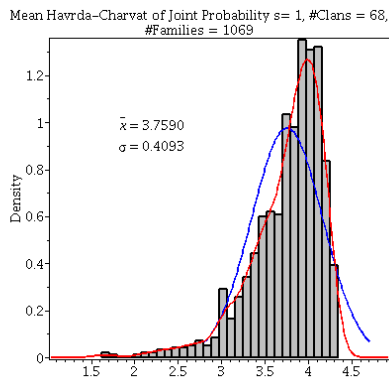
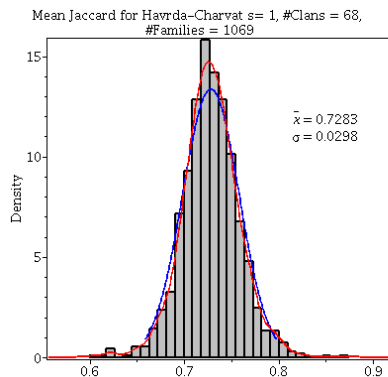


Figure: Histograms of Mean Jaccard (left side) and Mean Havrda-Charvat (right side) of 1069 families with $s = 1.0$.

Mean Jaccard X Mean Havrda-Charvat (Joint Probability)

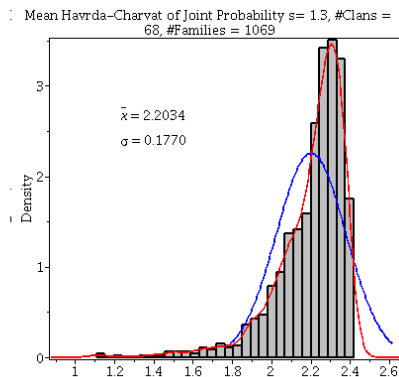
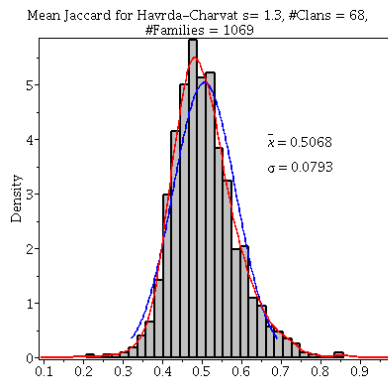
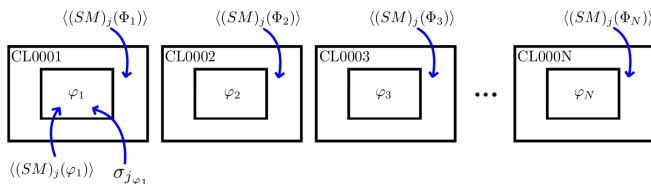


Figure: Histograms of Mean Jaccard (left side) and Mean Havrda-Charvat (right side) of 1069 families with $s = 1.3$.

F-test: $F \rightarrow$ Fisher - ANOVA

Groups of protein families (“Clans”) CL0001, CL0002, \dots , CL000N with $\Phi_1, \Phi_2, \dots, \Phi_N$ protein families, respectively and $\varphi_1, \varphi_2, \dots, \varphi_N$ the number of protein families on each statistical sample after the restriction to families containing $m \times n$ blocks of amino acids, respectively.



$\langle(SM)_j(\Phi_1)\rangle, \langle(SM)_j(\Phi_2)\rangle, \dots, \langle(SM)_j(\Phi_N)\rangle$ — generic means around the “Clans”.

$\langle(SM)_j(\varphi_1)\rangle, \langle(SM)_j(\varphi_2)\rangle, \dots, \langle(SM)_j(\varphi_N)\rangle$ — means per columns of the $(m \times n)$ blocks of amino acids.

F-test: $F \rightarrow$ Fisher - ANOVA

The entropy measure variables will be $(SM)_j^p(\varphi_l)$, where

$j = 1, 2, \dots, n$ (columns), $p = 1, 2, \dots, \varphi_l$ (families of the l^{th} "Clan"),
 $l = 1, 2, \dots, N$ ("Clans").

$$\langle (SM)_j \rangle = \frac{1}{\sum_{l=1}^N \varphi_l} \sum_{l=1}^N \sum_{p=1}^{\varphi_l} (SM)_j^p(\varphi_l) \text{ — overall mean per column of } (m \times n) \text{ blocks of amino acids}$$

$$\langle (SM)_j(\varphi_l) \rangle = \frac{1}{\varphi_l} \sum_{p=1}^{\varphi_l} (SM)_j^p(\varphi_l) \text{ — "Clan" mean per column } j \text{ of the } (m \times n) \text{ blocks of amino acids}$$

The standard deviations can be obtained from:

$$\left(\sum_{l=1}^N \varphi_l - 1 \right) \sigma_j^2 = \sum_{l=1}^N \sum_{p=1}^{\varphi_l} \left((SM)_j^p(\varphi_l) - \langle (SM)_j(\varphi_l) \rangle \right)^2$$

$$(\varphi_l - 1) \sigma_{j_{\varphi_l}}^2 = \sum_{p=1}^{\varphi_l} \left((SM)_j^p(\varphi_l) - \langle (SM)_j(\varphi_l) \rangle \right)^2$$

F-test: $F \rightarrow$ Fisher - ANOVA

We then have:

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

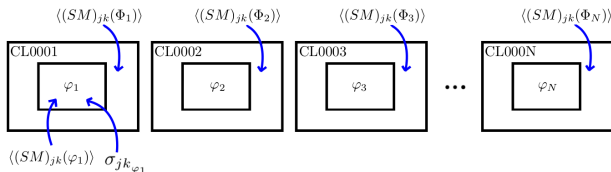
<p>Sum of squares total</p> <p>measures — variation of the</p> <p>data $(SM)_j^p(\varphi_l)$ around the</p> <p>overall mean $\langle (SM)_j \rangle$</p>	<p>Variability within group</p> <p>mean — variation of the</p> <p>data $(SM)_j^p(\varphi_l)$ around its</p> <p>group mean $\langle (SM)_j(\varphi_l) \rangle$</p>	<p>Variability between group means</p> <p>— variation of the group</p> <p>means $\langle (SM)_j(\varphi_l) \rangle$ around the</p> <p>overall mean $\langle (SM)_j \rangle$</p>
--	---	---

To check n^o of independent terms:

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

F-test: $F \rightarrow$ Fisher - ANOVA

Groups of protein families (“Clans”) $CL0001, CL0002, \dots, CL000N$ with $\Phi_1, \Phi_2, \dots, \Phi_N$ protein families, respectively and $\varphi_1, \varphi_2, \dots, \varphi_N$ the number of protein families on each statistical sample after the restriction to families containing $m \times n$ blocks of amino acids, respectively.



$\langle(SM)_{jk}(\Phi_1)\rangle, \langle(SM)_{jk}(\Phi_2)\rangle, \dots, \langle(SM)_{jk}(\Phi_N)\rangle$ — generic means around the “Clans”.

$\langle(SM)_{jk}(\varphi_1)\rangle, \langle(SM)_{jk}(\varphi_2)\rangle, \dots, \langle(SM)_{jk}(\varphi_N)\rangle$ — means by a pair of columns jk of the $(m \times n)$ blocks of amino acids.

F-test: $F \rightarrow$ Fisher - ANOVA

The entropy measure variables will be $(SM)_{jk}^p(\varphi_l)$, where $j = 1, 2, \dots, n-1$, $k = (j+1), (j+2), \dots, n$, $p = 1, 2, \dots, \varphi_l$ (families of the l^{th} "Clan"), $l = 1, 2, \dots, N$ ("Clans").

$$\langle (SM)_{jk} \rangle = \frac{1}{\sum_{l=1}^N \varphi_l} \sum_{l=1}^N \sum_{p=1}^{\varphi_l} (SM)_{jk}^p(\varphi_l) \text{ — overall mean per column of } (m \times n) \text{ blocks of amino acids}$$

$$\langle (SM)_{jk}(\varphi_l) \rangle = \frac{1}{\varphi_l} \sum_{p=1}^{\varphi_l} (SM)_{jk}^p(\varphi_l) \text{ — "Clan" mean per column } j \text{ of the } (m \times n) \text{ blocks of amino acids}$$

The standard deviations can be obtained from:

$$\left(\sum_{l=1}^N \varphi_l - 1 \right) \sigma_{jk}^2 = \sum_{l=1}^N \sum_{p=1}^{\varphi_l} \left((SM)_{jk}^p(\varphi_l) - \langle (SM)_{jk}(\varphi_l) \rangle \right)^2$$

$$(\varphi_l - 1) \sigma_{jk_{\varphi_l}}^2 = \sum_{p=1}^{\varphi_l} \left((SM)_{jk}^p(\varphi_l) - \langle (SM)_{jk}(\varphi_l) \rangle \right)^2$$

F-test: $F \rightarrow$ Fisher - ANOVA

We then have:

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

Sum of squares total Variability within group Variability between group means

measures — variation of the mean — variation of the — variation of the group

data $(SM)_{jk}^p(\varphi_l)$ around the data $(SM)_{jk}^p(\varphi_l)$ around its means $\langle (SM)_{jk}(\varphi_l) \rangle$ around the

overall mean $\langle (SM)_{jk} \rangle$ group mean $\langle (SM)_{jk}(\varphi_l) \rangle$ overall mean $\langle (SM)_{jk} \rangle$

To check n^o of independent terms:

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

F-test: $F \rightarrow$ Fisher - ANOVA

$$\underline{p_j(a), H_j(s), J_j(s)}$$

n ANOVA Tests on the $(m \times n)$ block samples

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

$$\underline{p_{jk}(a,b), H_{jk}(s), J_{jk}(s)}$$

$\frac{n(n-1)}{2}$ ANOVA Tests on the $(m \times n)$ block samples

$$F_{jk} = \frac{\frac{SSG}{N-1}}{\frac{SSE}{\sum_{l=1}^N \varphi_l - N}} = \left(\frac{\sum_{l=1}^N \varphi_l - N}{N-1} \right) \cdot \left(\frac{\left(\sum_{l=1}^N \varphi_l - 1 \right) \sigma_{jk}^2}{\sum_{l=1}^N (\varphi_l - 1) \sigma_{jk\varphi_l}^2} - 1 \right), \quad \begin{matrix} j=1,2,\dots,n-1 \\ k=(j+1),(j+2),\dots,n \end{matrix}$$

F-test: $F \rightarrow$ Fisher - ANOVA

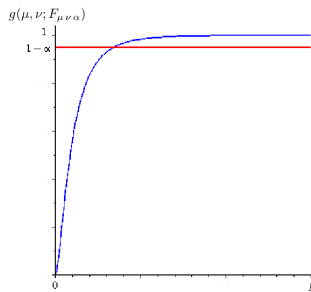
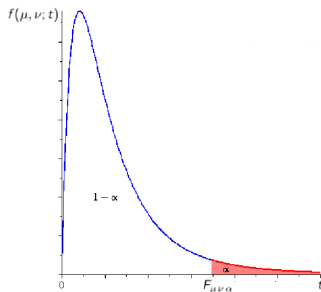
Comparison with:

pdf:

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

cdf:

$$g(\mu, \nu; F_{\mu \nu \alpha}) = \int_0^{F_{\mu \nu \alpha}} f(\mu, \nu; t) dt = 1 - \alpha$$



F-test: $F \rightarrow$ Fisher - ANOVA

μ = numerator degrees of freedom = $(N - 1)$,

ν = denominator degrees of freedom = $\left(\sum_{l=1}^N \varphi_l - N \right)$,

α = significance level

$$1 - \alpha = \frac{\Gamma(\frac{\mu+\nu}{2})}{\Gamma(\frac{\mu}{2})\Gamma(\frac{\nu}{2})} \mu^{\frac{\mu}{2}} \nu^{\frac{\nu}{2}} \int_0^{F_{\mu \nu \alpha}} \frac{t^{\frac{\mu}{2}-1}}{(\mu t + \nu)^{\frac{\mu+\nu}{2}}} dt$$

$$1 - \alpha = \frac{\Gamma(\frac{\mu+\nu}{2})}{\Gamma(\frac{\mu}{2})\Gamma(\frac{\nu}{2})} \left(\frac{\mu F_{\mu \nu \alpha}}{\nu} \right)^{\frac{\mu}{2}} \int_0^1 \frac{v^{\frac{\mu}{2}-1}}{(1 + \frac{\mu}{\nu} v F_{\mu \nu \alpha})^{\frac{\mu+\nu}{2}}} dv$$

F-test: $F \rightarrow$ Fisher - ANOVA

For $\mu \gg 1$, $\frac{\mu}{\nu} \ll 1$, we can write:

$$\left(1 + \frac{\mu}{\nu} v F_{\mu \nu \alpha}\right)^{\frac{\mu+\nu}{2}} \approx e^{\frac{\mu F_{\mu \nu \alpha}}{2} (1 + \frac{\mu}{\nu}) v} \approx e^{\frac{\mu F_{\mu \nu \alpha}}{2} v}$$

$$1 - \alpha \approx \frac{\Gamma(\frac{\mu+\nu}{2})}{\Gamma(\frac{\mu}{2})\Gamma(\frac{\nu}{2})} \left(\frac{\mu F_{\mu \nu \alpha}}{\nu}\right)^{\frac{\mu}{2}} \int_0^1 v^{\frac{\mu}{2}-1} e^{-\frac{\mu F_{\mu \nu \alpha}}{2} v} dv$$

$$\int_0^1 v^{A-1} e^{-Bv} dv = \frac{B^{-\frac{1}{2}(A+1)}}{A} e^{-\frac{1}{2}B} \text{Whittaker}\left(\frac{1}{2}(A-1), \frac{1}{2}A, B\right)$$

$$1 - \alpha \approx \frac{\Gamma(\frac{\mu+\nu}{2})}{\Gamma(\frac{\mu}{2})\Gamma(\frac{\nu}{2})} \left(\frac{\mu F_{\mu \nu \alpha}}{\nu}\right)^{\frac{\mu}{2}} \frac{\left(\frac{\mu F_{\mu \nu \alpha}}{2}\right)^{-\frac{1}{2}(\frac{\mu}{2}+1)}}{\frac{\mu}{2}} e^{-\frac{\mu F_{\mu \nu \alpha}}{4}} \\ \cdot \text{Whittaker}\left(\frac{1}{2}\left(\frac{\mu}{2} - 1\right), \frac{\mu}{4}, \frac{\mu F_{\mu \nu \alpha}}{2}\right)$$

Hypothesis Testing

Null hypothesis:

$H_0 : \langle (SM)_{jk}(\Phi_1) \rangle = \langle (SM)_{jk}(\Phi_2) \rangle = \dots = \langle (SM)_{jk}(\Phi_N) \rangle \Rightarrow$
invalidation of the “Clan” concept.

Alternative hypothesis:

$H_a : \langle (SM)_{jk}(\Phi_1) \rangle \neq \langle (SM)_{jk}(\Phi_2) \rangle \neq \dots \neq \langle (SM)_{jk}(\Phi_N) \rangle$ (not all
necessarily unequal) \Rightarrow existence of “Clans”.

Reject H_0 if $F_j > F_{\mu\nu\alpha} \Rightarrow$ Validity of the clan concept.

If $F_j < F_{\mu\nu\alpha}$ we cannot say unequivocally that “Clans” do not exist.

Some Technical Requirements for Data Validation

Assumptions for data to be used on ANOVA:

1. The $(m \times n)$ blocks from the N populations (“Clans”) are independent.
2. The $(m \times n)$ blocks should be normally distributed.
3. The $(m \times n)$ blocks should be selected from populations with equal variance $\sigma_{j\Phi_l}^2$.

Some comments are now in order:

Assumptions 2, 3 can be more or less relaxed by trusting on the robustness of ANOVA statistics and F-test.

We consider that assumption 3 is not violated if the “spreads” (differences between the extremum values of entropy measures for the $(m \times n)$ blocks on each “Clan”) are approximately the same.

F-test $\alpha = 0.01$

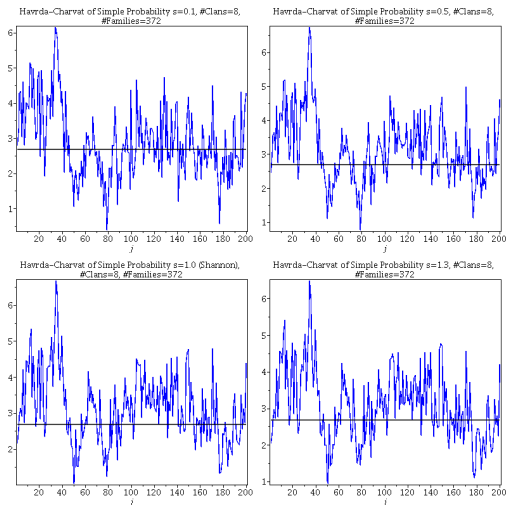


Figure: Variation of F experimental values with column number for a fixed number of clans = 8. F theoretical value is given by the height of the straight line (probabilities $p_j(a)$).

F-test $\alpha = 0.01$

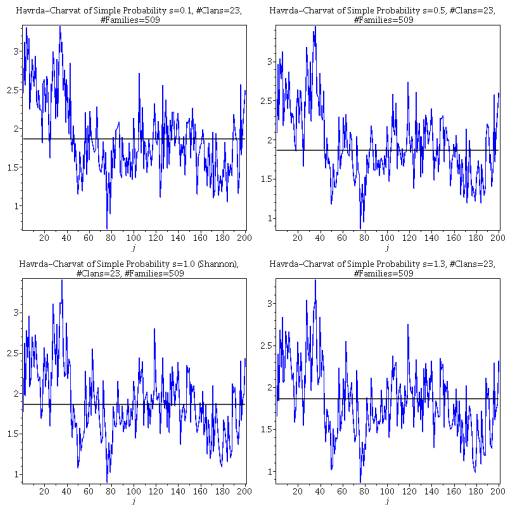


Figure: Variation of F experimental values with column number for a fixed number of clans = 23. F theoretical value is given by the height of the straight line (probabilities $p_j(a)$).

F-test $\alpha = 0.01$

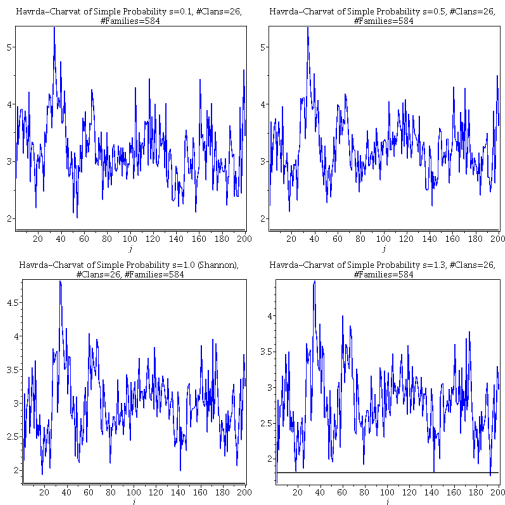
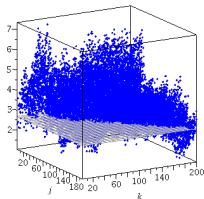


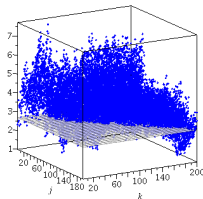
Figure: Variation of F experimental values with column number for a fixed number of clans = 26. F theoretical value is given by the height of the straight line (probabilities $p_j(a)$).

F-test $\alpha = 0.01$

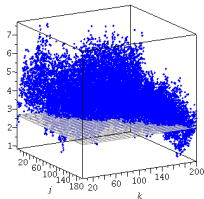
Havrda-Charvat of Joint Probability $s=0.1$, #Clans=8,
#Families=372



Havrda-Charvat of Joint Probability $s=0.5$, #Clans=8,
#Families=372



Havrda-Charvat of Joint Probability $s=1.0$ (Shannon),
#Clans=8, #Families=372



Havrda-Charvat of Joint Probability $s=1.3$, #Clans=8,
#Families=372

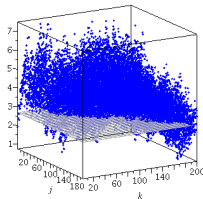
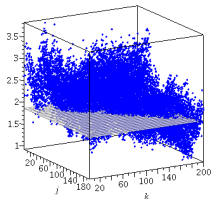


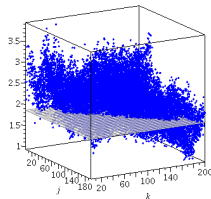
Figure: Variation of F experimental values with the ordered pair of columns for a fixed number of clans = 8. F theoretical value is given by the height of the plan (probabilities $p_{jk}(a, b)$).

F-test $\alpha = 0.01$

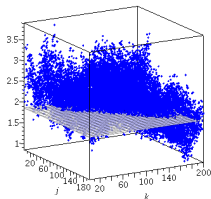
Havrda-Charvat of Joint Probability $s=0.1$, #Clans=23,
#Families=509



Havrda-Charvat of Joint Probability $s=0.5$, #Clans=23,
#Families=509



Havrda-Charvat of Joint Probability $s=1.0$ (Shannon),
#Clans=23, #Families=509



Havrda-Charvat of Joint Probability $s=1.3$, #Clans=23,
#Families=509

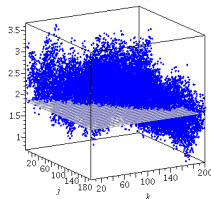
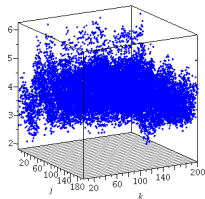


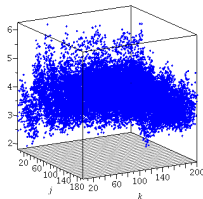
Figure: Variation of F experimental values with the ordered pair of columns for a fixed number of clans = 23. F theoretical value is given by the height of the plan (probabilities $p_{jk}(a, b)$).

F-test $\alpha = 0.01$

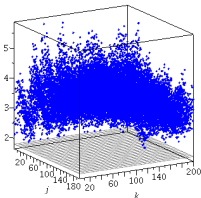
Havrda-Charvat of Joint Probability $s=0.1$, #Clans=26,
#Families=584



Havrda-Charvat of Joint Probability $s=0.5$, #Clans=26,
#Families=584



Havrda-Charvat of Joint Probability $s=1.0$ (Shannon),
#Clans=26, #Families=584



Havrda-Charvat of Joint Probability $s=1.3$, #Clans=26,
#Families=584

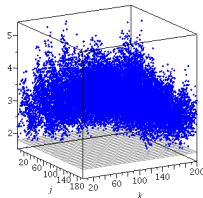


Figure: Variation of F experimental values with the ordered pair of columns for a fixed number of clans = 26. F theoretical value is given by the height of the plan (probabilities $p_{jk}(a, b)$).

Table: The number of “Clans” in successive experiments and the corresponding number of families.

n ^o of Clans	n ^o of Families
4	290
6	325
8	372
13	412
19	471
21	490
22	500
23	509
24	557
26	584
29	605
30	639
31	658
33	688
36	712
38	726
48	884
56	953
59	980
61	1029
68	1069

We have created “Pseudo-Clans” by exchanging families between the original “Clans”.

F-test $\alpha = 0.01$

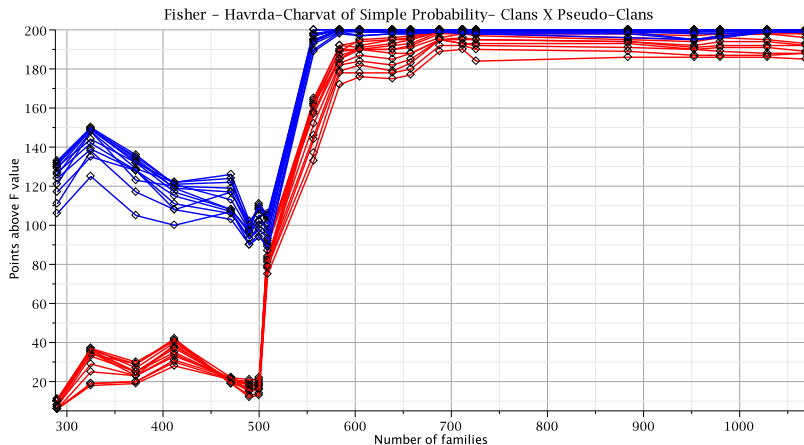


Figure: Number of F experimental values above the F theoretical value ($F_{\text{exp}} > F_{\text{theor}}$) for the cumulative number of families (probabilities $p_j(a)$). “Clans” are represented in blue and the “Pseudo-Clans” are represented in red.

F-test $\alpha = 0.01$

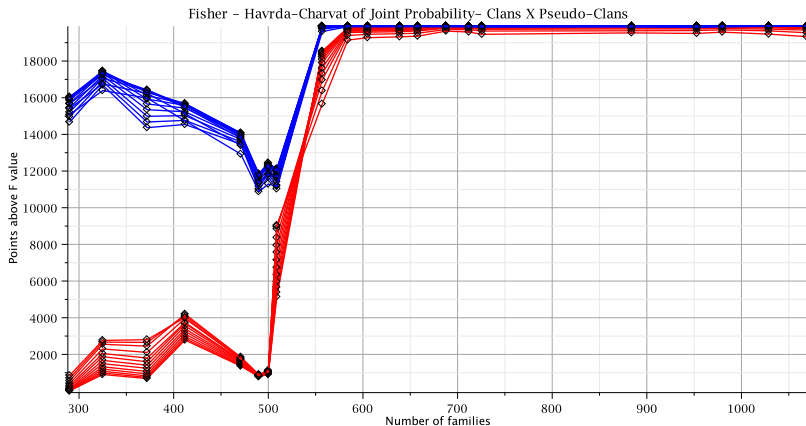


Figure: Number of F experimental values above the F theoretical value ($F_{\text{exp}} > F_{\text{theor}}$) for the cumulative number of families (probabilities $p_{jk}(a, b)$). “Clans” are represented in blue and the “Pseudo-Clans” are represented in red.

F-test $\alpha = 0.01$

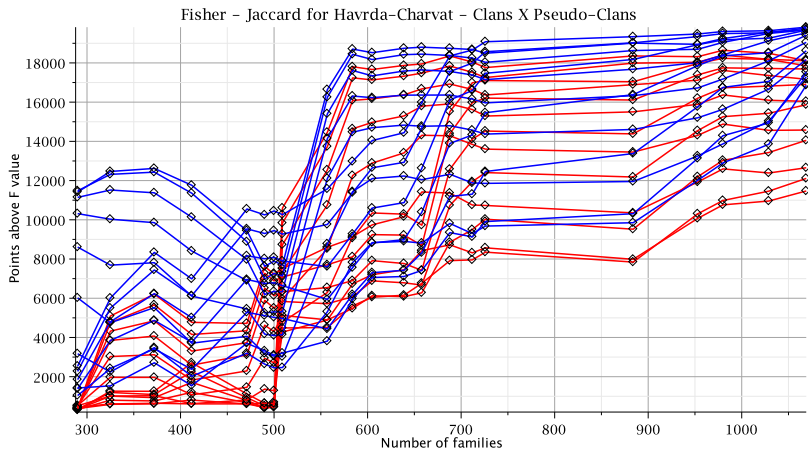


Figure: Number of F experimental values above the F theoretical value ($F_{\text{exp}} > F_{\text{theor}}$) for the cumulative number of families. “Clans” are represented in blue and the “Pseudo-Clans” are represented in red.

Conclusions and Suggestions for Improvement

- ▶ For blocks of (100×200) amino acids, we cannot say that these protein families are not classified into clans. This also means that we are not able to declare the existence of “clans”.
- ▶ The rejection of H_0 increases with the number of families. However, the rejection increases if natural clans are taken into account.
- ▶ ANOVA Statistics is not robust enough to the non-normality of data distribution. Use of other statistics to improve the results obtained by using Fisher's like Levine or Forsyth, could be advisable.
- ▶ A more rigorous validation of data for the F-test. Maybe the exclusion of “clans” with a greater spread of data.
- ▶ Considering an equal number of families on each clan.
- ▶ Greater number of families $\Rightarrow m < 100, n < 200$?