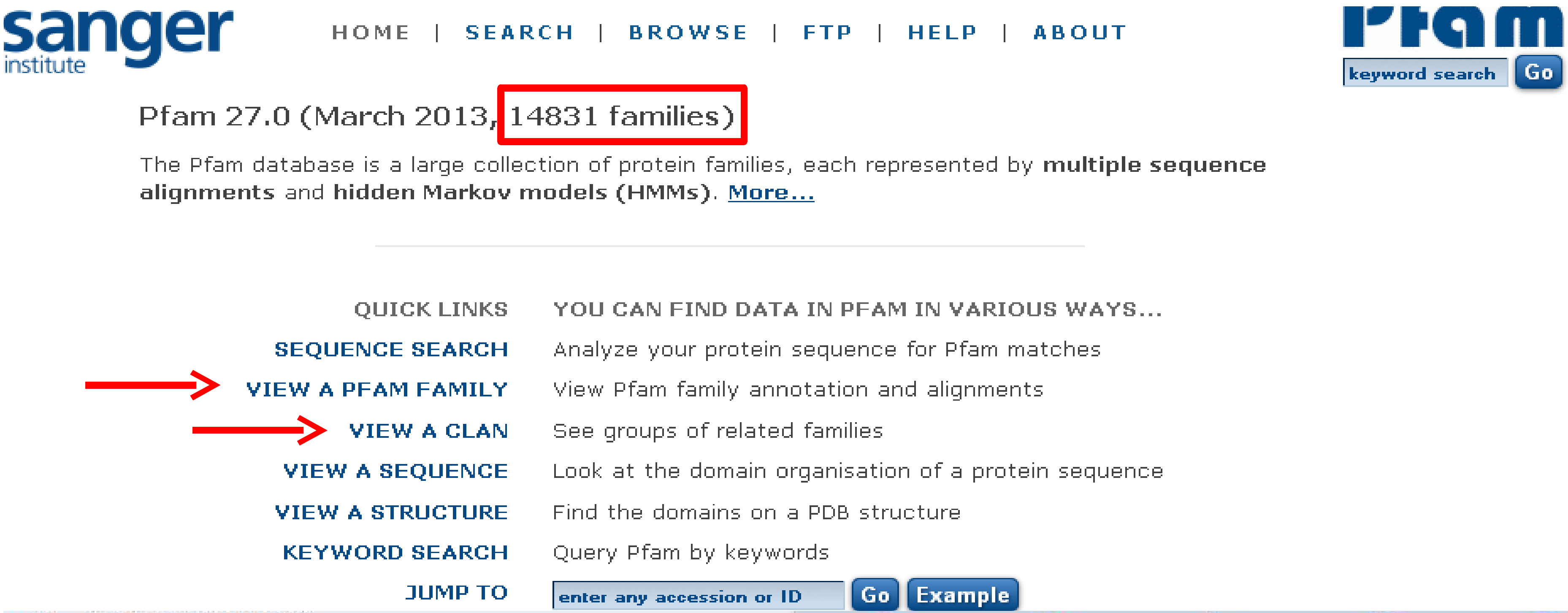
Assessment of protein domain space through entropy based methods

Nicolas Carels[1] and Rubem Mondaini[2]

[1] Ficoruz/CDTS, Rio de Janeiro, Brazil
[2] UFRJ, Rio de Janeiro, Brazil

nicolas.carels@gmail.com



sanger institute

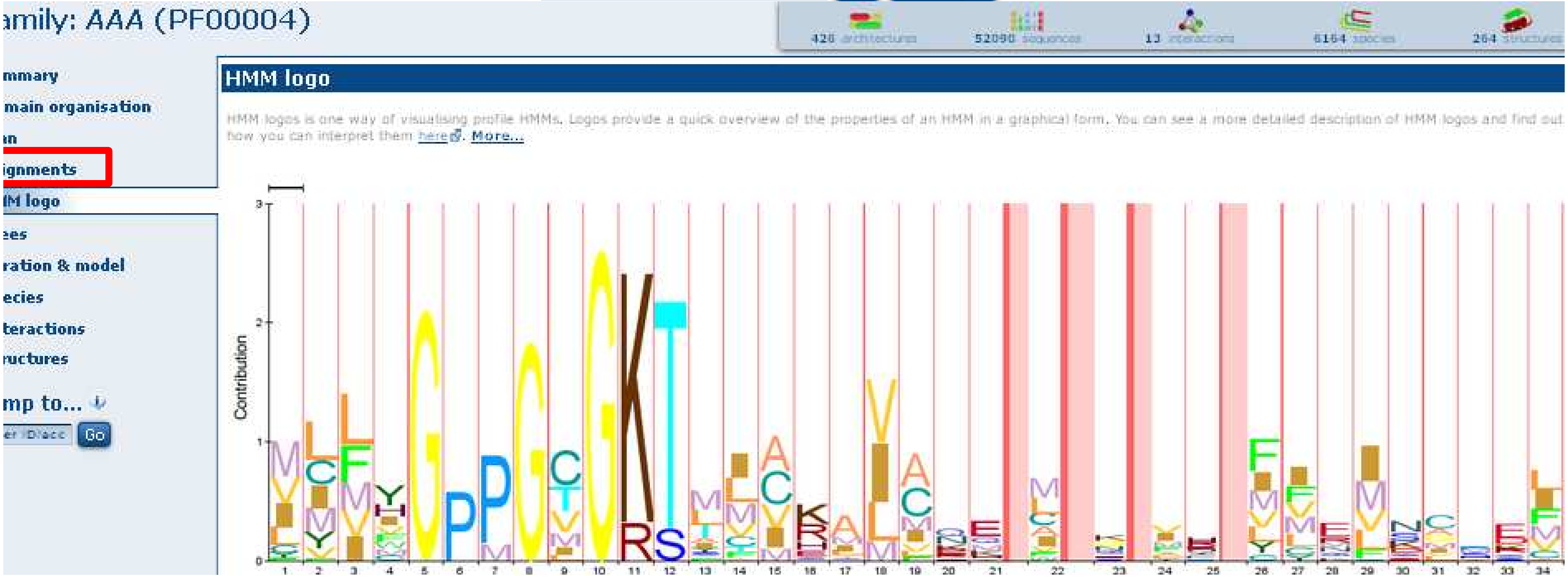
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

Pfam 27.0 (March 2013, 14831 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM FAMILY	View Pfam family annotation and alignments
VIEW A CLAN	See groups of related families
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	enter any accession or ID Go Example



amily: AAA (PF00004)

HMM logo

HMM logos is one way of visualising profile HMMs. Logos provide a quick overview of the properties of an HMM in a graphical form. You can see a more detailed description of HMM logos and find out how you can interpret them here. More...

How to translate domains in numbers?

- A domain has a **physico-chemical meaning**
  - Conserves **functionality** and **folding** after excision

=> Domain seqs. are **conserved** across a protein family

- **Domain seqs. form arrays** of residues (columns) per lines
- The **probability of a residue in a column** is ≠ from that of another
- The **column** makes sense since a protein domain is **+/- conserved** among species
- Because of the “+/-” -> **noise** -> entropy vs. **order**
- Family of **entropy measures -> characterize protein domains**
- PF07717 as an example:

off

```
>Q10752 DQLANLCERVEIELVTNSSESLDPIKKAITAGYFSNAARLDRSGDSYRTVKSNQTVYIHPSSSVAEKKPKVIIYFELVLTTKEYCRQITEIQPEWLLEISPH
>O22899 QQLVRIMSRFNLKMCSTDFNSRDYYVNIRKAMLAGYFMQVAHLERTGHYLTVKDNQVVHLHPSNCLDHKPEWVIYNEYVLTTRNFIRTVTDIRGEWLVDVAQH
>Q20875 TQLSRVMDKYNLRRVSTDFKSRDYYLNIRKALVAGFFMQVAHLERSGHYVTVKDNQLVNLHPSTVLDHKPEWALYNEFVLTTKNFIRTVTDVRPEWLLQIAPQ
>O35286 QQLSRIMDRFNLPRRSTDFTGRKYYINIRKALVTGYFMQVAHLERTGHYLTVKDNQVVQLHPSTVLDHKPEWVLYNEFVLTTKNYIRTCTDIKPEWLVKIAPQ
>O42945 KQLRRTMERQEVELISTPFDDKNYYVNIRRALVSGFFMQVAKKSANGKNYVTMKDNQVVSLHPSCGLSVTPEWVVYNEFVLTTKSFIRNVTAIRPEWLIELAPN
>O60231 EQLEGLLERVEVGLSSCQGDYIRVRKAITAGYFYHTARLTRSGYRTVKQQQTVFIHPNSSLFEQQPRWLLYHELVLTTKEFMRQVLEIESSWLLEVAPH
>Q9BKQ8 SQLVRLLKRFEIEKVSSRGLINCSENIRQCLVTGFFSQAAQYHYTGKYMTVKESFPFNMYKGSSIMFKKDYPKWVIFTEVMQDSIRDVTVIEPEWLYELAPH
>Q38953 KQLLSIMDKYKLDVVTAGKNFTKIRKAITAGFFFHGARKDPQEGYRTLVENQPVYIHPSSALFQRQPDWVIYHDLVMTTKEYMREVTVIDPKWLVELAPR
>O42643 KQLIRLMDRYRHPVVSCGRKRELILRALCSGYFTNVAKRDSHEGCYKTIVENAPVYMHPSGVLFGKAAEWVIYHELIQTSKEYMHTVSTVNPKWLVEVAPT
```

Measures

- **Shanon entropy**

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

- **Kullback–Leibler divergence**

$$D_{\mathrm{KL}}(P\|Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i)$$

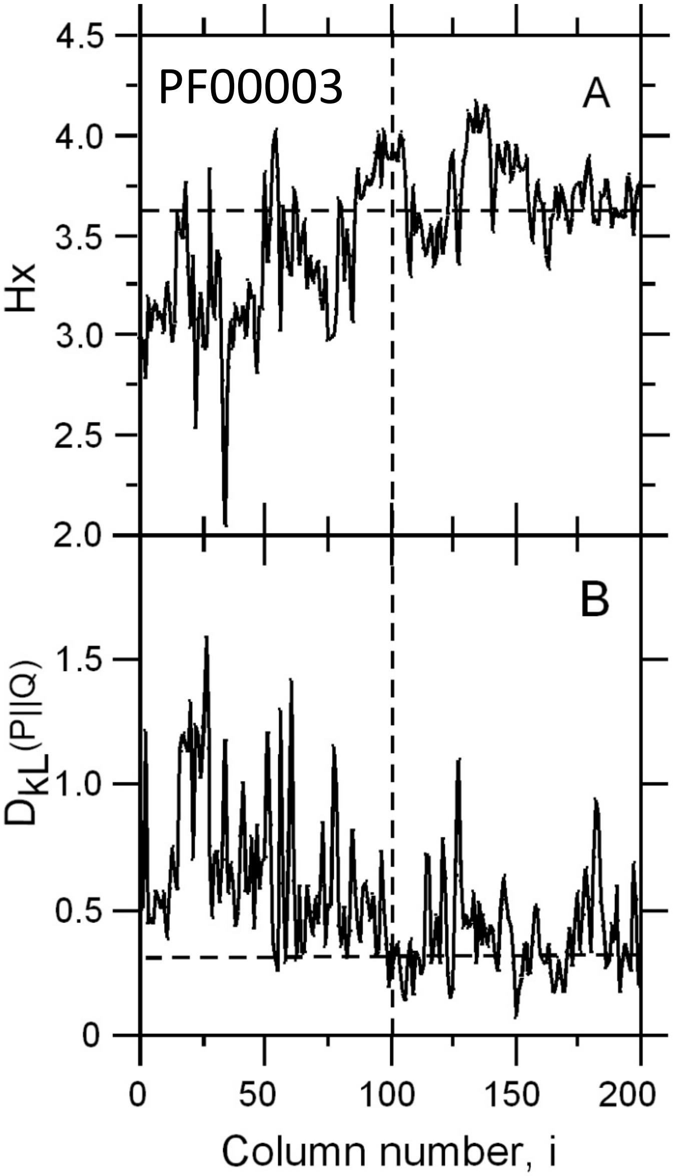
- **Mutual information**

$$I(X;Y) = \sum_{y\in Y} \sum_{x\in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

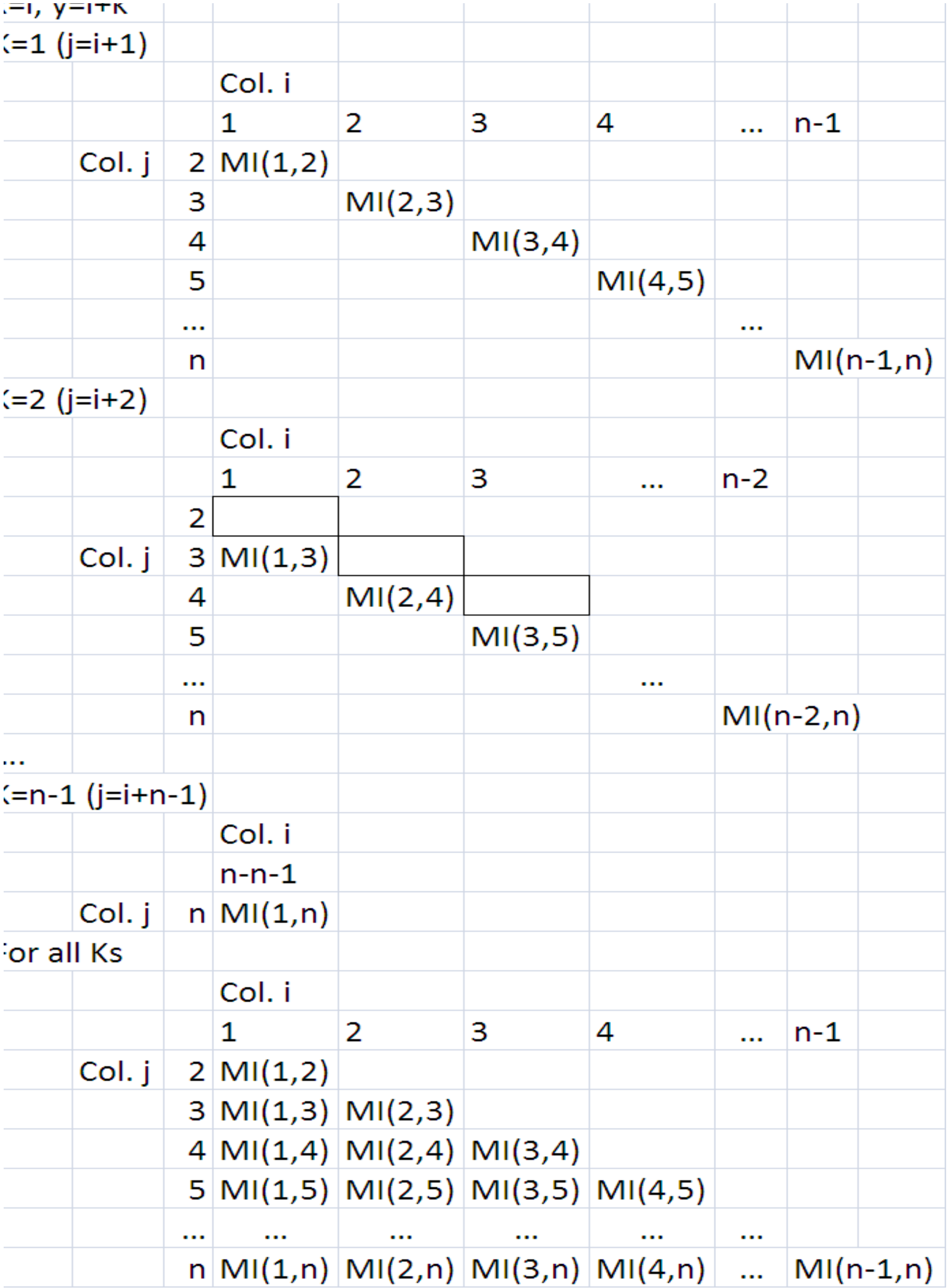
Entropy along sequences; robustn

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

$$D_{\mathrm{KL}}(P\|Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i)$$



Mutual information



=i, y=i+k									
=1 (j=i+1)									
			Col. i						
			1	2	3	4	...	n-1	
	Col. j	2	MI(1,2)						
		3		MI(2,3)					
		4			MI(3,4)				
		5				MI(4,5)			
		...					...		
		n						MI(n-1,n)	
=2 (j=i+2)									
			Col. i						
			1	2	3	...	n-2		
		2							
	Col. j	3	MI(1,3)						
		4		MI(2,4)					
		5			MI(3,5)				
		...				...			
		n					MI(n-2,n)		
...									
=n-1 (j=i+n-1)									
			Col. i						
			n-n-1						
	Col. j	n	MI(1,n)						
or all Ks									
			Col. i						
			1	2	3	4	...	n-1	
	Col. j	2	MI(1,2)						
		3	MI(1,3)	MI(2,3)					
		4	MI(1,4)	MI(2,4)	MI(3,4)				
		5	MI(1,5)	MI(2,5)	MI(3,5)	MI(4,5)			
		...	...	...	...	...	...		
		n	MI(1,n)	MI(2,n)	MI(3,n)	MI(4,n)	...	MI(n-1,n)	

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

- An average definition?

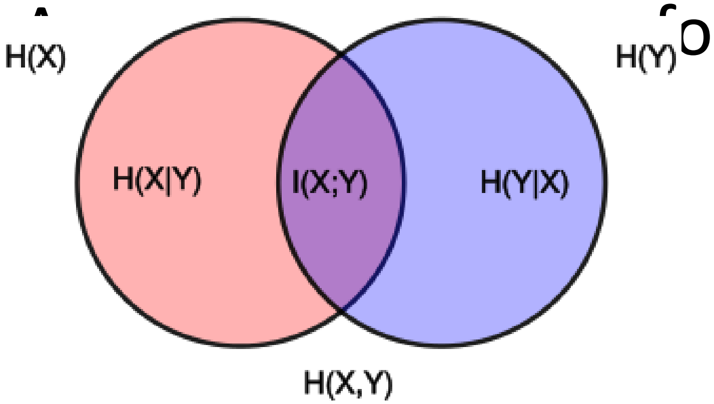
$$\Rightarrow M_{av} = \frac{2}{n(n-1)} \sum M_{i,k}$$

Other average measures

- An average measure for relative entropy (KL):

=> Jensen-Shanon: $JS = \frac{1}{n}\sum_{i=1}^{n}(D_{KL})_i$

- A for the Jaccard index:



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$I(X;Y) = H(X) - H(X|Y)$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

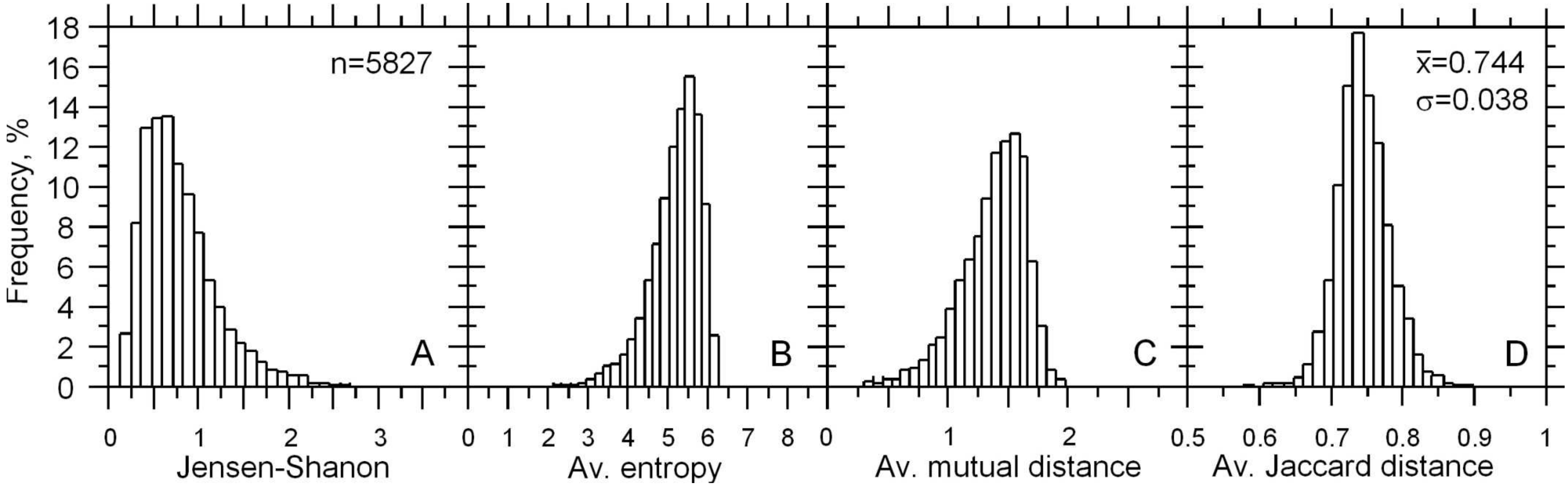
- Where *H(X|Y)* & *H(Y|X)* = the *conditional entropies* and *H(X,Y)* = the *joint entropy* of *X* & *Y*

- If $J_{i,k} = 1 - \frac{M_{i,k}}{H_{i,k}}$ then $J_{av} = \frac{2}{n(n-1)} \sum J_{i,k}$

Symmetry of probability distributions

$JS = \frac{1}{n} \sum (D_{KL})_i$

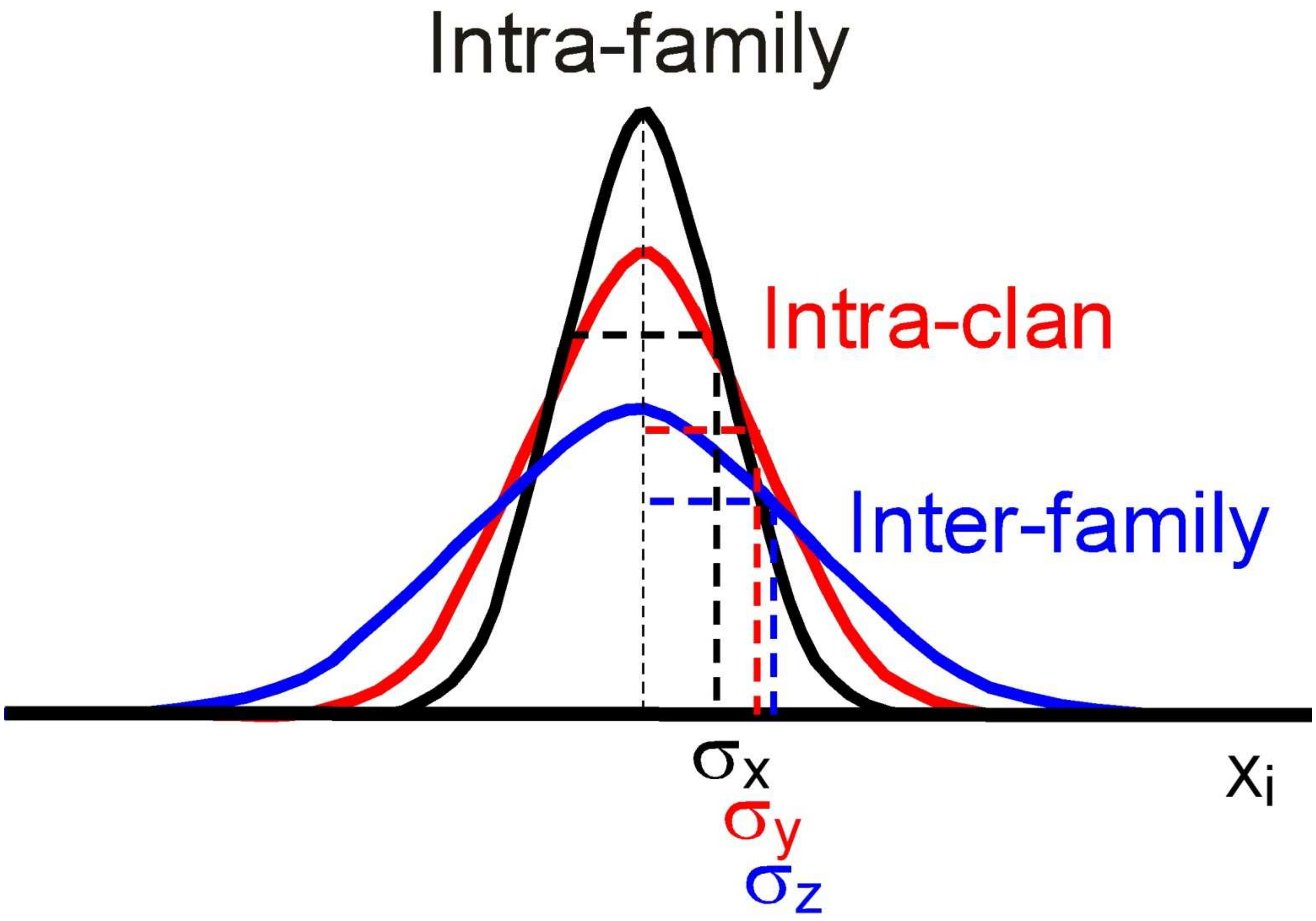
$M_{av} = \frac{2}{n(n-1)} \sum M_{i,k}$



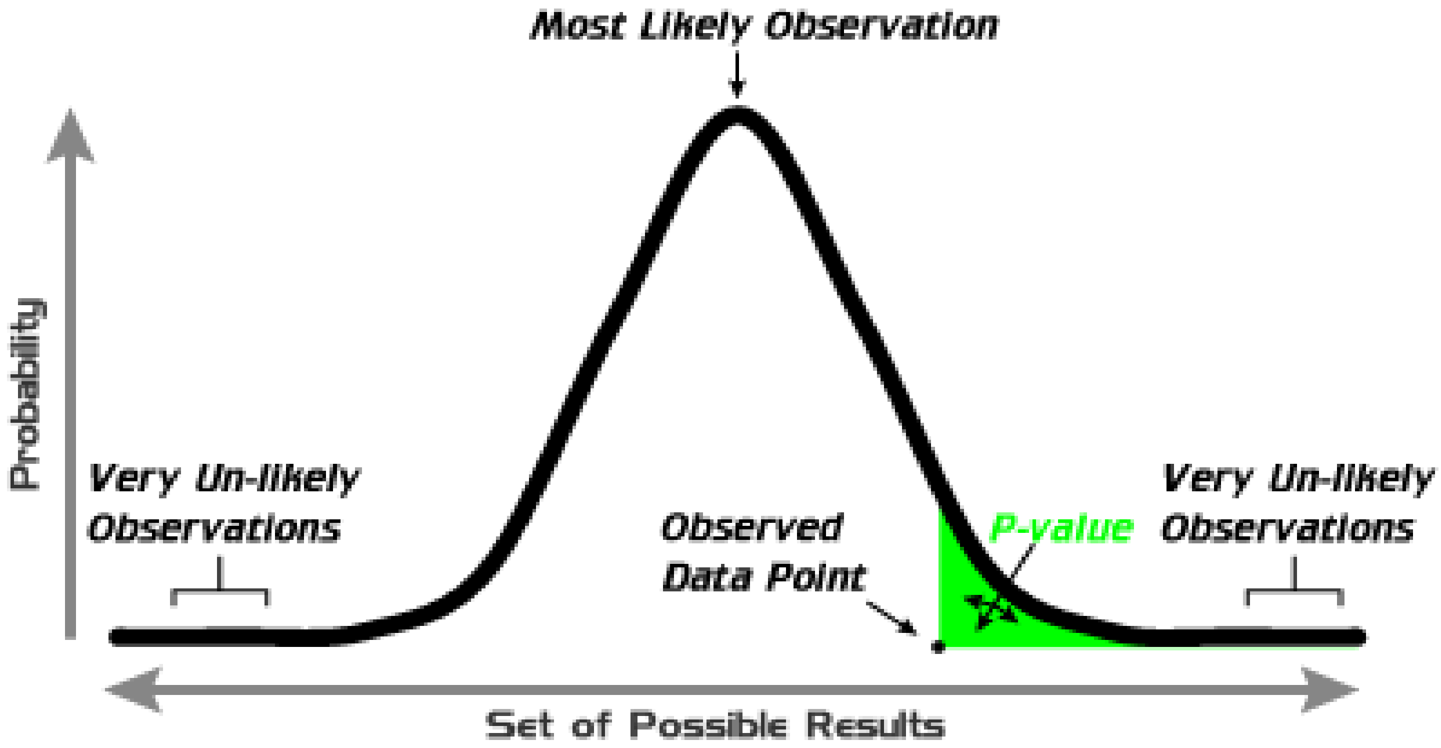
$H_{av} = \frac{2}{n(n-1)} \sum H_{i,k}$

$J_{av} = \frac{2}{n(n-1)} \sum J_{i,k}$

What to expect?



p-value



A p-value (shaded green area) is the probability of an observed (or more extreme) result arising by chance

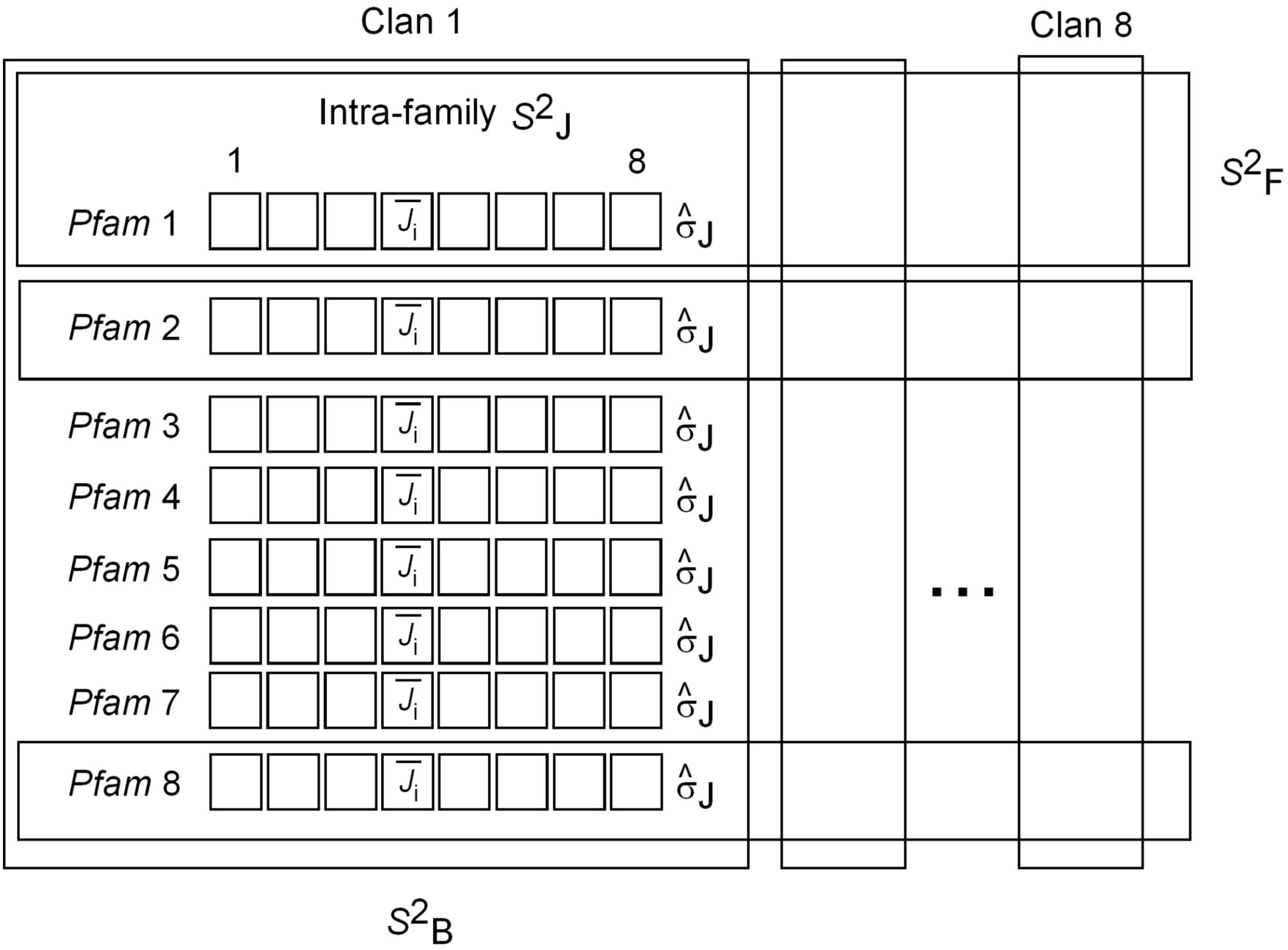
Does domain description with Av. Jaccard Distance makes sense? The F-test

- F-test particularly sensitive to variance ≠
- Condition: normality, n1=n2
- Given the variance of X and Y:

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \text{ and } S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^2$$

- The F-test is: $Fobs = \dfrac{\hat{\sigma}^2 max}{\hat{\sigma}^2 min}$
- With n1-1 (7) and n2-1 (7) degrees of freedom if the null hypothesis (*Ho*) *S2x = S2y* is true.
- *Ho* is true if *Fobs < Fth (1-α/2)=4.99 (α/2 = p*

Experimental design



Clan 1

Clan 8

Intra-family $S^2{}_J$

1

8

$S^2{}_F$

Pfam 1 $\overline{J}_i$ $\hat{\sigma}_J$

Pfam 2 $\overline{J}_i$ $\hat{\sigma}_J$

Pfam 3 $\overline{J}_i$ $\hat{\sigma}_J$

Pfam 4 $\overline{J}_i$ $\hat{\sigma}_J$

Pfam 5 $\overline{J}_i$ $\hat{\sigma}_J$

…

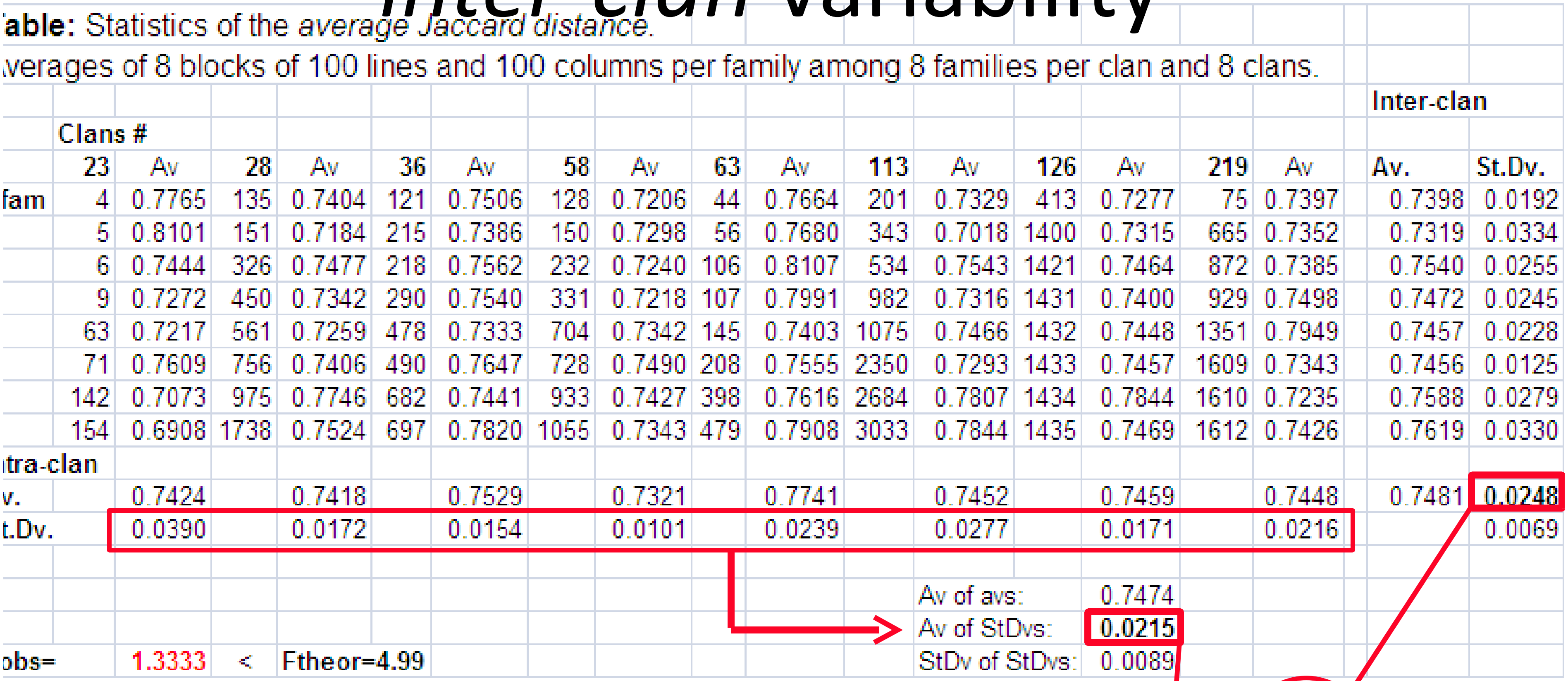
Pfam 6 $\overline{J}_i$ $\hat{\sigma}_J$

Pfam 7 $\overline{J}_i$ $\hat{\sigma}_J$

Pfam 8 $\overline{J}_i$ $\hat{\sigma}_J$

$S^2{}_B$

Av. Jaccard distance: *intra-clan* and *inter-clan* variability



able: Statistics of the *average Jaccard distance.*

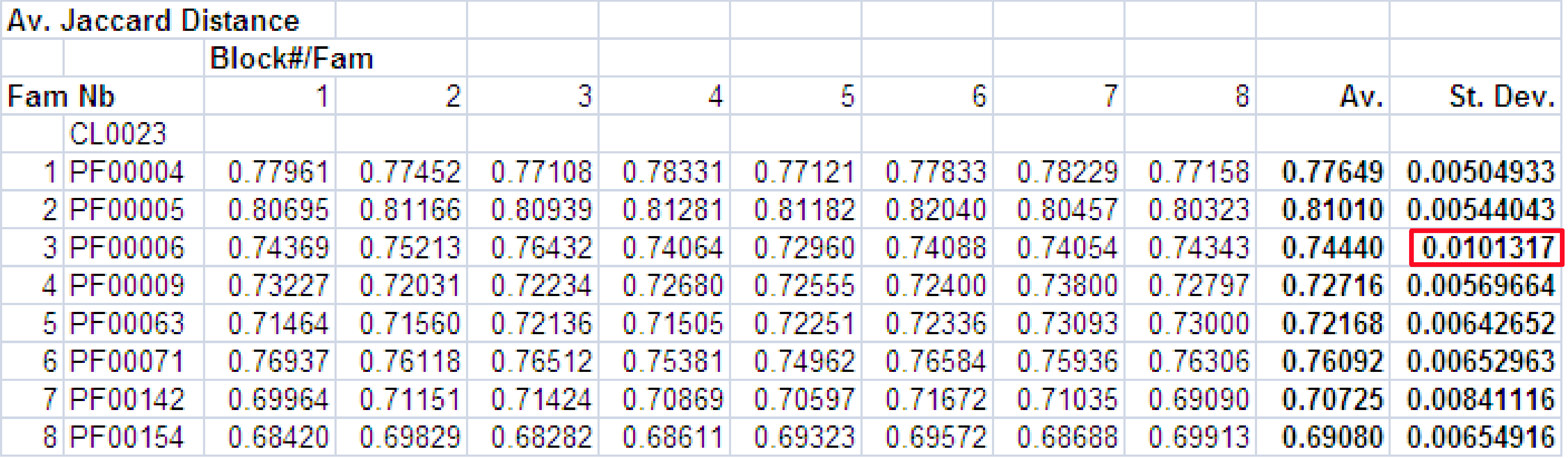
verages of 8 blocks of 100 lines and 100 columns per family among 8 families per clan and 8 clans.

																	Inter-clan	
	Clans #																	
	23	Av	28	Av	36	Av	58	Av	63	Av	113	Av	126	Av	219	Av	Av.	St.Dv.
fam	4	0.7765	135	0.7404	121	0.7506	128	0.7206	44	0.7664	201	0.7329	413	0.7277	75	0.7397	0.7398	0.0192
	5	0.8101	151	0.7184	215	0.7386	150	0.7298	56	0.7680	343	0.7018	1400	0.7315	665	0.7352	0.7319	0.0334
	6	0.7444	326	0.7477	218	0.7562	232	0.7240	106	0.8107	534	0.7543	1421	0.7464	872	0.7385	0.7540	0.0255
	9	0.7272	450	0.7342	290	0.7540	331	0.7218	107	0.7991	982	0.7316	1431	0.7400	929	0.7498	0.7472	0.0245
	63	0.7217	561	0.7259	478	0.7333	704	0.7342	145	0.7403	1075	0.7466	1432	0.7448	1351	0.7949	0.7457	0.0228
	71	0.7609	756	0.7406	490	0.7647	728	0.7490	208	0.7555	2350	0.7293	1433	0.7457	1609	0.7343	0.7456	0.0125
	142	0.7073	975	0.7746	682	0.7441	933	0.7427	398	0.7616	2684	0.7807	1434	0.7844	1610	0.7235	0.7588	0.0279
	154	0.6908	1738	0.7524	697	0.7820	1055	0.7343	479	0.7908	3033	0.7844	1435	0.7469	1612	0.7426	0.7619	0.0330
tra-clan																		
v.		0.7424		0.7418		0.7529		0.7321		0.7741		0.7452		0.7459		0.7448	0.7481	0.0248
t.Dv.		0.0390		0.0172		0.0154		0.0101		0.0239		0.0277		0.0171		0.0216		0.0069
												Av of avs:		0.7474				
												Av of StDvs:		0.0215				
obs=		1.3333	<	Ftheor=4.99								StDv of StDvs:		0.0089				

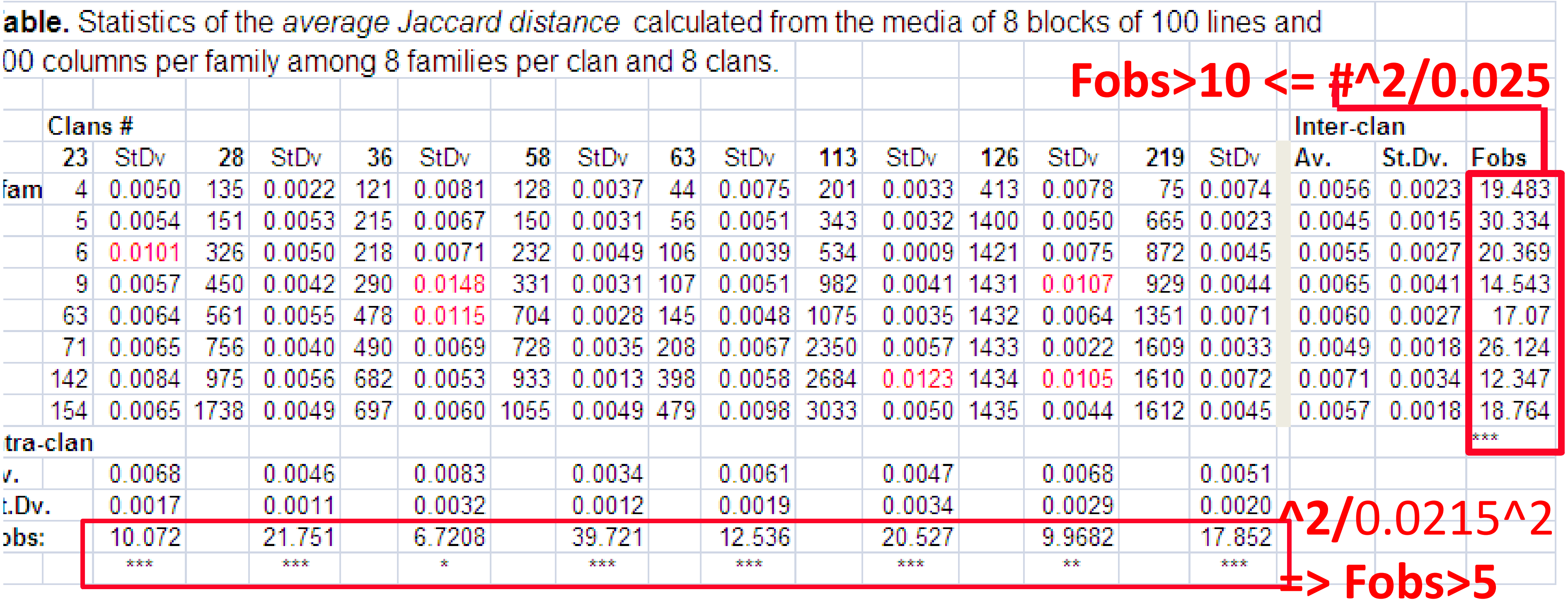
$\hat{\sigma}_F$ = 0.0248

The average standard deviation as an estimator: $\hat{\sigma}_{cl}$ = 0.0215

Av. Jaccard distance: the *intra-family* variability



Av. Jaccard Distance											
		Block#/Fam									
Fam Nb		1	2	3	4	5	6	7	8	Av.	St. Dev.
	CL0023										
1	PF00004	0.77961	0.77452	0.77108	0.78331	0.77121	0.77833	0.78229	0.77158	**0.77649**	**0.00504933**
2	PF00005	0.80695	0.81166	0.80939	0.81281	0.81182	0.82040	0.80457	0.80323	**0.81010**	**0.00544043**
3	PF00006	0.74369	0.75213	0.76432	0.74064	0.72960	0.74088	0.74054	0.74343	**0.74440**	**0.0101317**
4	PF00009	0.73227	0.72031	0.72234	0.72680	0.72555	0.72400	0.73800	0.72797	**0.72716**	**0.00569664**
5	PF00063	0.71464	0.71560	0.72136	0.71505	0.72251	0.72336	0.73093	0.73000	**0.72168**	**0.00642652**
6	PF00071	0.76937	0.76118	0.76512	0.75381	0.74962	0.76584	0.75936	0.76306	**0.76092**	**0.00652963**
7	PF00142	0.69964	0.71151	0.71424	0.70869	0.70597	0.71672	0.71035	0.69090	**0.70725**	**0.00841116**
8	PF00154	0.68420	0.69829	0.68282	0.68611	0.69323	0.69572	0.68688	0.69913	**0.69080**	**0.00654916**



able. Statistics of the *average Jaccard distance* calculated from the media of 8 blocks of 100 lines and 00 columns per family among 8 families per clan and 8 clans.

	Clans #																Inter-clan		
	23	StDv	**28**	StDv	**36**	StDv	**58**	StDv	**63**	StDv	**113**	StDv	**126**	StDv	**219**	StDv	Av.	St.Dv.	Fobs
fam	4	0.0050	135	0.0022	121	0.0081	128	0.0037	44	0.0075	201	0.0033	413	0.0078	75	0.0074	0.0056	0.0023	19.483
	5	0.0054	151	0.0053	215	0.0067	150	0.0031	56	0.0051	343	0.0032	1400	0.0050	665	0.0023	0.0045	0.0015	30.334
	6	0.0101	326	0.0050	218	0.0071	232	0.0049	106	0.0039	534	0.0009	1421	0.0075	872	0.0045	0.0055	0.0027	20.369
	9	0.0057	450	0.0042	290	0.0148	331	0.0031	107	0.0051	982	0.0041	1431	0.0107	929	0.0044	0.0065	0.0041	14.543
	63	0.0064	561	0.0055	478	0.0115	704	0.0028	145	0.0048	1075	0.0035	1432	0.0064	1351	0.0071	0.0060	0.0027	17.07
	71	0.0065	756	0.0040	490	0.0069	728	0.0035	208	0.0067	2350	0.0057	1433	0.0022	1609	0.0033	0.0049	0.0018	26.124
	142	0.0084	975	0.0056	682	0.0053	933	0.0013	398	0.0058	2684	0.0123	1434	0.0105	1610	0.0072	0.0071	0.0034	12.347
	154	0.0065	1738	0.0049	697	0.0060	1055	0.0049	479	0.0098	3033	0.0050	1435	0.0044	1612	0.0045	0.0057	0.0018	18.764
tra-clan																			***
v.		0.0068		0.0046		0.0083		0.0034		0.0061		0.0047		0.0068		0.0051			
t.Dv.		0.0017		0.0011		0.0032		0.0012		0.0019		0.0034		0.0029		0.0020			
bs:		10.072		21.751		6.7208		39.721		12.536		20.527		9.9682		17.852			
		***		***		*		***		***		***		**		***			

Fobs>10 <= #^2/0.025

^2/0.0215^2

=> Fobs>5

Intra-clan vs. inter-clan

- *Fobs* =1.33

=> *Fobs* < *Fth* (4.99)

- Ho cannot be rejected.
- Suggests that clans accumulate most of the inter-family variability?!
- One would expect: *S2intra-clan* < *S2inter-clan*
- Does Pfam **clans represent a division concept that reflects a biological reality**?
- **Difficult to conclude** given the small sample

Acknowledgements

- Thanks to

CNPq

Conselho Nacional de Desenvolvimento
Científico e Tecnológico