Supplementary Information

Viral Adaptation to Host: A proteome based analysis of codon usage and amino acid preferences

Iris Bahir[1], Menachem Fromer[2], Yosef Prat[2] and Michal Linial[1,3, *]

[1]Department of Biological Chemistry, Institute of Life Sciences,
[2]School of Computer Sciences and Engineering,
[3]The Sudarsky Center for Computational Biology, The Hebrew University of Jerusalem, 91904, Israel

The following supplementary information is available with the online version of this paper.

Supplementary Table S1 includes detailed information for all 121 known viruses that infect human. Data source for Figure 1B.

Supplementary Table S2 includes the detailed table for the viruses and the hosts that are analyzed in this study.

Supplementary data S1 shows the distance matrix for the similarity in amino acid distribution and codon usage between viruses and hosts mapped by their broad taxonomy by the 6 taxonomical groups: human, mammals - excluding human, vertebrate - excluding mammals, plants, insects, and bacteria. Data extension - Figure 4.

Supplementary Table S3 contains the GC content for the pairs of 30 viruses and their unique hosts (IDs are according to Table 1 and supplementary Table S2). Data source for Figure 6A.

Supplementary data S2 shows the complete distance matrix for the similarity in codon usage between pairs of 30 viruses and 30 hosts. Data extension - Figure 6B.

Supplementary Table S4 contains the data compiled for the partition of human infecting virus proteins to functional groups. Data source for Figure 7.

Supplementary Table S2

A list of 30 hosts and their organism ID according to NCBI taxonomy. The number of representative viruses that infect the specific hosts are listed (# Vir) as well as the sum of their proteins (# Pr). The statistical information includes the commutative number of the amino acids (# aa), the number of associated open reading frames (# ORF), the total number of the respective codons (# codons) and the average length of the viral proteins and the average number of codons for an ORF. The data only cover full length representative viruses from ViralZone (supported by SwissProt/UniProtKB). The data was used as a source for data in Figures 3-6.

	Viral hosts	Host name	NCBI TaxID	# Vir	# Pr	# aa	# ORF	# Codons	aa / Protein	Codon / ORF
HUM	Human	H. sapiens	9606	23	1104	471751	2027	793917	427	392
SQUM	Squirrel monkey	Saimiri	9520	1	75	33202	137	58837	443	429
MAC	Macaque	Macaca	9539	1	7	1704	16	2335	243	146
RAT	Rat	Rattus	10116	1	7	5023	29	17393	718	600
MOUS	Mouse	M. musculus	10090	1	7	2754	13	3556	393	274
PIG	Pig	S. scrofa	9823	3	295	93628	405	125470	317	310
BOVN	Bovine	B. bovis	9913	5	36	14567	61	23234	405	381
SHP	Sheep	O. aries	9940	1	28	8976	28	9004	321	322
HORS	Horse	E. caballus	9788	1	7	1136	24	2385	162	99
CAT	Cat	F. domesticus	9685	1	2	774	2	776	387	388
DOG	Dog	C. canis	9615	1	7	2357	14	5113	337	365
CHK	Chicken	G. gallus	9031	3	103	46673	532	189090	453	355
MOSQ	Tiger mosquito	A. Albopictus	7160	1	3	1532	3	1543	511	514
COMO	Codling moth	C. Pomonella	82600	1	143	36936	145	37370	258	258
CHIL	Rice stem borer	C. suppressalis	168631	1	468	77552	468	78020	166	167
AMMO	Noctuid moth	S. Frugiperda	7108	1	122	37391	136	43519	306	320
ARAB	Arabidopsis	C. Hayek	3701	1	5	1624	5	1629	325	326
LET	Lettuce	L. sativa	4236	1	6	3857	13	8186	643	630
RICE	Rice	O. sativa	4530	2	6	1436	6	1235	239	206
TOM	Tomato	S. lycopersicum	4081	4	16	3840	21	4993	240	238
BACI	Bacillus subtilis		1423	1	27	6064	32	6374	225	199
CHLM	Chlamydia psittaci		83554	1	8	1648	12	2268	206	189
ENCO	Enterococcus faecalis		1351	1	221	42234	126	14526	191	115
LACO	Lactococcus		1357	1	39	6900	40	7181	177	180
MYBC	Mycobacteria		1763	1	85	15251	86	15669	179	182
MYPL	Mycoplasma		2093	1	14	3272	14	3287	234	235
PSDO	Pseudomonas syringae		317	1	5	866	5	871	173	174
SALM	Salmonella typhimurim		602	1	56	10207	174	31106	182	179
STRP	Streptomyces coelicolor		1902	1	53	12454	67	15392	235	230
ECOL	Escherichia coli		562	1	708	154167	1375	273272	218	199
total				64	3663	1099776	6016	1777551	310	287

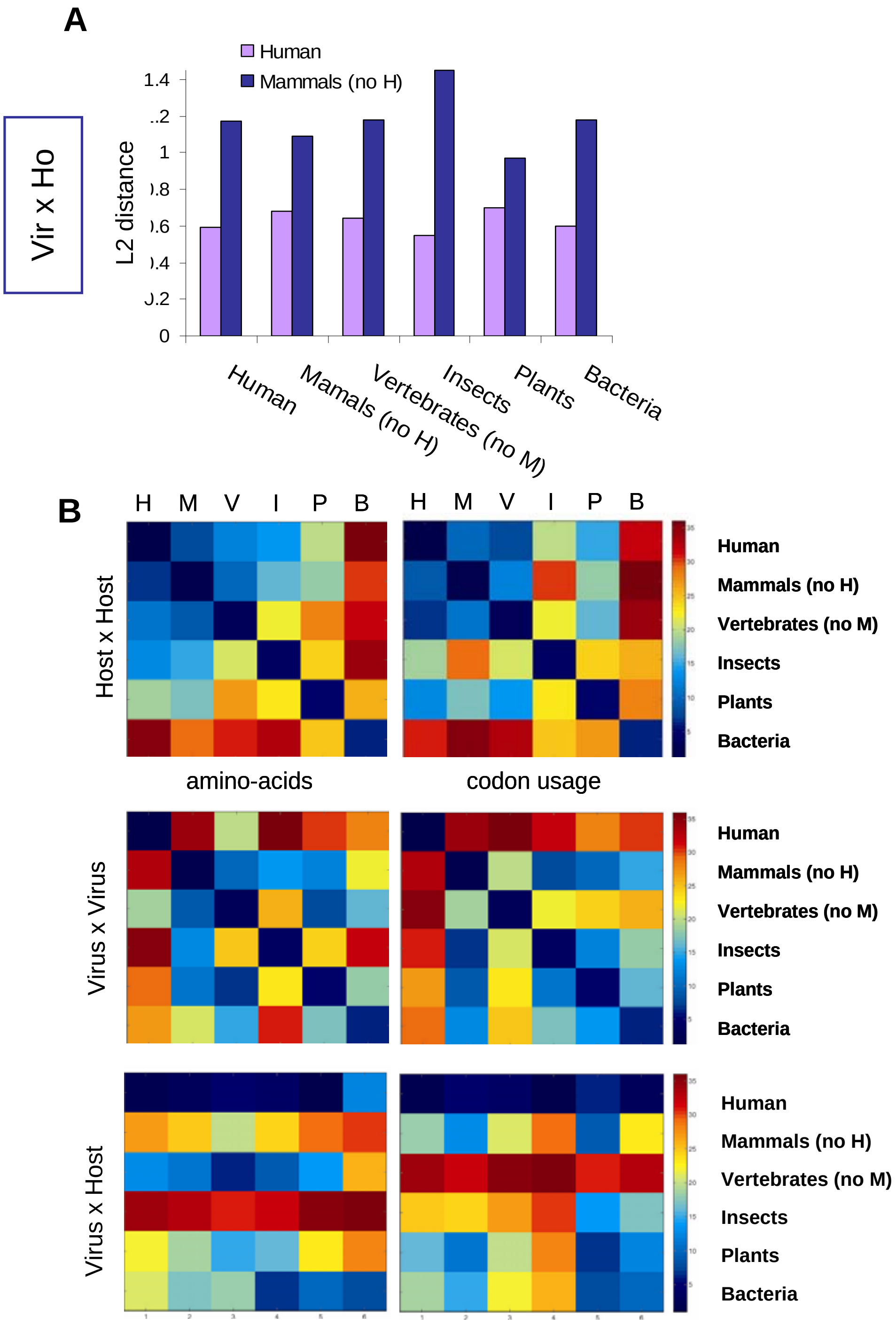
Supplementary Table S3

GC content (in percentage) for the pairs of 30 viruses and their hosts is shown. The abbreviate names refer to the list in Additional data file 2 and Table I.

	ID	GC - Host	GC-Virus
1	HUM	0.5066	0.5574
2	SQUM	0.5047	0.3555
3	MAC	0.4977	0.4194
4	RAT	0.5119	0.4689
5	MOUS	0.5200	0.4233
6	PIG	0.5254	0.3523
7	BOVN	0.5315	0.4180
8	SHP	0.5167	0.3405
9	HORS	0.5102	0.4759
10	CAT	0.5246	0.4372
11	DOG	0.5103	0.4332
12	CHK	0.5208	0.4399
13	MOSQ	0.4862	0.3969
14	COMO	0.4594	0.4679
15	CHIL	0.3999	0.2985
16	AMMO	0.5075	0.5137
17	ARAB	0.4471	0.3976
18	LET	0.4024	0.4385
19	RICE	0.5414	0.4607
20	TOM	0.4197	0.4165
21	BACI	0.3958	0.4016
22	CHLM	0.4141	0.3673
23	ENCO	0.3781	0.3575
24	LACO	0.3632	0.3673
25	MYBC	0.6683	0.625
26	MYPL	0.2984	0.3208
27	PSDO	0.5957	0.5694
28	SALM	0.5297	0.4823
29	STRP	0.7193	0.6391
30	ECOL	0.5149	0.3993

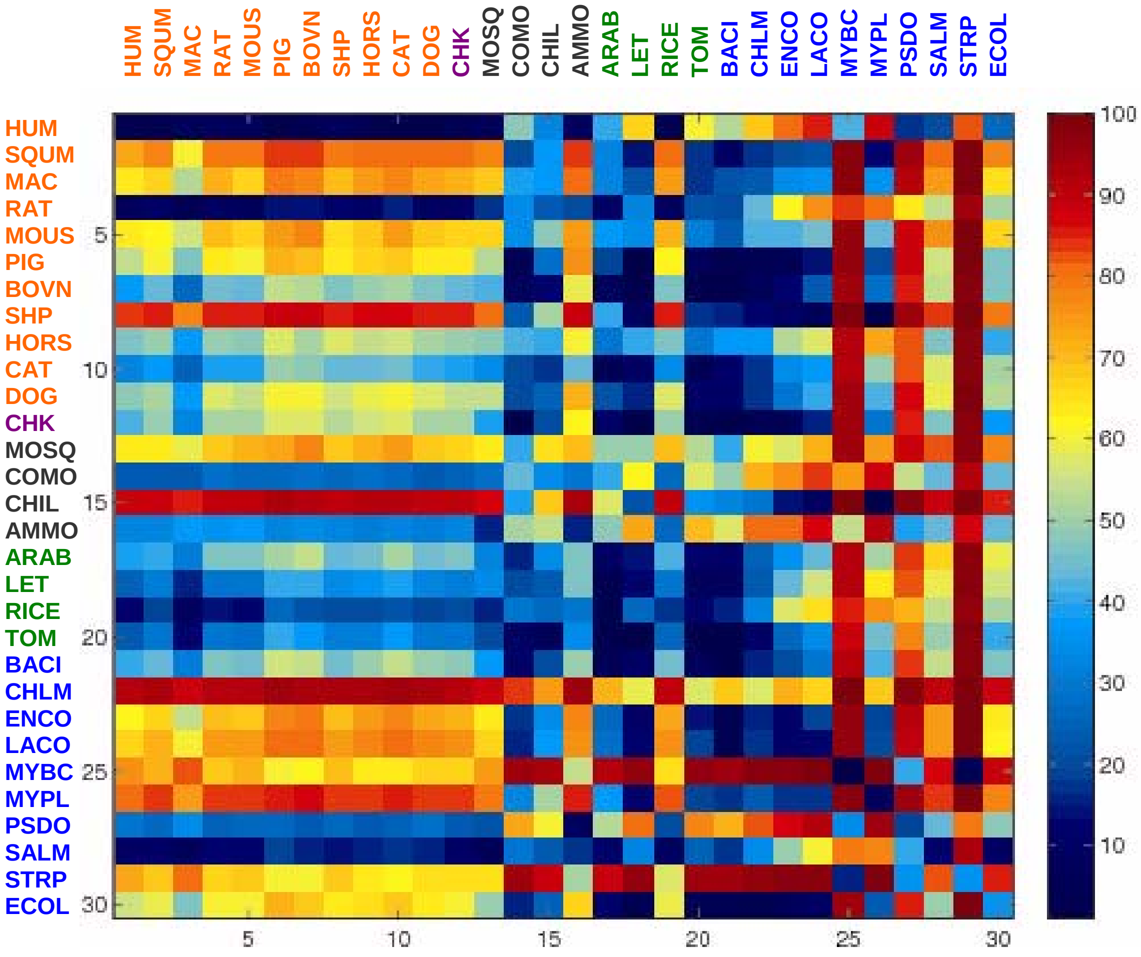
Supplementary data S1

Illustration for the L_2 distance values between human and mammalian viruses (excluding human) and 6 tested hosts groups as indicated (A). Note that a smaller L_2 value indicates high resemblance. The analysis was performed for all pairwise distances and the results were scaled according to the colored ruler (B). (see Materials and Methods and details in legend of Figure 4).



Supplementary data S2

A distance matrix based on the L_2 measure for the similarity in codon usage between 30 pairs of viruses and hosts. For details see legend of Figure 6. The abbreviation of the hosts (X-axis) and their viruses (Y-axis) are listed in Additional data file 2 and Table I. The colors used in the abbreviate names are as appears in legend of Figure 6A.


HUM
SQUM
MAC
RAT
MOUS
PIG
BOVN
SHP
HORS
CAT
DOG
CHK
MOSQ
COMO
CHIL
AMMO
ARAB
LET
RICE
TOM
BACI
CHLM
ENCO
LACO
MYBC
MYPL
PSDO
SALM
STRP
ECOL