

Triplet Codon Block Shannon Entropy (TCBShE) in terms of GC(1,2,3)% equates to Napier Constant for Model Organisms, and Harmonically Averages to same approximately: a Penta-Clado-genic Quantitative Survey across ~14.45 million Transcripts Clustered by 1118 Species

Praharshit Sharma^{1,*}, Kuralayanapalya Puttahonnappa Suresh^{1,\$}, Divakar Hemadri¹, S.S. Patil¹

¹ICAR- NIVEDI, PO Box 6450 Ramagondanahalli, Yelahanka, Bengaluru – 560 064 .

(Karnataka, India). *Email: sharmaji@iscb.org \$ suresh.kp@icar.gov.in

Abstract:

Background-

So far, several research efforts have tried to address the concept of Triplet Block Shannon entropy [TCBShE] computations pertaining to the context of Genetic codon [6]. Though dependence of block Shannon entropy values upon GC% was assessed, specifically GC-1% , GC-2% and GC-3% have not yet been taken into consideration – in this direction. Here, we utilize datasets from GC.evoBase (*Dapeng Wang, 2016*) to determine the typical TCBShE values and arrive at an interesting mathematical and numerical correlation, worthy of Biological interpretation.

Results-

Upon carrying out a comprehensive survey of 1118 species' GC-1,2,3 % values across 5 clades: namely 735 Fungi genomes, 68 Metazoa genomes, 44 Plant genomes, 186 Protist genomes and 85 Vertebrate Ensembl-release genomes respectively; from GC.evoBase datasets, we apply the appropriate formula based on 64 codon Trimers Binarily classified into 8 sets of 3 Blocks - { 000, 001, 010, 011, 100, 101, 110 and 111 } to compute TCBShE. It is observed that HM: Harmonic-Mean of these Entropy values, which in the language of Information theory and coding is the Ratio of “Mutual Information to complement of Normalized Variation of Information” ; and in the case of many Model Organisms the TCBShE values themselves – converge approximating to Napier’s constant/ Base of Natural logarithms. **HM of TCBShE for “Protists” is nearest to $e \sim 2.71828...$**

Conclusions-

Here, the approximation to Napier’s constant that we have attained by considering HM of TCBShE is a sort of Lower-bound and is clearly expressed in Bits. This may very well be corroborated with the direct implications of solving the HyperProteoGenomic–equation, as follows:

$$4^{4^x} = 20^{20^1} = 1.048576 \times 10^{26}$$

where in Equation above, 4 = Number of cDNA nucleotides (A|C|G|T) and

20 = Number of Amino-acids, and interestingly, $x = 99.9455\%$ Close to e , Napier constant.

Abbreviations Used-

TCBShE	Triplet Codon Block Shannon Entropy
HM	Harmonic Mean
TCRBE	Triplet Codon Rényi Block Entropy

1. INTRODUCTION

The entire end-to-end Pipeline accessing open access datasets from GC.evoBase has been highlighted in the Github repository link given below, herein referred to as “GC123e Github repo”:

<https://github.com/bioinformer/GC123e> [1]. After running all the Shell scripts in GC123e Github repo above, in Serial order (from 0_ to 9c_), we see that the 13th column is of interest to us, as the input into R-script described in 9d_ , whereupon we compute Harmonic-mean of TCBSHe values to approximate their HM to $e = 2.71828$, a fundamental Mathematical constant called Napier’s constant – which can also be addressed as the Base of Natural logarithms (\ln).

For the sake of easy Illustration, below is **TCBA** (Triplet Codon Block Alignment) for **228 bp** (Modulo 3, =0) envelope protein, belonging to RefSeq NC_045512.2 of Severe acute respiratory syndrome coronavirus 2, which is named **envelope.fa** (FastA header is deliberately excluded):

ATGTACTCATTCGTTTCGGAAGAGACAGGTACGTTAATAGTTAATAGCGTACTTCTTTTTT
CTTGCTTTTCGTGGTATTCTTGCTAGTTACACTAGCCATCCTTACTGCGCTTCGATTGTGT
GCGTACTGCTGCAATATTGTTAACGTGAGTCTTGTA AACCTTCTTTTTTACGTTTACTCT
CGTGTTAAAAATCTGAATTCTTCTAGAGTTCCTGATCTTCTGGTCTAA

123 Start->ATG	123 TTT	123 TTG	123 GTT
TAC	CTT	TGT	TAC
TCA	GCT	GCG	TCT
TTC	TTC	TAC	CGT
GTT	GTG	TGC	GTT
TCG	GTA	TGC	AAA
GAA	TTC	AAT	AAT
GAG	TTG	ATT	CTG
ACA	CTA	GTT	AAT
GGT	GTT	AAC	TCT
ACG	ACA	GTG	TCT
TTA	CTA	AGT	AGA
ATA	GCC	CTT	GTT
GTT	ATC	GTA	CCT
AAT	CTT	AAA	GAT
AGC	ACT	CCT	CTT
GTA	GCG	TCT	CTG
CTT	CTT	TTT	GTC
CTT	CGA	TAC	TAA->Stop
123	123	123	123

Figure-1: TCBA Script – `$ cat envelope.fa | grep -v "^>" | fold -w3`

1,2,3 being the Codon positions above, we define in a straight-forward manner:

$$GC1 = (G1 + C1) / (A1 + C1 + G1 + T1)$$

$$GC2 = (G2 + C2) / (A2 + C2 + G2 + T2)$$

$$GC3 = (G3 + C3) / (A3 + C1 + G1 + T1)$$

where,

B_i = Count of Base B (A|C|G|T) at position i (i = {1,2,3}) as per the Schematic Figure-1 above.

2. MATERIALS AND METHODS

A 5-iterations procedure is followed as outlined below to first compute individual TCBS_{HE} values, and subsequently their Harmonic Means, across the 5 clades: Fungi, Metazoa, Plsnts, Protists and Ensembl release vertebrates. Unless otherwise stated, all dry-run analysis were carried out on a Ubuntu 20.04.1 operating system LTS (Long-Term-Support) Desktop version, having x86_64 architecture, Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz workstation. Furthermore, a thorough Re-analysis is performed across several Transcripts belonging to a particular Clade, grouped by individual Species within each such Clade considered, as made available in the GC.evoBase [2].

2.1 Data Retrieval:

This is performed using the Shell-scripts mentioned in 1_GC123_Download.sh program, archived at GC123e Github repo. The total CPU time for the entire downstream analysis prior to R-script based computation of Harmonic-mean, inclusive of downloading All 5 datasets was recorded and the same is presented in Results section 3.1.

In terms of memory-sizes, the 735 Fungi genomes spanned 147MB, 68 Metazoa genomes: 29MB, whilst 44 Plant genomes and 186 Protist genomes occupied 44MB each and 85 Vertebrate Ensembl-release genomes amounted to 55MB of memory space. It may be noted that 1MB = 10,48,576 Bytes – as can be validated with the Log-file, 2_wgetLogFile_Download.txt in GC123e Github repo.

In each of the 5 datasets corresponding to the respective Clade, the Columns contained were: species_name, transcript_name, gene_name, effective_length, completeness (Ternary-status: Complete, Partial or Perfect) followed by the 5 compositions: gc%, gc1%, gc2%, gc3% and gc4d%.

2.2 Data Inclusion and Exclusion criteria:

Amongst all the 10 columns in each of 5 “raw” text flat files, we filter only the Species name (1st column) and GC-1% , GC-2% and GC-3% (that is, 7th , 8th and 9th Columns). Hence, rest of the data, not being relevant for our work, is excluded from further downstream analysis. This data filtration step is done using Iteration-1 mentioned in 6, 7, 8 and 9a_ , 9b_ shell scripts outlined in GC123e GitHub repo. Headers of the data flat files are also excluded right from this 1st Iteration itself, thereby this being a “Data-cleansing” step as well. Typical values of GC-1% , GC-2%, GC-3% based on this Step are tabulated in Results section 3.2, more extensive Supplementary data are also tabulated in XL format flat files. Herein the four columns of the new 5 files so generated become “species_name, GC-1% , GC-2% and GC-3%” grouped by Transcripts within same species.

2.3 Downstream Analysis to Compute TCBS_{HE} for individual Transcripts in all Species

Now, since All probabilities to be considered while computing TCBS_{HE} need to be in [0.1] range, the Percentage values of GC-1% , GC-2% and GC-3% each are normalized to [0,1] simply via. dividing them each by 100 – as described in Results section 3.3. It may be noted that due to the constraints imposed by the Nature of the Universal Genetic Code, it is not feasible for any biological organism to attain theoretically extreme GC-content limits of either 0% or 100%. Hence, we see that a stricter bound on the Range of Normalized GC-123% values is (0,1) and not [0,1] – as per standard Mathematical notations: (x,y) ; [x,y] indicate excluding and including x,y respectively.

Now the 8 TT (Three-Tuple) Combinations of Triplet Codon position-wise Probabilities are calculated as depicted in Table-1 below.

Assuming a Binary Status ($0_k = GC_k$, $1 = AT_k = (1 - GC_k)$) across 3 triplet Codon positions [(i.e.,) for $k = \{1, 2, 3\}$], we have 8 combinations as follows, which correspond to the following 8 Independent Probabilities expressed in terms of (0,1)- Normalized $GC_k = (GC_k\% / 100)$.

Table-1: Eight TT Combinations of Triplet Codon position-wise Probabilities wrt GC-1,2,3 %

Binary TT Codon Block	1 st Position Probability	2 nd Position Probability	3 rd Position Probability
000	GC1	GC2	GC3
001	GC1	GC2	1 - GC3
010	GC1	1 - GC2	GC3
011	GC1	1 - GC2	1 - GC3
100	1 - GC1	GC2	GC3
101	1 - GC1	GC2	1 - GC3
110	1 - GC1	1 - GC2	GC3
111	1 - GC1	1 - GC2	1 - GC3

Following this assignment of normalized GC_k probabilities as per above Table-1, the Block-wise probability is simply computed as the “product” of respective tri-positional probabilities (assuming that the GC_k values implicate Independent events, though its not quite the case actually observed in Nature: Due to inter-mono-Nucleotide interactions within dinucleotides, inter-di-nucleotide interactions within Codon (tri-nucleotides) – extended till inter-tri-nucleotide (Codon1-Codon2 interactions) within a Dicon (2 consecutive triplet Codons) [3]. For example,

Joint-Triplet Codon probability in case of **101**, $p = (1 - GC1) * GC2 * (1 - GC3)$

where, $GC1 = (GC-1\% / 100)$, $GC2 = (GC-2\% / 100)$ and $GC3 = (GC-3\% / 100)$.

This determination of 8-joint Binary TT-Codon Block probabilities is scripted in Iteration-3 of below files in GC123e Github repo:-

6_Fungi_GC123_b3ent.sh

7_Metazoa_GC123_b3ent.sh

8_Plants_GC123_b3ent.sh

9a_Protists_GC123_b3ent.sh

9b_Release_GC123_b3ent.sh

Whereas in Iteration-4 of above self-same Shell-scripts, individual Combinatorial Entropies are evaluated by LOG (Base-2) transformation [$(-p) * \log_2(p)$] onto a For-loop. Subsequently, All the 8 separate Entropies are added-up as per the Summation-definition of Shannon Entropy in so far as to obtain the final respective TCBSHE values: for individual Transcripts in all Species across the 5 Clades. This is done so with respect to the Web reference Video, <https://youtu.be/B3dVuP0Kzg0>

2.4 Determination of Harmonic Mean among TCBShe values within each of the 5 Clades

The Harmonic mean (HM) of a finite numbers of numerical terms can be computed by dividing the number of terms (Cardinality) by the sum of reciprocals each term. For a bi-Cardinal set {a,b} :

$$HM(a,b) = \frac{2ab}{a+b}$$

upon Simplification. We choose HM as the measure of central tendency for TCBShe values due to some of its important properties, that deem Merit:

HM is based on all the observations, and is rigidly defined. Harmonic mean gives less weightage to the large values and large weightage to the small values to balance the values correctly. In general, the harmonic mean is used when there is a necessity to give greater weight to the smaller items. So for the purpose of Entropy minimization, our choice of ascertaining HM of TCBShe is justified.

Here, we have used **pacman** , a requisite cran-R package manager for decisively computing the proximity of HM pertaining to **e** being. Initially pacman has been employed to check for **psych** package (Version 2.1.6), install the same if it doesnt exist in R/Rstudio, and load **psych Library** – wherein calculating HM becomes much simpler due to its Built-In **harmonic.mean()** Function.

The corresponding Results are tabulated in Results Section 3.4.

2.5 Re-Analysis for determining TCBShe across Multiple Transcripts within same Clade, grouped by individual Species/ Strains/ Sub-species

Herein, invoking the same package manager namely, **pacman** as described in 2.4 above, we had installed and configured **dplyr** package on R-studio version 4.0.4 that helps us “cluster” several unique Transcript accessions (2nd column of GC.evoBase datasets) belonging to the same Genus-species nomenclature pairs corresponding to a particular organism/ strain within a given Clade (1st column of GC.evoBase datasets). This action of clustering (or) ‘Group-By’ is facilitated by **%>%** Operator in cran-R / Rstudio 4.0.4 , upon loading the **dplyr** Library in R-programming instance. The respective results are mentioned in a Tabular format in Tables 14 to 18 of Results section 3.5.

3. RESULTS

3.1 CPU time for Data Extraction, followed by *apriori* R-script Downstream analyses

This is given in Table-1 as displayed below.

Table-2: CPU Time for entire Dry-run including Data Download, in GC123e Github repo

CPU Time	real-Time	29m + 14.758s
	user-Time	11m + 35.710s
	system-Time	3m + 50.274s

Digits after decimal-point succeeding '+' sign indicate time-taken for dry-run in 'milli-seconds' (ms). It can be inferred from above that, [Real-time > User-time > System-time] – which are obtained by prefixing **time** command while executing 0_TCBSHe_Run_PipeLine.sh Shell-script.

3.2 Typical values of GC-1% , GC-2% and GC-3% across 5 Clades

Multiple GC(1,2,3)% for same species actually indicates they are across multiple transcripts in each

3.2.1 FUNGI

Table-3: Typical values of GC-1% , GC-2% and GC-3% for FUNGI (Several Transcripts)

Species	GC-1%	GC-2%	GC-3%
Acidomyces_richmondensis_bfw	58.5	48.62	54.94
Acidomyces_richmondensis_bfw	56.46	40.11	59.37
Acidomyces_richmondensis_bfw	55.69	43.89	49.9
Acidomyces_richmondensis_bfw	61.41	52.54	64.13
Acidomyces_richmondensis_bfw	57.84	44.61	64.22

3.2.2 METAZOA

Table-4: Typical values of GC-1% , GC-2% and GC-3% for METAZOA (Several Transcripts)

Species	GC-1%	GC-2%	GC-3%
Bombyx_mori	43.85	38.46	29.23
Bombyx_mori	47.17	35.22	36.9
Bombyx_mori	65.09	58.02	64.15
Bombyx_mori	49.05	33.4	23.26
Bombyx_mori	55.3	48.48	51.52

3.2.3 PLANTS

Table-5: Typical values of GC-1% , GC-2% and GC-3% for PLANTS (Several Transcripts)

Species	GC-1%	GC-2%	GC-3%
Cyanidioschyzon_merolae	66.77	52.91	58.12
Cyanidioschyzon_merolae	66.49	53.68	58.38
Cyanidioschyzon_merolae	60.34	48.28	25.86
Cyanidioschyzon_merolae	60.49	46.17	58.6
Cyanidioschyzon_merolae	63.06	48.65	62.84

3.2.4 PROTISTS

Table-6: Typical values of GC-1% , GC-2% and GC-3% for PROTISTS (Several Transcripts)

Species	GC-1%	GC-2%	GC-3%
Albugo_laibachii	51.78	39.69	61.24
Albugo_laibachii	35.35	37.37	47.47
Albugo_laibachii	60.78	39.22	46.08
Albugo_laibachii	57.37	38.42	40
Albugo_laibachii	48.52	38.38	50.34

3.2.5 Ensembl RELEASE

Table-7: Typical values of GC-1%, GC-2%, GC-3% for VERTEBRAss (Several Transcripts)

Species	GC-1%	GC-2%	GC-3%
Anas_platyrhynchos	51.28	39.32	86.75
Anas_platyrhynchos	60.24	51.24	59.81
Anas_platyrhynchos	47.37	34.15	61.2
Anas_platyrhynchos	54.86	33.66	38.91
Anas_platyrhynchos	52.64	42.83	53.68

3.3 Typical TCBSHe values Close to Napier's Constant determined by Overall Summation

TCBSHe values for below Species (specific Transcripts) equal e upto 5 significant decimal places.

3.3.1 FUNGI (9e_Fungi_Napier.sh in <https://github.com/bioinform/GC123e>)

Table-8: TCBSHe values Close to Napier's Constant for certain FUNGAL model Organisms

Species	TCBSHe
Allomyces_macrozynus_atcc_38327	2.71828
Aureobasidium_namibiae_cbs_147_97	2.71828
Bipolaris_sorokiniana_nd90pr	2.71828
Blastomyces_dermatidis_atcc_18188	2.71828
Candida_albicans_12c	2.71828
Candida_albicans_19f	2.71828
Candida_albicans_ca6	2.71828
Candida_albicans_gc75	2.71828
Candida_albicans_l26	2.71828
Candida_albicans_p34048	2.71828
Candida_albicans_p34048	2.71828
Candida_albicans_p37005	2.71828
Candida_albicans_p37037	2.71828
Candida_albicans_p37039	2.71828
Candida_albicans_p57055	2.71828
Candida_albicans_p57072	2.71828
Candida_albicans_p60002	2.71828
Candida_albicans_p75010	2.71828
Candida_albicans_p75016	2.71828
Candida_albicans_p75063	2.71828
Candida_albicans_p76055	2.71828
Candida_albicans_p76067	2.71828
Candida_albicans_p78042	2.71828
Candida_albicans_p78048	2.71828
Candida_albicans_p87	2.71828
Candida_albicans_p94015	2.71828
Candida_albicans_sc5314	2.71828
Candida_albicans_sc5314_gca_000784655	2.71828
Candida_albicans_wo_1	2.71828
Ceraceosorus_bombacis	2.71828
Colletotrichum_incanum	2.71828

Colletotrichum_orbiculare	2.71828
Colletotrichum_simmondsii	2.71828
Coniophora_puteana_rwd_64_598_ss2	2.71828
Coniosporium_apollinis_cbs_100218	2.71828
Coniosporium_apollinis_cbs_100218	2.71828
Debaryomyces_fabryi	2.71828
Dothistroma_septosporum	2.71828
Dothistroma_septosporum	2.71828
Erysiphe_necator	2.71828
Fibulorhizoctonia_sp_cbs_109695	2.71828
Macrophomina_phaseolina_ms6	2.71828
Metarhizium_album_arsef_1941	2.71828
Naumovozyma_dairenensis_cbs_421	2.71828
Naumovozyma_dairenensis_cbs_421	2.71828
Neofusicoccum_parvum_ucrnp2	2.71828
Neurospora_crassa_gca_000786625	2.71828
Neurospora_tetrasperma_fgsc_2508	2.71828
Neurospora_tetrasperma_fgsc_2509	2.71828
Phialophora_attae	2.71828
Pisolithus_microcarpus_441	2.71828
Pseudoloma_neurophilia	2.71828
Purpureocillium_lilacinum	2.71828
Rasamsonia_emersonii_cbs_393_64	2.71828
Rhizophagus_irregularis_daom_197198w	2.71828
Rhodotorula_sp_jg_1b	2.71828
Serpula_lacrymans_var_lacrymans_s7_3	2.71828
Sporothrix_schenckii_1099_18	2.71828
Sporothrix_schenckii_atcc_58251	2.71828
Stachybotrys_chlorohalonata_ibt_40285	2.71828
Tsuchiyaea_wingfieldii_cbs_7118	2.71828
Vanderwaltozyma_polyspora_dsm_70294	2.71828
Vanderwaltozyma_polyspora_dsm_70294	2.71828
Verticillium_longisporum_gca_001268165	2.71828
Verticillium_longisporum	2.71828
Aspergillus_parasiticus_su_1	2.71828
Metarhizium_acridum_cqma_102	2.71828
Neurospora_crassa	2.71828
Verticillium_dahliaejr2	2.71828

3.3.2 METAZOA (9f_Metazoa_Napier.sh in <https://github.com/bioinformers/GC123e>)

Table-9: TCBSHe values Close to Napier's Constant for some METAZOAN model Organisms

Species	TCBSHe
Culex_quinquefasciatus	2.71828
Drosophila_pseudoobscura	2.71828
Ixodes_scapularis	2.71828
Ixodes_scapularis	2.71828
Loa_loa	2.71828

Lottia_gigantea	2.71828
Nasonia_vitripennis	2.71828
Octopus_bimaculoides	2.71828
Pediculus_humanus	2.71828
Solenopsis_invicta	2.71828
Stegodyphus_mimosarum	2.71828
Thelohanellus_kitaei	2.71828
Acyrtosiphon_pisum	2.71828
Acyrtosiphon_pisum	2.71828
Adineta_vaga	2.71828
Amphimedon_queenslandica	2.71828
Amphimedon_queenslandica	2.71828
Anoplophora_glabripennis	2.71828
Atta_cephalotes	2.71828

3.3.3 PLANTS (9g_Plants_Napier.sh in <https://github.com/bioinformer/GC123e>)

Table-10: TCBSHe values Close to Napier's Constant for certain PLANT model Organisms

Species	TCBSHe
Oryza_glumaepatula	2.71828
Oryza_punctata	2.71828
Oryza_rufipogon	2.71828
Selaginella_moellendorffii	2.71828
Selaginella_moellendorffii	2.71828
Selaginella_moellendorffii	2.71828
Aegilops_tauschii	2.71828
Arabidopsis_thaliana	2.71828
Arabidopsis_thaliana	2.71828
Beta_vulgaris	2.71828

3.3.4 PROTISTS (9h_Protists_Napier.sh in <https://github.com/bioinformer/GC123e>)

Table-11: TCBSHe values Close to Napier's Constant for certain PROTIST model Organisms

Species	TCBSHe
Angomonas_deanei	2.71828
Aphanomyces_invadans	2.71828
Capsaspora_owczarzaki_atcc_30864	2.71828
Eimeria_necatrix	2.71828
Leishmania_mexicana_mhom_gt_2001_u1103	2.71828
Leishmania_panamensis	2.71828
Naegleria_gruberi	2.71828
Paramecium_tetraurelia	2.71828
Phytophthora_parasitica_p1976	2.71828
Phytophthora_sojae	2.71828
Pythium_arrhenomanes	2.71828
Pythium_iwayamai	2.71828

Trichomonas_vaginalis_g3	2.71828
Trichomonas_vaginalis_g3	2.71828
Trypanosoma_rangeli_sc58	2.71828
Hyaloperonospora_arabidopsidis	2.71828
Leptomonas_seymouri	2.71828
Saprolegnia_diclina_vs20	2.71828

3.3.5 Ensembl RELEASE (9i_Vertebrata_Napier.sh in <https://github.com/bioinformer/GC123e>)

Table-12: TCBSHe values ~ Napier's Constant for a few VERTEBRATE model Organisms

Species	TCBSHe
Callithrix_jacchus	2.71828
Felis_catus	2.71828
Homo_sapiens	2.71828
Mus_musculus_129s1svimj	2.71828
Mus_musculus_lpj	2.71828
Mus_musculus_c57bl6nj	2.71828
Mus_musculus_nzohltj	2.71828
Mustela_putorius_furo	2.71828
Myotis_lucifugus	2.71828
Oryctolagus_cuniculus	2.71828
Otolemur_garnettii	2.71828
Sus_scrofa	2.71828
Takifugu_rubripes	2.71828
Tarsius_syrichta	2.71828

3.4 Harmonic Mean (HM) of TCBSHe values computed using R-studio (R version 4.0.4)

All of the 5 Clades: Fungi, Metazoa, Plants, Protists and Ensembl release, are given in Table-13.

Table-13: Cardinality of Total Species, Transcripts and TCBSHe-HM across 5 Clades

Clade	Total Species	Total Transcripts	HM of TCBSHe	Deviation from e
Fungi	735	6783327	2.873189	+0.154907
Metazoa	68	1364948	2.836593	+0.118311
Plants	44	1917118	2.821708	+0.103426
Protists	186	2024435	2.716689	-0.00159283
Vertebrates	85	2360029	2.847058	+0.128776

From Table-13 above, we observe that interestingly, Plants < Metazoa < Vertebrates < Fungi expressed in increasing Order of deviation from Ideal value, $e = 2.71828...$ which show Positive deviation, whereas only Protists, in form of exception, exhibits Negative deviation from e , although specific Transcripts across several Species in each Clade yield TCBSHe --> e as per 3.5 below .

3.5 HM across several Transcripts grouped by individual Species within a Clade

It is observed that HM of TCBSHe for Transcripts in Some species tend to e = Napier constant.

Table-14: Typical HM of TCBSHe for 6,783,327 annotated Transcripts vs. 735 Fungal Species

Fungal-Species	TCBSHe-HM grouped by Species
<i>_candida_auris</i>	2.90367107889668
<i>_candida_glabrata</i>	2.86790562410742
<i>_candida_glabrata_gca_001466525</i>	2.86342352430567
<i>_candida_glabrata_gca_001466535</i>	2.8634437132242
<i>_candida_glabrata_gca_001466565</i>	2.86492410540444
<i>_candida_glabrata_gca_001466575</i>	2.86330430501295
<i>_candida_glabrata_gca_001466635</i>	2.8640227199129
<i>_candida_glabrata_gca_001466685</i>	2.8650522492127
<i>_candida_tenuis_atcc_10573</i>	2.89738848276511
<i>Absidia_glauca</i>	2.95044505959613

Additional rows including Above entries are collated in Supplementary table **fg735.csv**

Table-15: Typical HM of TCBSHe for 1,364,948 annotated Transcripts vs. 68 Metazoa Spp.

Metazoan-Species	TCBSHe-HM grouped by Species
<i>Acyrtosiphon_pisum</i>	2.78209642876382
<i>Adineta_vaga</i>	2.57537585742343
<i>Aedes_aegypti</i>	2.89702417547141
<i>Amphimedon_queenslandica</i>	2.83080298500206
<i>Anopheles_darlingi</i>	2.86523081876304
<i>Anopheles_gambiae</i>	2.77878890789813
<i>Anoplophora_glabripennis</i>	2.84509923863477
<i>Apis_mellifera</i>	2.70957019274186
<i>Atta_cephalotes</i>	2.8727969728819
<i>Belgica_antarctica</i>	2.90982144075001

Additional rows including Above entries are collated in Supplementary table **mz068.csv**

Table-16: Typical HM of TCBSHe for 1,917,118 annotated Transcripts vs. 44 Plant Varieties

Plant Varieties	TCBSHe-HM grouped by Species
<i>Aegilops_tauschii</i>	2.76050540892975
<i>Amborella_trichopoda</i>	2.933548594834
<i>Arabidopsis_lyrata</i>	2.92278885937872
<i>Arabidopsis_thaliana</i>	2.92624852052254
<i>Beta_vulgaris</i>	2.89839899917588
<i>Brachypodium_distachyon</i>	2.75666834032552
<i>Brassica_napus</i>	2.93309716247077
<i>Brassica_oleracea</i>	2.93554340658407
<i>Brassica_rapa</i>	2.94132888109909
<i>Chlamydomonas_reinhardtii</i>	2.51621134866737

Additional rows including Above entries are collated in Supplementary table **nt044.csv**

Table-17: Typical HM of TCBSHE for 2,024,435 annotated Transcripts vs. 186 Protist Types

Species	TCBSHE-HM grouped by Species
Acanthamoeba_castellanii_str_neff	2.59413442401166
Albugo_candida	2.9340816207376
Albugo_laibachii	2.93870711814763
Angomonas_deanei	2.8109408672605
Aphanomyces_astaci	2.82672569741714
Aphanomyces_invadans	2.83724248069306
Aureococcus_anophagefferens	2.12432344507822
Babesia_bigemina	2.82880558827141
Babesia_bovis	2.91233932787662
Babesia_microti_strain_ri	2.84076785193199

Additional rows including Above entries are collated in Supplementary table **pr186.csv**

Table-18: Typical HM of TCBSHE for 2,360,029 annotated Transcripts vs. 85 Vertebrates

Species	TCBSHE-HM grouped by Species
Ailuropoda_melanoleuca	2.80840566575861
Anas_platyrhynchos	2.79699229594752
Anolis_carolinensis	2.86242541018576
Astyanax_mexicanus	2.88849817686598
Bos_taurus	2.80398218424545
Caenorhabditis_elegans	2.88799317522863
Callithrix_jacchus	2.84028752411799
Canis_familiaris	2.79700604479059
Cavia_porcellus	2.82656385779042

Additional rows including Above entries are collated in Supplementary table **vb085.csv**

4. DISCUSSION

In the present work, we have essentially considered the Genetic codon as a Triplet-block, Zero-memory source and computed the TCBSHE values (Triplet-Codon Block Shannon Entropy) for each species belonging to 5 clades across several Transcripts by taking into account Eight 3-tuple Binary combinations of GC/AT taking into account all 3 codonic positions, assuming AT = (1-GC). Also, we have considered the Harmonic Mean of TCBSHE values as a true indicator of Entropy lower-bounds, also keeping in mind the following significance of “HM of Entropy values”, by definition.

It may be noted that whenever we consider Harmonic Mean for a set of Entropy values, such a HM can be expressed in terms of pre-defined crucial quantities, namely = ratio of Mutual Information

to (1 – normalized Variation of Information), as per the simple and elegant proof in non-trivial case of 2 Entropy values $H(X)$ and $H(Y)$ below. The proof below can be naturally extended to ≥ 3 Cardinality of Entropy values by applying the principle of Mathematical Induction.

By Definition, Mutual Information of 2 entropies $H(X)$ and $H(Y)$,

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

Normalized version of $I(X;Y)$ in terms of Harmonic-Mean (HM) of entropies,

$$I_{HM} = \frac{H(X) + H(Y) - H(X,Y)}{\frac{2 * H(X) * H(Y)}{H(X) + H(Y)}}$$

By definition, “Normalized” Variation of Information [4]

$$VI_{norm}(X, Y) = 0.5 * \frac{H(X|Y)}{H(X)} + 0.5 * \frac{H(Y|X)}{H(Y)}$$

Simplifying above 2 equations yields, using the Relation, $H(Y|X) = H(X,Y) - H(X)$

HM of 2 Entropies $H(X)$ & $H(Y)$ (can be Generalized for Several Entropies, whose Harmonic-Mean is being computed) is equivalent to

$$\frac{I(X;Y)}{1 - VI_{norm}(X, Y)}$$

5. CONCLUSIONS AND FURTHER WORK

Wherever the HM of TCBSHe values tends close to e = Napier’s constant – as in the case of specific transcripts across several Model organisms, we are done, in so far as the Hypothesis purported in Title of this paper has been concretely validated using real-world datasets pertaining to GC-1%, GC-2% and GC-3%. This is just a special case when TCRBE (Triplet Codon Rényi Block Entropy) order approaches 1. For other species or transcripts “non-compliant” with TCBSHe $\rightarrow e$, our natural detour is then to determine non-negative, non-unity, non-trivial Rényi entropy orders in order to mathematically predict generalized fractal dimensions of “CODOMES”. In this context, we define CODOME as that which codes for 100% Universal Proteome – for a particular instance of Translation as per the Central Dogma. We may also epitomize our notion of the “Codome” as a Singular Representative-compositional Abstraction of a given CDS: CoDing Sequence, which consists of multiple ‘Codons,’ occurring as Triplets. It is named in line with Genome, Proteome, Lipidome, Metabolome, Kinome, Degradome, Transcriptome, and other such -Omes.

For instance, it has been found that the Real number solution for Rényi entropy order, $x = 18.6726$ so that TCRBE = e (Napier’s constant), in the case of Transcript **KYG41716** for fungus *Acidomyces richmondensis*, as per the below equation, wherein the TCBSHe deviant from e is **2.97145**

solve	$\frac{1}{1-x} \log_2(0.156264^x + 0.128163^x + 0.165135^x + 0.135438^x + 0.110854^x + 0.0909189^x + 0.117147^x + 0.0960801^x) = e$
-------	---

Figure-2: Computing Rényi entropy order for a Typical Fungal Transcript, with TCRBE = e

Hence, one logically acceptable further work in this direction that calls for immediate attention is towards ascertaining individual Rényi entropy orders (α) adjusted such that $RHS = e$ so as to obtain confirmatory results in accordance with Big Data Genetic Coding [5] as outlined above. Certainly, this would demand much time and huge effort, being computationally-intensive, spanning in total ~14,449,857 transcripts grouped by 1118 Species culled from the GC.evoBase.

Author Contributions

First author has contributed to the Original ideation, drafting the Manuscript, proof-reading and also Source-coding of the entire Data Analysis pipelines. Other authors have provided their valuable Comments and Suggestions for improvement, able Guidance and Expert Support from time-to-time.

Acknowledgements

We would like to thank ICAR: Indian Council of Agricultural Research for facilitating this Work, specifically under the Aegis of NADCP: National Animal Disease Control Program. Thanks are due to Director, NIVEDI for his patronage and encouragement.

Conflict of Interest

To the best of their knowledge and belief, All the Authors hereby declare that there are absolutely NO competing interests that exist amongst each other, as far as this Original Research is concerned.

REFERENCES:

- [1] Praharshit Sharma, & Kuralayanapalya Puttahonnappa Suresh. (2021). Source-Code: Triplet Codon Block Shannon Entropy (TCBShE) in terms of GC(1,2,3)% equates to Napier's Constant for Model Organisms, and Harmonically Averages to same approximately: a Penta-Clado-genic Quantitative Survey across ~14.45 million Transcripts Clustered by 1118 Species (v-NAPIERconst). Zenodo. <https://doi.org/10.5281/zenodo.5179552>
- [2] Wang D. (2018). GCevobase: an evolution-based database for GC content in eukaryotic genomes. *Bioinformatics (Oxford, England)*, 34(12), 2129–2131. <https://doi.org/10.1093/bioinformatics/bty068>
- [3] B V L S Prasad and Mohan C Vemuri (September, 1998). Genome analysis for nucleotide interactions in fully sequenced genomes of selective prokaryotes. *Journal of Biosciences*. Volume 23 (Issue 3) pp: 255-263. <https://www.ias.ac.in/article/fulltext/jbsc/023/03/0255-0263>
- [4] Appendix-B. Andrea Lancichinetti et al 2009 New J. Phys. 11 033015Andrea Lancichinetti et al 2009 New J. Phys. 11 033015
- [5] <https://www.itsoc.org/profile/9590>
- [6] **Background Paper**= Nigatu, D., Henkel, W., Sobetzko, P. *et al.* Relationship between digital information and thermodynamic stability in bacterial genomes. *J Bioinform Sys Biology* 2016, 4 (2016). <https://doi.org/10.1186/s13637-016-0037-x>