from the pillar fell in the direction of *Anantaśāyin* Vishnu in cave 13. Moreover, on the important summer solstice day at Udayagiri, the motion of the sun across the sky was in line with the specially cut passageway and the evening setting sun completely illuminated the *Anantaśā yin* Vishnu panel in cave 13. The image that was probably atop the Delhi iron pillar capital has been deduced to be disc-shaped, approximately 20″ in diameter and 2″ thick. The nature of the top surface of the iron pillar capital indicated that the disc-shaped object was fitted vertically on a flat, circular base, which was subsequently welded onto the top of the cylinder, around which the components of the decorative bell capital were shrunk fit.

1. Anantharaman, T. R., *The Rustless Wonder – A Study of the Delhi Iron Pillar*, Vigyan Prasar, New Delhi, India, 1997.
2. Balasubramaniam, R., *Delhi Iron Pillar: New Insights*, Indian Institute of Advanced Study, Shimla, 2002, pp. 8–46.
3. Balasubramaniam, R., Identity of *Chandra* and *Vishnupadagiri* of the Delhi iron pillar inscription: numismatic, archaeological and literary evidence. *Bull. Met. Mus.*, 2000, **32**, 42–64.
4. Dass, M. I., *Udayagiri – Rise of a sacred hill, its art, architecture and landscale – A study*. Ph D thesis, DeMontfort University, Leicestershire, UK, 2001, pp. 105–155.
5. Dass, M. I. and Balasubramaniam, R., Estimation of the original erection site of the Delhi iron pillar at Udayagiri. *Indian J. Hist. Sci.*, 2004, **39.1**, 51–74.
6. Willis, M., Inscriptions at Udayagiri: Locating domains of devotion, patronage and power in the eleventh century. *South Asian Stud.*, 2001, **17**, 48.
7. Lahiri, N., *Archaeology of Ancient Indian Trade Routes upto c. 300 BC*, Oxford University Press, New Delhi, India, 1992, pp. 368–381.
8. Dass, M. I. and Willis, M., The lion capital from Udayagiri and the antiquity of sun worship in central India. *South Asian Stud.*, 2002, **18**, 25–45.
9. Gangooly, P. (ed.), *Surya Siddhanta, Translated with Notes and Appendix by Rev. Ebenezer Burgess (1880)*, Motilal Banarsidass, New Delhi, 1960.
10. Thibaut, G., Contribution to the explanation of the *Jyotisa-Vedanga*. *J. Asiat. Soc. Bengal*, 1977, **46**, 411–437.
11. Krishna Deva, Gupta and their feudatories. In *Encyclopaedia of Indian Temple Architecture, North India: Foundations of North Indian Style c 250 BC–1100 AD* (eds Meister, M. W., Dhaky M. A. and Krishna Deva), American Institute of Indian Studies, Princeton University Press, 1988, vol. II, part I, pp. 18–57.
12. Patil, D. R., *The Monuments of the Udayagiri Hill*, Gwalior, 1948, pp. 377–428.
13. Mitra, D., Varāha cave at Udayagiri – an iconographic study. *J. Asiat. Soc.*, 1963, **5**, 99–103.
14. Williams, J., *The Art of Gupta India: Empire and Province*, Princeton University Press, Princeton, USA, 1982, pp. 37–49.
15. Fleet, J. F., Inscriptions of the early Gupta kings and their successors. *Corpus Inscriptionum Indicarum*, 1888, vol. III, p. 25.
16. Majumdar, R. C., In *The History and Culture of the Indian People – Classical Age*, Bharatiya Vidya Bhavan, Mumbai, 1954, vol. 3, pp. 321–323.
17. Balasubramaniam, R., Dass, M. I. and Raven, E. M., On the original image atop the Delhi iron pillar. *Indian J. Hist. Sci.*, 2004, **39.2**, in press.
18. Balasubramaniam, R., Decorative bell capital of the Delhi iron pillar. *J. Met.*, 1998, **50**, 40–47.
19. Balasubramaniam, R., New insights on the corrosion of the Delhi iron pillar based on historical and dimensional analysis. *Curr. Sci.*, 1997, **73**, 1057–1067.
20. Agrawal, V. S., *Chakra-dhwaja: The Wheel Flag of India*, Prithvi Prakashan, Varanasi, 1964, pp. 1–59.

# Shannon's uncertainty principle and gene expression levels

## V. Subramaniyam[†], S. K. Gupta and T. C. Ghosh*

Bioinformatics Centre, Bose Institute, P/12, CIT Scheme VII M, Kolkata 700 054, India
[†]Present address: Indian Institute of Information Technology, Allahabad 211 012, India

**Shannon's uncertainty principle has been applied to measure the degree of constraints in codon bias in the coding sequences of *Escherichia coli*, *Saccharomyces cerevisiae* and *Haemophilus influenzae*. Our study shows a high degree of correlation between Shannon's uncertainty values and codon adaptation index (CAI). This result suggests that the degree of constraints determined by Shannon's uncertainty principle reflects the level of gene expression and we propose that Shannon's uncertainty principle can be used as an alternative method for predicting the level of gene expression. The main advantage of Shannon's uncertainty values over CAI is that for calculating Shannon's uncertainty values, one does not require any reference set of genes as required in CAI. This gives a potential use of Shannon's uncertainty principle over CAI in predicting the level of gene expression, specially for the newly sequenced genomes where genes are not properly annotated.**

THE existence of non-random uses of synonymous codons is well documented. Moreover, codon usage patterns differ significantly among different genes within the same taxon[1]. It has been widely accepted that compositional biases are the only dictators in shaping the codon usage variation among the genes in the extremely AT or GC-rich unicellular organisms[2–4]. It has been suggested that translational selection determines the codon usage bias of highly expressed genes and subsequently, it has been advocated that preferred codons in highly expressed genes are recognized by most abundant tRNAs[5–7]. Recently, it was

---

*For correspondence. (e-mail: tapash@bic.boseinst.ernet.in)

observed that in *Pseudomonas aeruginosa*, codon usage bias is mainly dictated by translational selection rather than mutational biases, though it is an extremely GC-rich organism[8]. In some unicellular organisms, it was observed that both translational and compositional constraints are operational in dictating the codon usage variation among the genes in those organisms[5,9–11]. In *Borrelia burgdorferi*, it was observed that replicational–transcriptional selection is responsible for the codon usage variation among the genes[12]. Recently, it has been reported that the cellular as well as the physical location of the gene products can also reflect the codon usage patterns[13,14]. In *Mycobacteria*, it has been reported that codon usage bias was mainly dictated by hydrophobicity of each gene[15]. Codon usage was found to be affected by the base composition of the neighbouring sites[16]. Possible relationships between synonymous codon usage and protein secondary structures have also been reported in the literature[17–20].

There may be other unknown constraints coming into play in selecting the codon usage variation in different organisms and availability of large number of complete genomes opens up a tremendous opportunity to reveal those unknown factors. Several mathematical indices have been proposed to estimate the degree of codon bias and until now, the best method to capture gene expression using translational bias is codon adaptation index (CAI), as described by Sharp and Li[21]. CAI takes into consideration that the highly expressed genes use optimal codons more frequently than other genes in the genome and computes a weight for each codon and combines them to define the CAI value for each gene in the genome.

Although CAI has been widely used in predicting highly expressed genes, the interpretation of the results requires much more care. In species where translational selection is either absent or ineffective and mutational bias is more pronounced, the CAI values for individual genes are related to its base composition rather than its expression level. The other drawback of CAI is that it requires prior knowledge of optimal codons or a reference set of highly expressed genes in the species.

It has become necessary to devise new methods, so that one can predict the level of gene expression without any prior knowledge of optimal codons in the highly expressed genes. With this aim, we have used Shannon's uncertainty principle to predict the level of gene expression. Shannon's uncertainty value in bits per codon, is used to know the degree of constraint present in the usage of codons while coding the sequence[22–24].

The complete genomes of *Escherichia coli*, *Saccharomyces cerevisiae* and *Haemophilus influenzae* have been downloaded from ftp.ncbi.nlm.nih.gov/genbank/genomes. Our own program developed in C was used to retrieve the coding sequences from the complete genomes. To minimize sampling errors, we have chosen only those sequences that are greater than or equal to 300 bp and have correct initial and termination codons. Finally, 3892, 4725 and 1568 genes were selected for *E. coli*, *S. cervisiae* and *H. influenzae* respectively.

Shannon's uncertainty values (*H*) were computed using

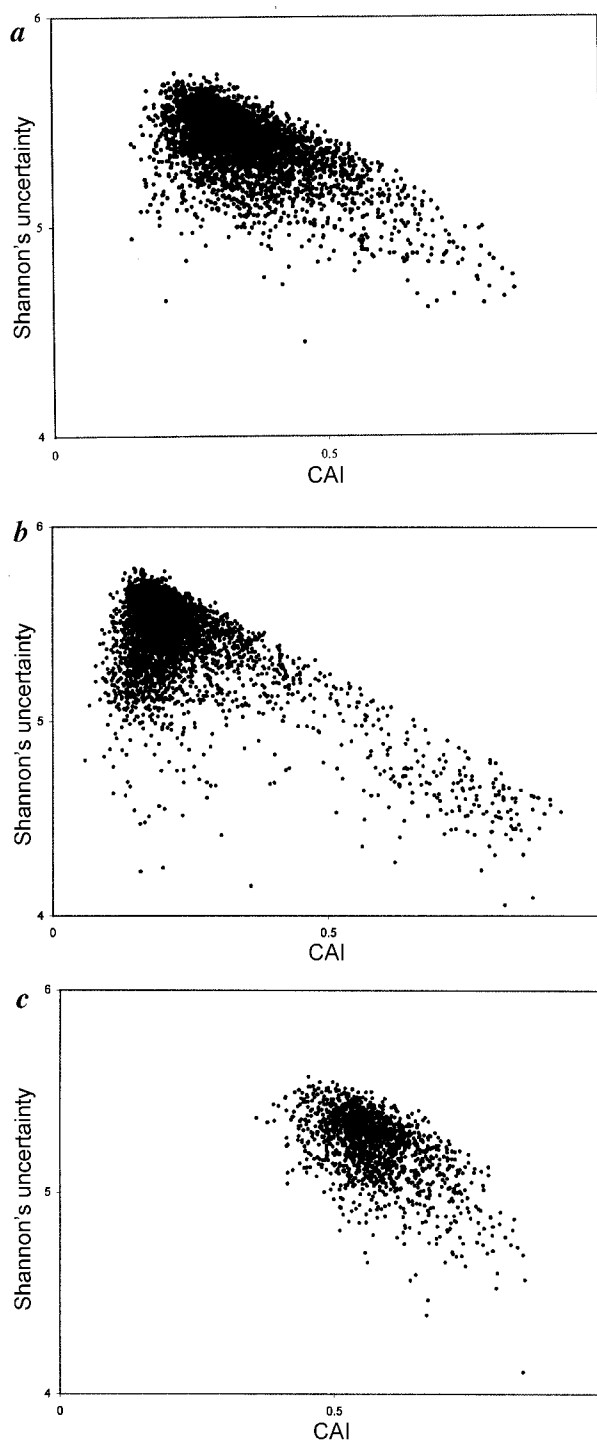$$H = \left( -\sum_{i=1}^{i=61} p_i \log_2 P_i \right), \qquad (1)$$



**Figure 1.** Scatter plot between uncertainty and CAI values of (*a*) *E. coli*, (*b*) *S. cerevisiae* and (*c*) *H. influenzae* genes.

where $p_i$ is the probability of a codon present in the given gene and $i$ runs from 1 to 61. When a codon is found to be absent for a gene, we introduced a Laplace correction of $1.0 \times 10^{-6}$ for $p_i$. We have excluded three termination codons in our analysis. $H$ gives the uncertainty or degree of constraint in bits per codon present in a given gene in accordance with their codon usage[22]. In a gene, when all

the codons are used in equal probability then according to the eq. (1), the uncertainty value will be 5.93 bits per codon and any value less than 5.93 bits per codon implies that certain degrees of constraint are present in that gene.

CAI was calculated for all the genes in the study according to the method adopted by Sharp and Li[21]. The reference subset to calculate CAI for *E. coli* and *S. cere-*

**Table 1.** Uncertainty and CAI values of 50 genes located at the extreme end in the uncertainty scale of lowest uncertainty values

| | E. coli | | | S. cerevisiae | | | H. influenzae | |
|---|---|---|---|---|---|---|---|---|
| Gene | Uncertainty in bits per codon | CAI | Gene | Uncertainty in bits per codon | CAI | Gene | Uncertainty in bits per codon | CAI |
| rplL | 3.99 | 0.85 | RPP2A | 4.06 | 0.82 | rpL7/L12 | 4.11 | 0.85 |
| asr | 4.45 | 0.46 | CCW12 | 4.09 | 0.88 | Outer membrane protein | 4.53 | 0.80 |
| rplT | 4.61 | 0.69 | RPL42B | 4.23 | 0.78 | rpL9 | 4.57 | 0.85 |
| rpsI | 4.63 | 0.79 | RPL36B | 4.31 | 0.81 | rpL22 | 4.60 | 0.80 |
| rplO | 4.64 | 0.70 | RPS12 | 4.31 | 0.86 | rpL15 | 4.63 | 0.74 |
| ompC | 4.66 | 0.83 | FIT3 | 4.36 | 0.56 | rpL20 | 4.68 | 0.73 |
| rplI | 4.68 | 0.73 | RPS26A | 4.36 | 0.80 | rpL21 | 4.68 | 0.74 |
| rplW | 4.68 | 0.67 | RPS20 | 4.39 | 0.84 | rpL1 | 4.69 | 0.85 |
| eno | 4.70 | 0.84 | RPL30 | 4.39 | 0.87 | bolA | 4.69 | 0.72 |
| mopA | 4.71 | 0.80 | TIR1 | 4.40 | 0.63 | rrf | 4.70 | 0.78 |
| glnB | 4.72 | 0.42 | RPL24B | 4.42 | 0.77 | atpF | 4.70 | 0.72 |
| tsf | 4.74 | 0.78 | RPS24A | 4.42 | 0.83 | rpL19 | 4.71 | 0.79 |
| rplA | 4.76 | 0.78 | RPS26B | 4.42 | 0.71 | pal | 4.73 | 0.84 |
| gapA | 4.77 | 0.84 | RPS23B | 4.43 | 0.80 | tsf | 4.74 | 0.83 |
| tufA | 4.79 | 0.82 | RPL24A | 4.43 | 0.76 | rpL11 | 4.74 | 0.78 |
| rpsA | 4.81 | 0.78 | RPS10B | 4.43 | 0.80 | hns | 4.75 | 0.69 |
| tig | 4.82 | 0.74 | RPS6B | 4.44 | 0.85 | secG | 4.75 | 0.68 |
| yfiD | 4.82 | 0.71 | RPS25A | 4.44 | 0.73 | iscU | 4.75 | 0.76 |
| slyD | 4.83 | 0.69 | HHF1 | 4.45 | 0.74 | rpL24 | 4.77 | 0.69 |
| ompX | 4.84 | 0.75 | RPS10A | 4.45 | 0.84 | rplQ | 4.78 | 0.73 |
| rplJ | 4.84 | 0.65 | RPS6A | 4.45 | 0.84 | hupA | 4.78 | 0.70 |
| ahpC | 4.84 | 0.81 | RPL22A | 4.45 | 0.89 | rpL13 | 4.79 | 0.79 |
| hlpA | 4.84 | 0.64 | RPS9B | 4.46 | 0.84 | rpS9 | 4.80 | 0.77 |
| rplU | 4.85 | 0.65 | RPL8B | 4.47 | 0.85 | secE | 4.80 | 0.63 |
| rplX | 4.85 | 0.59 | RPS14A | 4.48 | 0.80 | rsgA | 4.81 | 0.71 |
| rplK | 4.86 | 0.72 | HSP12 | 4.49 | 0.64 | rpS1 | 4.81 | 0.83 |
| tufB | 4.86 | 0.80 | TIR2 | 4.49 | 0.57 | rps12 | 4.82 | 0.72 |
| rplS | 4.86 | 0.65 | RPL8A | 4.50 | 0.84 | trxM | 4.82 | 0.78 |
| rpsF | 4.86 | 0.69 | RPL18A | 4.50 | 0.82 | rpL18 | 4.83 | 0.69 |
| rpsJ | 4.86 | 0.58 | RPL17A | 4.50 | 0.81 | rpL3 | 4.83 | 0.79 |
| gcvH | 4.86 | 0.55 | RPL13B | 4.51 | 0.75 | mscL | 4.84 | 0.63 |
| yaiU | 4.87 | 0.72 | RPS17A | 4.51 | 0.81 | eno | 4.84 | 0.81 |
| spy | 4.87 | 0.66 | RPL36A | 4.51 | 0.68 | nusG | 4.85 | 0.72 |
| rpsL | 4.87 | 0.67 | RPL25 | 4.52 | 0.76 | rpS7 | 4.85 | 0.77 |
| crr | 4.87 | 0.62 | SRP40 | 4.52 | 0.24 | recX | 4.85 | 0.66 |
| rplP | 4.87 | 0.62 | RPL19B | 4.52 | 0.72 | greA | 4.86 | 0.67 |
| rplC | 4.87 | 0.72 | RPL1B | 4.52 | 0.85 | rpL10 | 4.86 | 0.73 |
| frdD | 4.88 | 0.51 | RPL34B | 4.52 | 0.79 | fic | 4.86 | 0.72 |
| hns | 4.88 | 0.60 | ENO2 | 4.52 | 0.90 | gapdH | 4.87 | 0.83 |
| csgA | 4.89 | 0.57 | TIR3 | 4.53 | 0.52 | rpS11 | 4.87 | 0.75 |
| tpiA | 4.89 | 0.75 | TDH3 | 4.53 | 0.93 | glnB | 4.88 | 0.58 |
| pal | 4.89 | 0.69 | RPS24B | 4.53 | 0.76 | hlpA | 4.88 | 0.79 |
| tolA | 4.89 | 0.40 | RPL32 | 4.53 | 0.83 | Opacity protein | 4.88 | 0.63 |
| rpsC | 4.89 | 0.74 | RPS31 | 4.53 | 0.81 | tig | 4.88 | 0.79 |
| rpsB | 4.90 | 0.78 | RPL15A | 4.53 | 0.79 | tolR | 4.88 | 0.60 |
| rplV | 4.90 | 0.56 | RPS23A | 4.54 | 0.73 | FkbP-type peptidyl-prolyl cis-trans isomerase | 4.88 | 0.69 |
| nirD | 4.90 | 0.49 | STM1 | 4.54 | 0.74 | slyD | 4.89 | 0.72 |
| rplM | 4.90 | 0.68 | RPL19A | 4.54 | 0.71 | rpS14 | 4.89 | 0.68 |
| adk | 4.91 | 0.65 | RPL1A | 4.54 | 0.84 | rpS5 | 4.90 | 0.73 |
| pgk | 4.92 | 0.73 | ENO1 | 4.55 | 0.88 | exbD | 4.90 | 0.63 |

**Table 2.** Uncertainty and CAI values of 50 genes located at the extreme end in the uncertainty scale of highest uncertainty values

| E. coli | | | S. cerevisiae | | | H. influenzae | | |
|---|---|---|---|---|---|---|---|---|
| Gene | Uncertainty | CAI | Gene | Uncertainty | CAI | Gene | Uncertainty | CAI |
| pqiA | 5.60 | 0.27 | MAL13 | 5.69 | 0.17 | hrpa | 5.47 | 0.51 |
| idnD | 5.60 | 0.31 | BUB2 | 5.69 | 0.16 | Predicted serine protease | 5.47 | 0.54 |
| ilvI | 5.61 | 0.30 | MIP1 | 5.69 | 0.17 | sdaA | 5.47 | 0.55 |
| plsX | 5.61 | 0.23 | RIB2 | 5.69 | 0.20 | comM | 5.47 | 0.47 |
| sgcR | 5.61 | 0.29 | MHP1 | 5.69 | 0.21 | mepA | 5.47 | 0.52 |
| yicK | 5.61 | 0.25 | BNA2 | 5.69 | 0.19 | Heme-hemopexin utilization protein C | 5.47 | 0.54 |
| wecB | 5.61 | 0.32 | AQR2 | 5.69 | 0.19 | Biotin synthesis protein | 5.47 | 0.45 |
| rseB | 5.61 | 0.29 | BOP1 | 5.69 | 0.21 | Integrase/recombinase | 5.47 | 0.41 |
| NadB | 5.61 | 0.31 | AKR2 | 5.70 | 0.17 | rbsR | 5.47 | 0.48 |
| FucK | 5.61 | 0.30 | RGT1 | 5.70 | 0.17 | nifR3 | 5.48 | 0.53 |
| PyrC | 5.62 | 0.37 | AYT1 | 5.70 | 0.20 | topB | 5.48 | 0.51 |
| FdhD | 5.62 | 0.29 | TAH11 | 5.70 | 0.15 | dnaE | 5.48 | 0.53 |
| PhrB | 5.62 | 0.29 | THI2 | 5.70 | 0.17 | Conserved hypothetical protein | 5.48 | 0.45 |
| DacB | 5.62 | 0.28 | GPI16 | 5.70 | 0.21 | truB | 5.49 | 0.47 |
| rpiR | 5.62 | 0.27 | UBP7 | 5.70 | 0.19 | rec2 | 5.49 | 0.46 |
| MalY | 5.62 | 0.35 | SNG1 | 5.70 | 0.19 | emrB | 5.49 | 0.52 |
| intA | 5.62 | 0.30 | FRE6 | 5.70 | 0.17 | spoT | 5.49 | 0.53 |
| MolR_1 | 5.62 | 0.27 | BUD7 | 5.70 | 0.18 | Conserved hypothetical transmembrane protein | 5.49 | 0.53 |
| YdaU | 5.62 | 0.23 | DPH2 | 5.70 | 0.18 | modC | 5.49 | 0.46 |
| iap | 5.62 | 0.26 | PTK1 | 5.70 | 0.18 | Predicted Zn-dependent protease | 5.49 | 0.52 |
| hyfI | 5.62 | 0.26 | DAL1 | 5.70 | 0.19 | rffG | 5.49 | 0.52 |
| MolR_2 | 5.62 | 0.25 | MMP1 | 5.71 | 0.19 | HI1500 | 5.49 | 0.44 |
| rcsC | 5.62 | 0.29 | CDC14 | 5.71 | 0.18 | hisB | 5.49 | 0.48 |
| AcnA | 5.62 | 0.34 | HST3 | 5.71 | 0.19 | trmD | 5.49 | 0.43 |
| ThiH | 5.63 | 0.30 | CRR1 | 5.71 | 0.17 | menC | 5.49 | 0.52 |
| rfaQ | 5.63 | 0.26 | UBP5 | 5.71 | 0.19 | argH | 5.50 | 0.56 |
| ampC | 5.63 | 0.29 | PTR3 | 5.71 | 0.18 | cyaA | 5.50 | 0.55 |
| intE | 5.63 | 0.24 | SGA1 | 5.71 | 0.18 | HI0129 | 5.50 | 0.49 |
| PabC | 5.63 | 0.25 | TOP3 | 5.72 | 0.18 | purL | 5.50 | 0.57 |
| DnaG | 5.64 | 0.29 | ISC1 | 5.72 | 0.20 | nifS protein | 5.50 | 0.49 |
| smf | 5.64 | 0.20 | MET30 | 5.72 | 0.20 | ung | 5.50 | 0.51 |
| nac | 5.64 | 0.24 | ZMS1 | 5.72 | 0.19 | HI1410 | 5.50 | 0.49 |
| fimI | 5.64 | 0.21 | SMF2 | 5.72 | 0.15 | hsdR | 5.50 | 0.47 |
| FimD | 5.64 | 0.25 | NDC1 | 5.72 | 0.17 | glgA | 5.51 | 0.51 |
| HofB | 5.65 | 0.27 | GAL4 | 5.72 | 0.19 | muL | 5.51 | 0.47 |
| fecE | 5.65 | 0.23 | LSB6 | 5.73 | 0.18 | Integrase | 5.51 | 0.44 |
| intB | 5.65 | 0.24 | SYG1 | 5.73 | 0.21 | Conserved hypothetical protein | 5.51 | 0.49 |
| amn | 5.66 | 0.30 | CAC2 | 5.73 | 0.13 | glgX | 5.51 | 0.54 |
| lhr | 5.66 | 0.26 | PDR10 | 5.73 | 0.19 | glgB | 5.51 | 0.53 |
| AppA | 5.66 | 0.27 | PHM7 | 5.73 | 0.21 | Potassium/copper-transporting ATPase | 5.51 | 0.50 |
| umuC | 5.66 | 0.25 | RAX1 | 5.73 | 0.16 | Conserved hypothetical protein | 5.52 | 0.52 |
| intF | 5.66 | 0.26 | GFD2 | 5.73 | 0.16 | dprA | 5.52 | 0.49 |
| DsdC | 5.66 | 0.27 | STB5 | 5.74 | 0.20 | Conserved hypothetical protein | 5.52 | 0.47 |
| CadC | 5.66 | 0.27 | GUT1 | 5.74 | 0.22 | priA | 5.52 | 0.49 |
| PhoH | 5.66 | 0.27 | TDP1 | 5.74 | 0.17 | bioF | 5.52 | 0.46 |
| fes | 5.67 | 0.28 | FMT1 | 5.74 | 0.14 | Conserved hypothetical protein | 5.52 | 0.45 |
| EvgS | 5.67 | 0.24 | CVT17 | 5.74 | 0.18 | Transport protein | 5.53 | 0.46 |
| tra8_1 | 5.68 | 0.21 | RDR1 | 5.76 | 0.16 | Conserved hypothetical protein | 5.53 | 0.51 |
| HipA | 5.68 | 0.25 | PUS2 | 5.76 | 0.13 | hisC | 5.55 | 0.48 |
| SgcX | 5.69 | 0.29 | TRK2 | 5.77 | 0.21 | HI1522 | 5.55 | 0.50 |

visiae was taken from Sharp and Li[21] and Velculescu *et al.*[25] respectively. Due to non-availability of information regarding experimentally known highly expressed genes in *H. influenzae*, we have used all the ribosomal proteins present in the genome as a reference subset to calculate CAI.

Shannon's uncertainty values in bits per codon as well as their corresponding CAI values were calculated for *E. coli*, *S. cerevisiae* and *H. influenzae* genes. Figure 1 *a–c* shows the scattered plots between uncertainty and CAI values for the genes of *E. coli*, *S. cerevisiae* and *H. influenzae* respectively. Pearson correlation coefficients at 0.01

**Table 3.** Uncertainty and CAI values for ribosomal proteins of *E. coli*, *S. cerevisiae* and *H. influenzae*

| | *S. cerevisiae* | | | *E. coli* | | | *H. influenzae* | |
| Gene | Uncertainty in bits per codon | CAI | Gene | Uncertainty in bits per codon | CAI | Gene | Uncertainty in bits per codon | CAI |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| RPP2A | 4.06 | 0.82 | rpIL | 3.99 | 0.85 | rpL7/L12 | 4.11 | 0.85 |
| RPL42B | 4.23 | 0.78 | rpIT | 4.61 | 0.69 | rpL9 | 4.57 | 0.85 |
| RPL36B | 4.31 | 0.81 | rpsI | 4.63 | 0.79 | rpL22 | 4.60 | 0.80 |
| RPS12 | 4.31 | 0.86 | rpIO | 4.64 | 0.70 | rpL15 | 4.63 | 0.74 |
| RPS26A | 4.36 | 0.80 | rpII | 4.68 | 0.73 | rpL20 | 4.68 | 0.73 |
| RPS20 | 4.39 | 0.84 | rpIW | 4.68 | 0.67 | rpL21 | 4.68 | 0.74 |
| RPL30 | 4.39 | 0.87 | rpIA | 4.76 | 0.78 | rpL1 | 4.69 | 0.85 |
| RPL24B | 4.42 | 0.77 | rpsA | 4.81 | 0.78 | rpL19 | 4.71 | 0.79 |
| RPS24A | 4.42 | 0.83 | rpIJ | 4.84 | 0.65 | rpL11 | 4.74 | 0.78 |
| RPS26B | 4.42 | 0.71 | rpIU | 4.85 | 0.65 | rpL24 | 4.77 | 0.69 |
| RPS23B | 4.43 | 0.80 | rpIX | 4.85 | 0.59 | rpIQ | 4.78 | 0.73 |
| RPL24A | 4.43 | 0.76 | rpIK | 4.86 | 0.72 | rpL13 | 4.79 | 0.79 |
| RPS10B | 4.43 | 0.80 | rpIS | 4.86 | 0.65 | rpS9 | 4.80 | 0.77 |
| RPS6B | 4.44 | 0.85 | rpsF | 4.86 | 0.69 | rpS1 | 4.81 | 0.83 |
| RPS25A | 4.44 | 0.73 | rpsJ | 4.86 | 0.58 | rps12 | 4.82 | 0.72 |
| RPS10A | 4.45 | 0.84 | rpsL | 4.87 | 0.67 | rpL18 | 4.83 | 0.69 |
| RPS6A | 4.45 | 0.84 | rpIP | 4.87 | 0.62 | rpL3 | 4.83 | 0.79 |
| RPL22A | 4.45 | 0.89 | rpIC | 4.87 | 0.72 | rpS7 | 4.85 | 0.77 |
| RPS9B | 4.46 | 0.84 | rpsC | 4.89 | 0.74 | rpL10 | 4.86 | 0.73 |
| RPL8B | 4.47 | 0.85 | rpsB | 4.90 | 0.78 | rpS11 | 4.87 | 0.75 |
| RPS14A | 4.48 | 0.80 | rpIV | 4.90 | 0.56 | rpS14 | 4.89 | 0.68 |
| RPL8A | 4.50 | 0.84 | rpIM | 4.90 | 0.68 | rpS5 | 4.90 | 0.73 |
| RPL18A | 4.50 | 0.82 | rpIQ | 4.92 | 0.56 | rpL16 | 4.91 | 0.75 |
| RPL17A | 4.50 | 0.81 | rpIR | 4.93 | 0.62 | rpL2 | 4.92 | 0.77 |
| RPL13B | 4.51 | 0.75 | rpsE | 4.94 | 0.60 | rpS8 | 4.92 | 0.75 |
| RPS17A | 4.51 | 0.81 | rpID | 4.95 | 0.70 | rpL4 | 4.93 | 0.73 |
| RPL36A | 4.51 | 0.68 | rpIF | 4.96 | 0.62 | rbfA | 4.93 | 0.58 |
| RPL25 | 4.52 | 0.76 | rpsN | 4.96 | 0.55 | rpS10 | 5.01 | 0.68 |
| RPL19B | 4.52 | 0.72 | rpIB | 4.97 | 0.72 | rpS2 | 5.02 | 0.78 |

level of confidence were found to be –0.672, –0.669 and –0.563 for *E. coli*, *S. cerevisiae* and *H. influenzae* respectively. From Figure 1 it can be concluded that the uncertainty values are highly correlated with gene expression levels (CAI). One should keep in mind that the uncertainty values were calculated for all the genomes under study using the simple Shannon's uncertainty formula as described earlier, but to calculate CAI one would require a different reference subset of highly expressed genes from their respective genomes.

It is also interesting to note that the uncertainty value ranges from 3.99 to 5.73 for *E. coli* genes, 4.05 to 5.77 for *S. cerevisiae* genes and 4.10 to 5.57 *H. influenzae* genes. We found no genes with uncertainty value of 5.93 bits per codon, indicating that all genes have some degree of internal constraints in their codon usage. One should not view the low correlation values for *H. influenzae* seriously, since uncertainty values try to capture all the hidden constraints (bias) that nature would face in getting the gene better-adopted to a given genome, while CAI values were calculated from the codon usage of optimal codons of a given gene with respect to a highly expressed gene present in the reference subset for the respective genome.

The low correlation value between the uncertainty and CAI values in the case of *H. influenzae* may be due to the fact that mutational bias is more effective than translational selection in detecting the codon bias in this highly skewed organism, particularly in the presumably highly expressed genes of ribosomal genes.

Genes are sorted according to the uncertainty values. Tables 1 and 2 show the uncertainty and their corresponding CAI values for 50 genes for each organism located at the two extreme ends in their uncertainty scale. From Tables 1 and 2 it is obvious that there is a high degree of correlation between uncertainty values and their corresponding CAI values in *E. coli* and *S. cerevisiae* genes located at the extreme end of the lowest uncertainty values. From Table 2 it was observed that the CAI for *H. influenzae* is comparatively higher than the *E. coli* and *S. cerevisiae* counterparts. The comparatively higher value of CAI of *H. influenzae* may be due to the fact that CAI values for its individual genes are related to its base composition, rather than its expression level. To find out whether genomic GC has any effect on uncertainty values, we have calculated the uncertainty values for the same functional category of genes such as ribosomal proteins of *E. coli*, *S.*

**Table 4.** Uncertainty values in bits per codon for experimentally known highly expressed genes of *S. cerevisiae*

| Gene | Uncertainty in bits per codon |
| --- | --- |
| *ENO2* | 4.52 |
| *TDH3* | 4.53 |
| *RPS31* | 4.53 |
| *RPL1A* | 4.54 |
| *RPL27A* | 4.56 |
| *TDH2* | 4.57 |
| *PGK1* | 4.60 |
| *PDC1* | 4.60 |
| *RPL2A* | 4.60 |
| *RPL5* | 4.60 |
| *FBA1* | 4.61 |
| *GPM1* | 4.61 |
| *TEF2* | 4.61 |
| *TEF1* | 4.62 |
| *RPL4A* | 4.64 |
| *ADH1* | 4.65 |
| *RPL16A* | 4.69 |
| *RPS18B* | 4.70 |
| *ADH2* | 5.10 |

*cerevisiae* and *H. influenzae* (Table 3). It was observed that the uncertainty values range between 3.99 and 5.02 bits per codon, emphasizing that majority of the genes of the same functional categories share the same degree of constraint irrespective of their genomic GC composition.

Uncertainty values of experimentally known highly expressed genes of *S. cerevisiae* were calculated to compare them with the other highly expressed genes based on CAI values (Table 4). It is evident from Table 4 that all the experimentally known highly expressed genes fall in the range of 4.52 to 5.09 bits per codon. These values fall in the range of other highly expressed genes, as shown for the ribosomal proteins in Table 3.

From these results we can conclude that Shannon's uncertainty principle can be used as an alternative method for predicting the level of gene expression. In other words, it can be said that Shannon's uncertainty values determine the degree of constraint present in the given gene, which is strongly correlated to the level of gene expression. This method does not require any prior reference subset of highly expressed genes, as required to calculate CAI values. To have an accurate CAI value, one would require knowing all highly expressed genes of the genome, whereas Shannon's uncertainty is immune to this limitation. For the vast amount of sequenced gene data present for different genomes, one can use the uncertainty values to determine the level of gene expression without knowing their proper functional annotation.

1. Wada, K., Aota, S., Tsuchiya, R., Ishibashi, F., Gojobori. T. and Ikemura, T., *Nucleic Acids Res.* (*Suppl.*), 1990, **18**, 2367–2411.

2. Ohkubo, S., Muto, A., Kawauchi, Y., Yamao, F. and Osawa, S., *Mol. Gen. Genet.*, 1987, **210**, 314–322.
3. Wright, F. and Bibb, M. J., *Gene*, 1992, **113**, 55–65.
4. Gupta, S. K., Bhattacharyya, T. K. and Ghosh, T. C., *Indian J. Biochem. Biophys.*, 2002, **39**, 35–48.
5. Ikemura, T., *J. Mol. Biol.*, 1981, **146**, 1–21.
6. Ikemura, T., *J. Mol. Biol.*, 1982, **158**, 573–587.
7. Bennetzen, J. L. and Hall, B. D., *J. Biol. Chem.*, 1982, **257**, 3026–3031.
8. Gupta, S. K. and Ghosh, T. C., *Gene*, 2001, **273**, 63–70.
9. Gouy, M. and Gautier, C., *Nucleic Acids Res.*, 1982, **10**, 7055–7074.
10. Andersson, S. G. and Sharp, P. M., *Microbiology*, 1996, **142**, 915–925.
11. Ghosh, T. C., Gupta, S. K. and Majumdar, S., *Int. J. Parasitol.*, 2000, **30**, 715–722.
12. McInerney, J. O., *Proc. Natl. Acad. Sci.*, 1998, **95**, 10698–10703.
13. Chiapello, H., Ollivier, E., Landes-Devauchelle, C., Nitschke, P. and Risler, J-L., *Nucleic Acids Res.*, 1999, **27**, 2848–2851.
14. Kerr, A. R. W., Peden, J. F. and Sharp, P. M., *Mol. Microbiol.*, 1997, **25**, 1177–1179.
15. Miranda, A. B. de., Valin, F. A., Jabbari, K., Degrave, W. M. and Bernardi, G., *J. Mol. Evol.*, 2000, **50**, 45–55.
16. Lio, P., Ruffo, S. and Buiatti, M., *J. Theor. Biol.*, 1994, **171**, 215–223.
17. Gupta, S. K., Majumdar, S., Bhattacharya, T. K. and Ghosh, T. C., *Biochem. Biophys. Res. Commun.*, 2000, **269**, 692–696.
18. D'Onofrio, G., Ghosh, T. C. and Bernardi, G., *Gene*, 2002, **300**, 179–187.
19. Adzhubei, A. A., Adzhubei, I. A., Krasheninnikov, I. A. and Neidle, S., *FEBS Lett.*, 1996, **399**, 78–82.
20. Tao, X. and Dafu, D., *FEBS Lett.*, 1998, **434**, 93–96.
21. Sharp, P. M. and Li, W-H., *Nucleic Acids Res.*, 1987, **15**, 1281–1295.
22. Schneider, T. D., Stormo, G. D., Gold, L. and Ehrenfeucht, A., *J. Mol. Biol.*, 1986, **188**, 415–431.
23. Hamming, R. W., Entropy and Shannon's first theorem. In *Coding and Information Theory*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1980, p. 107.
24. Shannon, C. E., *Bell Syst. Tech. J.*, 1948, **27**, 279–423, 623–656.
25. Velculescu, V. E. *et al.*, *Cell*, 1997, **88**, 243–251.