Unusual codon usage of HIV

SIR—It is interesting that although codon usage bias is organism-specific or gene-specific as far as the third degenerate position is concerned[1,2], it is almost general in the first two non-dengenerate codon positions[3]. Human immunodeficiency virus (HIV), the aetiological agent of AIDS (acquired immune deficiency syndrome), follows the general pattern in the first two positions but the bias is shifted towards a distinct preference for adenine at the expense of pyrimidine bases, in particular cytosine. The shift is markedly amplified in the third codon position of HIV genes while adenine is rare in the third codon position of genes in most other organisms and eukaryotes in particular. Purines predominate over pyrimidines in all codon positions of all HIV genes, which is unprecedented.

In an analysis of a recent compilation[4] of codon usage in 1,638 genes we have found no gene that is as deviant from mean codon usage as the genes of HIV. Nonetheless there are a few other viruses (influenza, human papilloma, cauliflower mosaic, papova, visna) showing a similar trend of adenine preference in their genes. Sharp[5] has found that HIV shares a surprisingly high number of the most preferred codons with influenza and cauliflower mosaic viruses. We add that most of these shared codons contain adenine in the degenerate position.

What accounts for the overrepresentation of adenine in HIV? It cannot be genome type or a factor specific for the host organism because these vary greatly among different adenine-rich viruses. Neither is it an involvement of reverse transcriptase in the viral reproduction cycle because the Moloney murine leukaemia retrovirus, for example, does not show a preference for adenine. Although HIV is not extreme in its A+T content (37 per cent A + 21 per cent T) we examined codon usage in A+T rich human genes and found correspondingly increased amounts of adenine in the third position of their codons. But the amount of adenine is far less than in HIV genes.

We have also considered the possibility that the excessive adenine is related to the great genetic variability of HIV. In this respect it is of interest that the mutation rate in vertebrate immunoglobulin genes correlates positively with the local A+T content[6]. But the overrepresentation of adenine in HIV is highest in the *pol* gene and lowest in the more variable *env* gene. Also, adenine distribution does not appear to be concentrated in the hypervariable segments of *env*.

In summary, we are unable to offer a plausible explanation for the existence of a group of viruses (but no bacteriophage), of which HIV is the most extreme representative, that seem to prefer adenine to alternative bases. Finding an explanation for this peculiar coding strategy could contribute to a better understanding of the evolution and pathogenesis of HIV.

JAROSLAV KYPR
JAN MRÁZEK

Institute of Biophysics,
Czechoslovak Academy of Sciences,
612 65 Brno, Czechoslovakia

[1] Grantham, R., Perrin, P. & Mouchiroud, D. *Oxford Surv. Evol. Biol.* **3**, 48-81 (1986).
[2] Ikemura, T. *Molec. evol. Biol.* **2**, 13-34 (1985).
[3] Kypr, J. & Mrázek, J. *Int. J. biol. Macromol.* **9**, 49-53 (1987).
[4] Maruyama, T., Gojobori, T., Aota, S. & Ikemura, T. *Nucleic Acids Res.* **14**, r151-197 (1986).
[5] Sharp, P.M. *Nature* **324**, 114 (1986).
[6] Perrin, P. *Nucleic Acids Res.* **12**, 5515-5527 (1984).

How many reactor accidents will there be?

SIR—The argument of Islam and Lindgren[1] that there is some difficulty in reconciling the predictions of future nuclear incidents based on observations of historical accidents with the claims made by nuclear designers that the chance of a nuclear reactor incident is one in 1 million per reactor year is based on the assumption that the incidents can be accurately described by the Poisson distribution involving two parameters—the failure rate, *r*, and years of reactor operation, *T*. Edwards[2] correctly points out that they have unwittingly adopted a bayesian approach and implicitly used a uniform prior probability density for the failure rate. He then proposes estimation based on transforming the likelihood for *r* to a likelihood for the probability of one or more incidents, $P = P(r) = 1 - e^{-rT}$. The difference between the two analyses is that, in a bayesian framework, Islam and Lindgren use a uniform prior on *r*, while Edwards implicitly uses a uniform prior on *P*, which is equally dubious. None of these authors has combined engineering expertise with data derived from operational experience, especially the highly relevant prior information contained in studies on nuclear reactor safety such as the Rasmussen[3] report for the Nuclear Regulatory Commission or articles by Rasmussen[4], Lewis[5] and Groer[6].

If one treats *r* as a random variable, the only coherent way to incorporate the extensive engineering information is to use the bayesian methodology with a prior distribution based on the best knowledge available at the time. Rasmussen[4] and Lewis[5] argue that the logarithm of *r* (to the base 10) for complete core meltdowns should be normally distributed with mean 4 and variance 1; alternatively, to the base e, this prior probability is $\ln r \sim N(-9.21, 5.30)$. Conditional on *r*, the density of time to the next incident is exponential with $p(x|r) = r\,e^{-rx}$; thus the posterior density of the failure rate (or of $P(r)$) can be obtained by direct integration:

$$p(r|x) = C \int p(x|r)\,p(r|0)\,dr$$
$$= C \int r\,e^{-rx}\,p(r|0)\,dr$$

where $p(r|0)$ denotes the prior density at time 0 without operating history and C is a normalizing constant. With the lognormal prior, integration is straightforward but messy; it has been reported, in a slightly different context, by Lewis[5] and Groer[6].

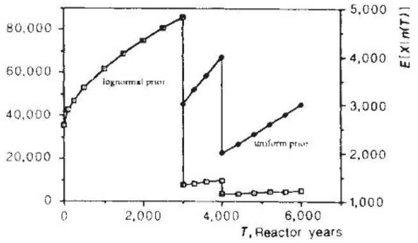

Fig. 1 Expected time to the next incident.

Combining information for partial core meltdowns[6] of PWR-9 with the information for complete core meltdowns[5], we find by a linear scaling of ln *r* that the prior for partial core meltdowns is $\ln r \sim N(-7.82, 5.3)$. This implies that the rate of partial to complete core meltdowns is about 4 to 1. Based on this lognormal prior it can be shown that the probability there is at least one partial core meltdown incident in ten years (374 reactors in operation from 1986 to 1996) is equal to 0.75. Because of the optimistic Rasmussen, Groer and Lewis priors, the estimate is reduced from the 0.86 figure found in Islam and Lindgren. However, it is much better to do this than to use a uniform prior which places equal weight on all values of *r*. Eventually the data will dominate any prior, but it is still too early to tell.

Insights on how operational experience combines with engineering expertise can be easily demonstrated if one uses, instead of the lognormal, a gamma distribution of the prior probability for which analytical results are available. When the prior density is approximated by a member of the gamma family with parameters (α, β), expected time to the next incident given $n(T)$ incidents by T is

$$E[X|n] = (T + \beta)[n + \alpha - 1]^{-1}$$
$$= (T/n)[n/(n + \alpha - 1)] +$$
$$(\beta/\alpha - 1)[1 - n/(n + \alpha - 1)]$$

provided $n + \alpha > 1$. With $0 \leq n + \alpha \leq 1$, the conditional expectation is infinite. As one might suspect, the expected time to the next incident is a weighted sum of the arithmetic mean T/n and the prior expectation of *X*. Unfortunately, it is not possible to fit the gamma well (simul-