Available online at www.sciencedirect.com
SCIENCE DIRECT
Biochemical and Biophysical Research Communications 344 (2006) 612–616
BBRC
www.elsevier.com/locate/ybbrc

# An investigation of genomic base distribution

David Mitchell [a,*], Robert Bridge [b]

[a] Vice Deanery of Genetics and Microbiology, Trinity College, Dublin, Ireland
[b] School of Chemistry, Trinity College, Dublin, Ireland

Received 22 March 2006
Available online 6 April 2006

**Abstract**

While veritable oceans of ink have been spilled over the base distributions within genes, the literature is virtually silent on large scale intra genomic base distribution. To address this issue, we have examined ~3400 chromosomal sequences from ~2000 entire genomes—including DNA and RNA, single- and double-stranded, coding and non-coding genomes. For each sequence the mean, variance, skewness, and kurtosis for each base were computed along with the genome base composition. The main findings are: (1) there is no simple relationship between these statistics and the base composition of the genome, (2) in non-viral genomes, base distribution is non-uniform, (3) base distribution in non-eukaryotic genomes obeys a number of simple rules, (4) these rules are not dependent on the presence of coding sequences, (5) bacterial genomes in particular are unusually compliant with these rules, and (6) eukaryotes have a unique pattern of base distribution.


*Keywords:* DNA; RNA; Genome; Virus; Bacteria; Eukaryote; Archebacteria; Organelle; Viroid

Since its discovery in 1869 [1], the resolution of its composition and structure [2–4], and the recognition of its importance in biology [5], DNA has played a central role in biological thought. Yet to the best of our knowledge, the large scale distribution of bases with genomes has not been previously been addressed. In published analyses of genomes, a common but largely unspoken assumption is that bases within the sequence can be—at least to a first approximation—treated as a 'random' string of characters. While in the case of some viral genomes this question has not been completely resolved, the findings presented here strongly suggest otherwise.

## Materials and methods

The genome sequences were obtained from the NCBI server and consisted of 3402 sequences from 1495 viral genomes, 835 organellar genomes, 231 bacterial, 20 archaeal genomes, and 14 eukaryotes: *Anopheles gambiense*, *Aradopsis thaliana*, *Candida albicans*, *Canis familiaris*, *Cryptococcus neoformans*, *Drosophilia melanogaster*, *Encephalitozoon cuniculi*, *Eremothecium gossypii*, *Kluyveromyces lactis*, *Homo sapiens*, *Leishmania major*, *Schizosaccharomyces pombe*, *Trypanosoma brucei*, and *Yarrowia lipolytica*. The viruses were initially divided into a number of types based on their genome (RNA/DNA, single/double stranded) but it was found—rather unexpectedly—that there were no obvious qualitative differences between these genome types. As a consequence the viruses have here been treated as a single group. The viroids were examined separately as these do not encode proteins.

The values of the moments of the uniform distribution are known exactly: the mean (0.5); the variance (1/12); the skewness (0); and the kurtosis (1.8). The mean, variance, skewness, and kurtosis of the bases within the sequences were calculated with the following formulae:

$$\text{Mean } (m) = (1/n) * \sum x_i,$$
$$\text{Variance } (s^2) = \sum (x_i - \mathrm{m})^2/(n-1),$$
$$\text{Skewness} = \sum (x_i - m)^3/[(n-1) * s^3],$$
$$\text{SE skewness} = (6/n)^{0.5},$$
$$\text{Kurtosis} = \sum (x_i - \mathrm{m})^4/[(n-1) * s^4],$$
$$\text{SE kurtosis} = (24/n)^{0.5}.$$

Here $x_i$ is the position of each base, $n$ is the number of bases, and SE is the standard error. The sums are taken over all the bases. The large number of bases meant that the asymptotic estimators could be used here rather than the finite sample estimators. The mean, skewness, and kurtosis

* Corresponding author. Fax: +353 1 679 9294.
*E-mail addresses:* dmitchel@tcd.ie (D. Mitchell), bridgero@tcd.ie (R. Bridge).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.03.182

were tested against the null hypothesis of a uniform distribution with Student's $t$ test [6] and the variances were compared with the $\chi^2$ test [7,8]. The test statistic used for the variance was

$$\chi^2 = (n-1) * s^2/\sigma^2,$$

where $s^2$ is the sample variance, $\sigma^2$ the theoretical variance, and $n$ is the number of bases used to estimate the sample variance. The number of degrees of freedom are $n-1$. As the test statistic is expected to be greater than 1, where $s < \sigma$ the numerator and denominator were exchanged.

A statistic based on generating functions [9] was also used. Let $X$ be a random variable that assumes the values 0, 1, 2, ... with probabilities $p_0$, $p_1$, $p_2$, ..., respectively. Let $P(s)$ be a power series

$$P(s) = p_0 + p_1 s + p_2 s^2 + p_3 s^3 + \cdots$$

Since the $p_i$ are probabilities, $P(1) = 1$. The derivative

$$P'(s) = \sum k p_k s^{k-1},$$

when evaluated at $s = 1$ converges formally to the mean value of $X$. The variance of $X$ is

$$\mathrm{Var}(X) = P''(1) + P(1)' - [P'(1)]^2,$$

where $P(1)''$ is the derivative of $P(s)$ at $s = 1$.

Although the generating function(s) for the base distributions are not yet known, the sample means and variances are. By rearranging the formula above, the $P''(1)$ values were computed:

$$P''(1) = P(1)' - [P'(1)]^2 - \mathrm{Var}(X).$$

This last statistic—$P''(1)$—currently lacks an official title and has been designated here as the 'reduced variance.' The proportion of unidentified bases was also recorded.

## Results

The intra group mean varied considerably with the range maximal in the organellar genomes (range: 0.37–0.60). One value (0.26) from sequence NC_000024—the *Homo sapiens* Y chromosome—was an outlier. This chromosome is only partly (43%) sequenced. We believe this problem to be responsible for the anomalous behaviour observed.

No consistent correlation between the means and the composition of the sequences was found here. The strongest relationship ($R^2 = 0.2254$) was found in the adenosine bases in the archaebacteria (Fig. 1). All the plots are visibly heteroscedastic, suggesting that the regressions are likely to be missing important variables—an observation that is not surprising on purely biological grounds. While the mean positions differed significantly from the null hypothesis (0.5) in many of the sequences (Table 1), in the majority of the viral and viroid genomes it was not possible to reject the null hypothesis with these tests. This pattern was also

Table 1
Percentage of genomes that differ significantly from a uniform distribution

| | A | C | G | T |
|---|---|---|---|---|
| Arch | | | | |
| Mean | 85.0 | 65.0 | 80.0 | 75.0 |
| Var | 80.0 | 75.0 | 75.0 | 70.0 |
| Skew | 65.0 | 70.0 | 80.0 | 70.0 |
| Kurt | 25.0 | 15.0 | 25.0 | 30.0 |
| Bact | | | | |
| Mean | 92.2 | 95.2 | 93.5 | 87.4 |
| Var | 76.6 | 71.0 | 74.0 | 74.5 |
| Skew | 86.1 | 92.6 | 91.8 | 82.7 |
| Kurt | 45.5 | 49.4 | 53.7 | 44.6 |
| Euks | | | | |
| Mean | 83.9 | 84.6 | 86.6 | 79.9 |
| Var | 79.9 | 81.2 | 81.9 | 78.5 |
| Skew | 67.1 | 72.5 | 79.2 | 70.5 |
| Kurt | 60.4 | 61.7 | 62.4 | 57.0 |
| Orgs | | | | |
| Mean | 45.4 | 53.7 | 85.0 | 65.1 |
| Var | 56.2 | 36.5 | 21.1 | 56.4 |
| Skew | 23.0 | 21.6 | 72.3 | 16.8 |
| Kurt | 2.8 | 1.6 | 1.8 | 1.8 |
| Viruses | | | | |
| Mean | 28.4 | 29.9 | 26.4 | 32.3 |
| Var | 13.1 | 9.6 | 11.0 | 14.4 |
| Skew | 13.4 | 12.5 | 11.6 | 15.5 |
| Kurt | 1.0 | 1.2 | 2.5 | 2.6 |
| Viroids | | | | |
| Mean | 25.0 | 5.6 | 8.3 | 19.4 |
| Var | 0.0 | 0.0 | 0.0 | 0.0 |
| Skew | 5.6 | 0.0 | 0.0 | 2.8 |
| Kurt | 0.0 | 0.0 | 0.0 | 0.0 |

*Abbreviations:* A, adenosine; C, cytosine; G, guanine; T, thymidine; Euks, eukaryotic genomes; Arch, archaebacterial genomes; Org, organellar genomes; Viruses, viral genomes; Viroids, viroid genomes. Example: 92.2% of the bacterial adenosine means are significantly different from a uniform distribution.

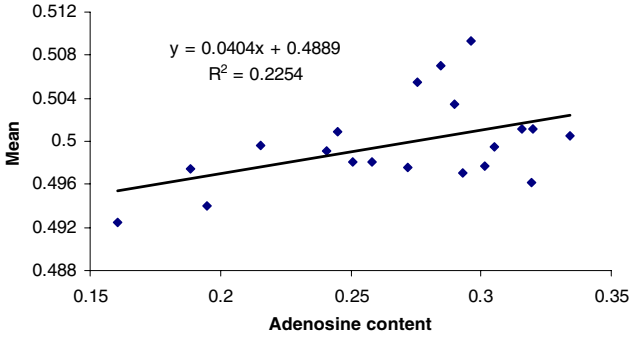


Fig. 1. Correlation of the genome content and mean of the adenosine distribution in 20 archaebacterial genomes. Heteroscedasticity is present. $F = 5.2$, $p < 0.04$.

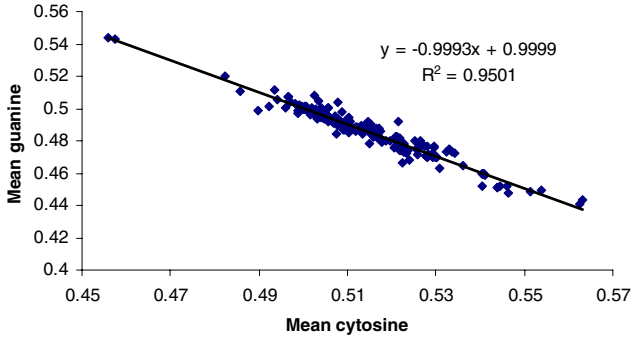


Fig. 2. The means of the cytosine and guanine distributions in 231 bacterial genomes. $F = 4376.0$, $p < 10^{-150}$.