On random sequence spacing distributions

C.T.J. Dodson
School of Mathematics, Manchester University
Manchester M60 1QD, UK ctdodson@manchester.ac.uk

Abstract

The random model for allocation of an amino acid A at locations along a protein chain is controlled by an underlying binomial distribution for occurrence of A, with density given by the relative abundance of A. Here we derive analytically the distribution of lengths of inter-A spaces in sequences of arbitrary length n, with arbitrary relative abundance p of occurrencies of A. This provides a well-defined reference structure for comparison with observations. We derive also the distribution function for inter-AA spaces.

It turns out that the standard deviation of inter-A space length is approximately proportional to the mean inter-A space length, independently of sequence length n and relative abundance p. This means that, to the extent that the space length can be approximated by a continuous random variable, the distribution of space lengths is represented by a gamma distribution. The significance of this latter approximation is that the gamma family may provide models for spacing length distributions when the underlying process is not binomial, for example, when clustering or evening out of amino acids occurs. The new results show that actual sequences of amino acids also do have standard deviation of spacing lengths proportional to the mean but all exhibit more self-clustering than expected for finite sequences from a random process.

1 Binomial model

We think of a protein chain as a sequence of amino acids among which we have distinguished one, represented by the letter A, while all others are represented by X. The relative abundance of A is given by the probability p that an arbitrarily chosen location has an occurrence of A. Then $1-p$ is the probability that the location contains a different amino acid from A; here, all locations are occupied by some amino acid. If the A locations are chosen with uniform probability subject to the constraint that the net density of A in the chain is p, then either A happens or it does not; we have a binomial process.

Therefore, in a sequence of n amino acids, the mean or expected number of occurrences of A is np and its variance is $np(1-p)$, but it is not immediately clear what will be the distribution of lengths of spaces between occurrencies of A. Evidently the distribution of such lengths r, measured in units of one location length also is controlled by the underlying binomial distribution.

We are interested in the probability of finding in a sequence of n amino acids a subsequence of form

$$\underbrace{\cdots XA\overbrace{X\cdots X}AX\cdots},$$

where the overbrace $\overbrace{\ }$ encompasses precisely r amino acids that are not A and the underbrace $\underbrace{\ }$ encompasses precisely n amino acids, the whole sequence.

2 Derivation of the distributions

In a sequence of n locations filled by amino acids we consider the probability of finding a subsequence containing two A's separated by exactly r non-A X's, that is the occurrence of an inter-A space length r.

It is not difficult to see that the probability distribution function $\mathbb{P}(r,p,n)$ for inter-A space length r reduces to the first expression below (1), which is a geometric distribution and simplifies to (2)

$$\begin{aligned}\mathbb{P}(r,p,n) &= \frac{\left(p^2(1-p)^r(n-r-2)\right)}{\sum_{r=0}^{n-2}(p^2(1-p)^r(n-r-2))}, &(1)\\ &= \frac{(1-p)^{1+r}\,p^2\,(n-r-2)}{-1+(1-p)^n+p\,(n+p-n\,p)}, &(2)\\ &\qquad\qquad \text{for}\;\; r=0,1,\ldots,(n-2).\end{aligned}$$

By setting $r=0$, equation (2) gives $\mathbb{P}(0,p,n)$, the probability of finding a pair of form AA.
The mean $\bar{r}$ and standard deviation σ_r of the distribution (2) are given for $r=0,1,\ldots,(n-2)$, by

$$\begin{aligned}\bar{r} &= \sum_{r=0}^{n-2} r\,\mathbb{P}(r,p,n)\\ &= \frac{((1-p)^n\,(2+(-3+n)\,p))+(-1+p)^2\,(-2+(-1+n)\,p)}{p\,((1-p)^n+p\,(n+p-n\,p)-1)} &(3)\\ \sigma_r &= \sqrt{\left(\sum_{r=0}^{n-2} r^2\,\mathbb{P}(r,p,n)\right)-\bar{r}^2}\\ &= \sqrt{\frac{(p-1)\,\left(-2\,(1-p)^{2\,n}-(1-p)^n\,K(r,p,n)-(p-1)^2\,(2+(n-1)\,p\,(n\,p-4))\right)}{p^2\,((1-p)^n+p\,(n+p-n\,p)-1)^2}}, &(4)\end{aligned}$$

where we make the abbreviation

$$\begin{aligned}K(r,p,n) &= \left(4\,n\,p-4+(n-6)\,(n-1)\,p^2+(n-2)^2\,(n-1)\,p^3\right) &(5)\\ &\qquad\qquad \text{for } r=0,1,\ldots,(n-2).\end{aligned}$$

Example distributions are shown in Figure 1, for a sequence of $n=200$ amino acids with mean probability $p=0.01$ (left, red) which has from equations (3), (4) $\bar{r}=46.9$, $\sigma_r=39.3$, and $p=0.02$ (right, blue) which has from equations (3), (4) $\bar{r}=34.1$, $\sigma_r=31.5$.

3 Properties of the distributions

The coefficient of variation is given by

$$cv_r=\frac{\sigma_r}{\bar{r}}=\frac{p\,((1-p)^n-1+p\,(n+p-n\,p))\,L(r,p,n)}{(1-p)^n\,(2+(n-3)\,p)+(p-1)^2\,((-1+n)\,p-2)}$$

where we make the abbreviations

$$\begin{aligned}L(r,p,n) &= \sqrt{\frac{(1-p)\,\left(2\,(1-p)^{2\,n}+(1-p)^n\,M(r,p,n)+(p-1)^2\,(2+(n-1)\,p\,(n\,p-4))\right)}{p^2\,((1-p)^n+p\,(n+p-n\,p-1))^2}} &(6)\\ M(r,p,n) &= \left(4(\,n\,p-1)+(n-6)\,(n-1)\,p^2+(n-2)^2\,(n-1)\,p^3\right). &(7)\end{aligned}$$

The two main variables are the number of amino acids, n, in the sequence and the abundance probability p of occurrence of A; their effects on the statistics of the distribution of inter-A space lengths are illustrated in Figure 2 and Figure 3, respectively.

4 Inter-pair spacing distributions

Now we are interested in the probability of finding in a sequence of n amino acids a subsequence of form

$$\underbrace{\cdots XAA\overbrace{X\cdots X}AAX\cdots}_{},$$

where the overbrace $\overbrace{\ \ }$ encompasses precisely r amino acids that are not A and the underbrace $\underbrace{\ \ }$ encompasses precisely n amino acids, the whole sequence.

Following a similar approach to that in section 2 we can obtain the probability distribution of inter-AA spacings, in sequences of length n with relative abundance of A given by p. This distribution is:

$$\mathbb{Q}(r,p,n) = \frac{(1-p)^{3+r}\, p^2\, (4-n+r)}{-(1-p)^n + (p-1)^3\, ((n-3)\, p-1)} \quad \text{for } r = 0,1,\ldots,(n-4). \tag{8}$$

The mean inter-AA spacing is given by

$$\bar{r}_{AA} = -\left(\frac{(1-p)^n\, (2+(n-5)\, p) + (p-1)^4\, (-2+(-3+n)\, p)}{-((1-p)^n\, p) + (-1+p)^3\, p\, (-1+(-3+n)\, p)} \right) \quad \text{for } r = 0,1,\ldots,(n-4). \tag{9}$$

The standard deviation is known but it is rather cumbersome to express. Qualitative properties of the distribution (8) of inter-AA space lengths are similar to those of the inter-A space lengths (2).

5 Relation to gamma distributions

Roughly, the standard deviation of the inter-A space lengths is proportional to the mean; mean and standard deviation increase monotonically with increasing n and decrease monotonically with increasing p. In fact, for both cases shown in Figures 2 and 3, the numerical values lie close to the line of slope 1, namely

$$\sigma_r = \bar{r}. \tag{10}$$

It is known that the property of having coefficient of variation independent of the mean actually characterizes gamma distributions, as shown recently by Hwang and Hu [2]. They proved, for $m \geq 3$ independent positive random variables $x_1, x_2, \ldots, x_m$ with a common continuous probability density function h, that having independence of the sample mean $\bar{x}$ and sample coefficient of variation is equivalent to h being a gamma distribution.

The family of gamma distributions has continuous event space over the positive real numbers $\mathbb{R}^+$, its two parameters are $\tau, \nu \in \mathbb{R}^+$ and probability density functions are given by

$$f(t;\tau,\nu) = \left(\frac{\nu}{\tau}\right)^{\nu} \frac{t^{\nu-1}}{\Gamma(\nu)}\, e^{-t\nu/\tau}. \tag{11}$$

Here, Γ is the gamma function defined by

$$\Gamma(\nu) = \int_0^\infty s^{\nu-1}\, e^{-s}\, ds, \quad \text{(for positive integer } n,\ \Gamma(n) = (n-1)!).$$

Then $E(t) = \tau$ is the mean and $Var(t) = \tau^2/\nu$ is the variance, so the coefficient of variation $\sqrt{Var(t)}/\tau = 1/\sqrt{\nu}$ is independent of the mean.

Figure 4 shows the plot of gamma parameter $\nu = \left(\frac{\bar{r}}{\sigma_r}\right)^2$ in (11) against mean $\bar{r}$ for inter-A space length distributions (2). The mean probability for the occurrence of A is $p = 0.01$ (red) and $p = 0.02$ (blue), corresponding to the cases in Figures 1, 2. For asymptotically long sequences, $n \to \infty$, we expect then to have $\nu \to 1$, the random case. For random spacings in an infinite sequence we have a gamma distribution with $\nu = 1$, which is precisely the exponential distribution complementary to a Poisson process of events.

In Figure 5 we show the joint effects of p and n on ν; for large n, as expected, we seem to have $\nu \to 1$. However, we do find limiting values less than 1. For example

$$\text{At } p = 0.01, \quad \lim_{n\to\infty} \nu = 0.99 \tag{12}$$

$$\text{At } p = 0.1, \quad \lim_{n\to\infty} \nu = 0.9. \tag{13}$$

Indeed, we find that $\nu < 1$ for some sequence lengths in the range of practical interest; for example, the parameter ν passes from above 1 at small p, n through the plane at height 1, and then remains at about 0.9. The two Figures 4, 5 illustrate the apparent departure from randomness that arises from sampling an underlying random process with finite sequences.

Clearly, from Figure 4, the estimate of the gamma parameter, hence also the coefficient of variation, is not independent of the mean so we do not expect a gamma distribution to be a perfect model. Significantly however, all values shown there indicate $\nu > 1$, which differs from the random case $\nu = 1$; it differs also from the observed spacings of the twenty amino acids in 6294 proteins with sequence lengths up to $n = 4092$, all had $\nu < 1$ [1]. Thus, we have that finite sequences drawn from an infinite random sequence exhibit $\nu > 1$, which corresponds to an apparent smoothing or evening out of occurrencies whereas actual sequences of amino acids exhibit $\nu < 1$, which corresponds to self-clustering [1]. Furthermore, the experimental data in [1] gave standard deviation and mean values lying around the line

$$\sigma_r = \frac{4}{3}\,\bar{r} \tag{14}$$

with relative abundance values in the range $0.01 \leq p \leq 0.1$ and sequence lengths up to $n = 4092$. We see that this equation differs from the random case in equation (10) by having a slightly steeper slope.

The significance of this is that the gamma family may indeed provide models for spacing length distributions when the underlying process is not binomial, for example, when clustering or evening out of amino acids occurs. The new results show that though also actual sequences of amino acids do have standard deviation of spacing lengths proportional to the mean, nevertheless their spacings exhibited more self-clustering than expected for finite sequences from a random process.

References

[1] Y. Cai, C.T.J. Dodson, A.J. Doig and O. Wolkenhauer. Information theoretic analysis of protein sequences shows that amino acids self cluster. *J. Theor. Biol.* 218, 4 (2002) 409-418. http://www.ma.umist.ac.uk/kd/PREPRINTS/amino.pdf

[2] T-Y. Hwang and C-Y. Hu. On a characterization of the gamma distribution: The independence of the sample mean and the sample coefficient of variation. *Annals Inst. Statist. Math.* 51, 4 (1999) 749-753.

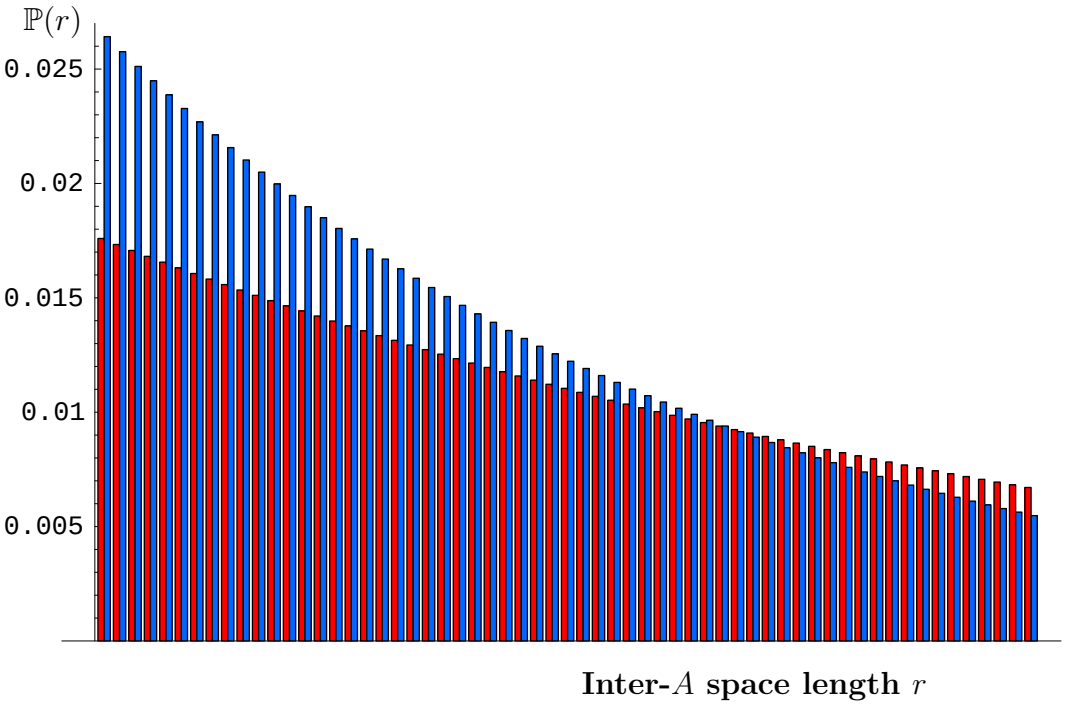

Figure 1: *Probability distribution of space length r between occurrences of amino acid A shown for the range $0 \leq r \leq 60$ in a random sequence of $n = 200$ amino acids. The mean probability for the occurrence of A is $p = 0.01$ (left, red) where we find $\bar{r} = 46.9$, $\sigma_r = 39.3$, and $p = 0.02$ (right, blue) where we find $\bar{r} = 34.1$, $\sigma_r = 31.5$.*

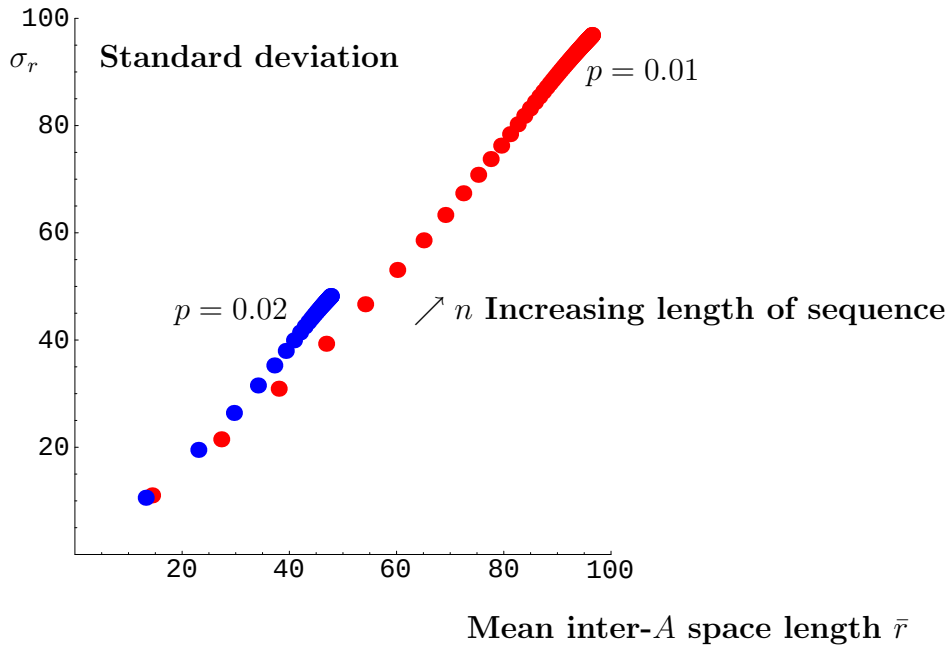

Figure 2: *Effect of sequence length n in random amino acid sequences of length from $n = 50$ to $n = 4000$ in steps of 50. Plot of standard deviation σ_r against mean $\bar{r}$ for inter-A space length distributions (2). The mean probability for the occurrence of A is $p = 0.01$ (right, red) and $p = 0.02$ (left, blue), corresponding to the cases in Figure 1. The standard deviation is roughly equal to the mean; mean and standard deviation increase monotonically with increasing n.*

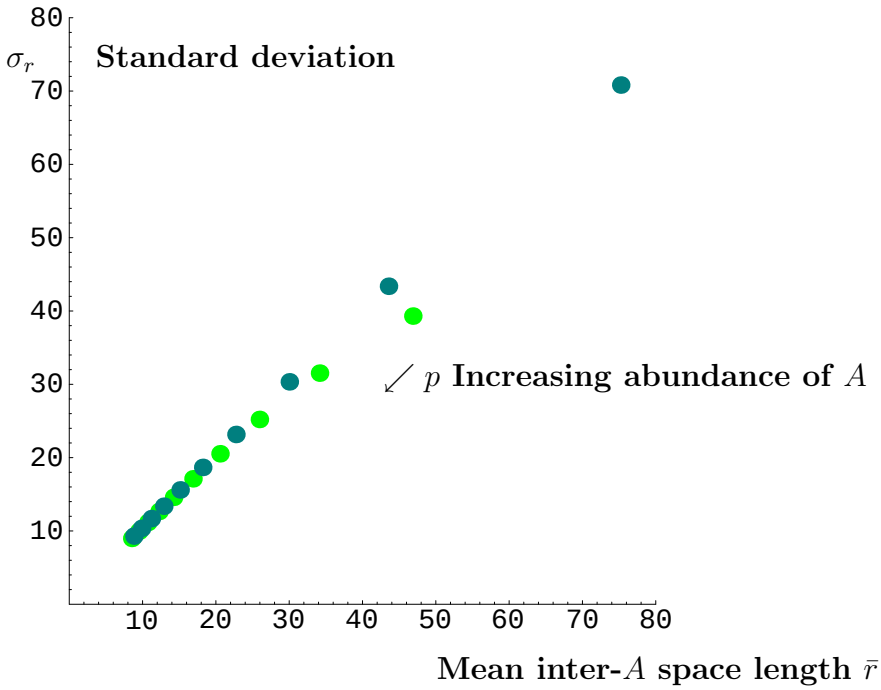

Figure 3: *Effect of relative abundance, probability* p*, over the range* $0.01 \leq p \leq 0.1$ *in steps of* 0.01. *Plot of standard deviation* σ_r *against mean* $\bar{r}$ *for inter-*A *space length in random sequences of length* $n = 200$ *amino acids (light green) and length* $n = 500$ *amino acids (dark green), with probability* $0.01 \leq p \leq 0.1$ *for occurrence of* A. *The standard deviation is for many practical purposes proportional to the mean; mean and standard deviation decrease monotonically with increasing* p.

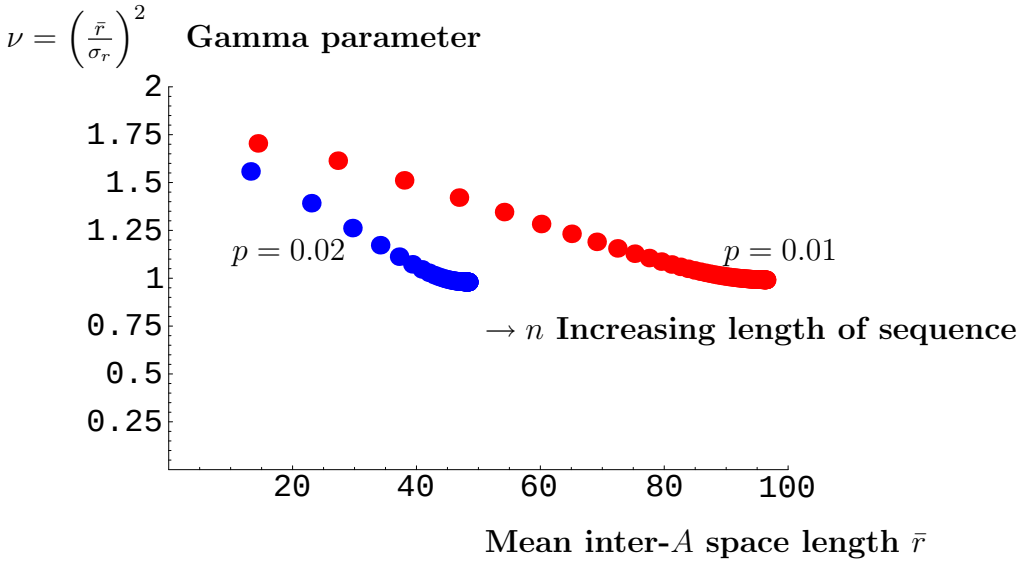

Figure 4: *Effect of sequence length* n *in random amino acid sequences of length from* $n = 50$ *to* $n = 4000$ *in steps of* 50. *Plot of gamma parameter* ν *from (11) against mean* $\bar{r}$ *for inter-*A *space length distributions (2). The mean probability for the occurrence of* A *is* $p = 0.01$ *(right, red) and* $p = 0.02$ *(left, blue), corresponding to the cases in Figures 1 and 2. We expect that, as* $n \rightarrow \infty$, *so* $\nu \rightarrow 1$, *the random case.*

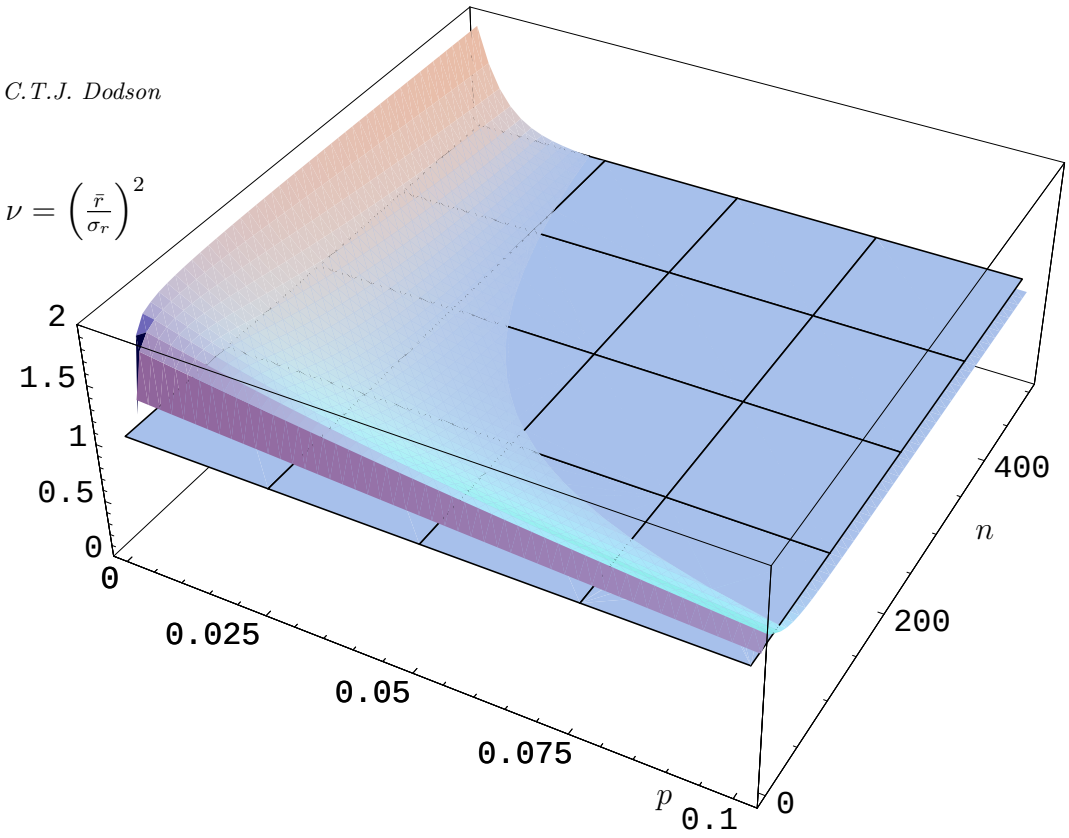

Figure 5: *Effect of relative abundance p, and sequence length n in random amino acid sequences of length from $n = 50$ to $n = 500$. Plot of gamma parameter ν from (11) for inter-A space length distributions (2). This illustrates apparent departures from randomness when we sample the process with finite sequences; in the case of infinite sequences we expect to recover the random case, shown by the plane $\nu = 1$. In fact, here the parameter ν passes from above at small p, n through the plane at height* 1, *and then remains at about* 0.9.

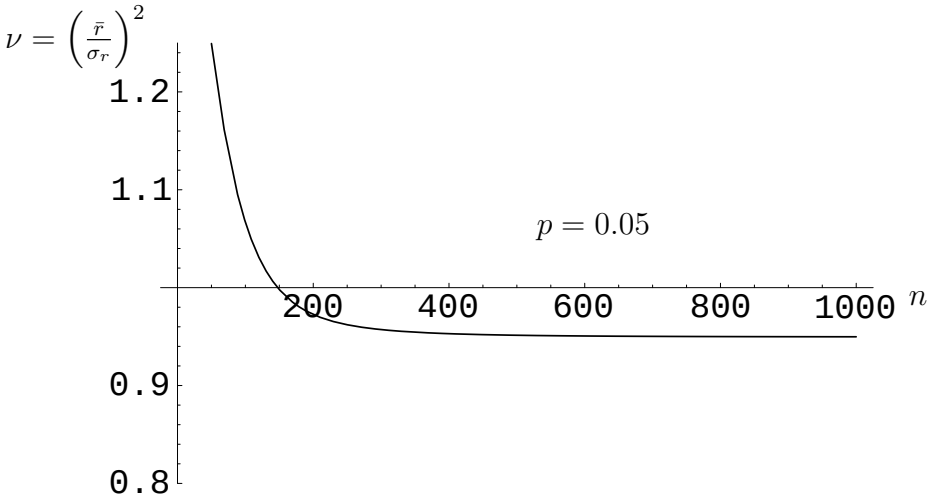

Figure 6: *A section through Figure 5. Effect on gamma parameter ν of sequence length n in random amino acid sequences for fixed relative abundance $p = 0.05$. The value $p = 0.05$ was the average relative abundance of the 20 amino acids for the data reported in [1] but that data yielded values* $0.59 \leq \nu \leq 0.95$, *indicating self-clustering.*