Information theoretic analysis of protein sequences shows that amino acids self cluster

Yudong Cai[1], C.T.J. Dodson[2], Andrew J. Doig[1] and Olaf Wolkenhauer[1,3]
[1,3]Department of Biomolecular Sciences, [2]Department of Mathematics,
[3]Department of Electrical Engineering and Electronics
UMIST Manchester M60 1QD, UK

May 6, 2002

Abstract

We analyse for each of 20 amino acids X the statistics of spacings between consecutive occurrences of X within the well-characterised *Saccharomyces cerevisiae* genome. The occurrences of amino acids may exhibit near random, clustered or smoothed out behaviour, like 1-dimensional stochastic processes along the protein chain. If amino acids are distributed randomly within a sequence then they follow a Poisson process and a histogram of the number of observations of each gap size would asymptotically follow a negative exponential distribution. The novelty of the present approach lies in the use of differential geometric methods to quantify information on sequencing of amino acids and groups of amino acids, via the sequences of intervals between their occurrences. The differential geometry arises from an information-theoretic distance function on the 2-dimensional space of stochastic processes subordinate to gamma distributions—which latter include the random process as a special case. We find that maximum-likelihood estimates of parametric statistics show that all 20 amino acids tend to cluster, some substantially. In other words, the frequencies of short gap lengths tends to be higher and the variance of the gap lengths is greater than expected by chance. This may be because localising amino acids with the same properties may favour secondary structure formation or transmembrane domains. Gap sizes of 1 or 2 are generally disfavoured, 1 strongly so. The only exceptions to this are Gln and Ser, as a result of poly(Gln) or poly(Ser) sequences. There are preferences for gaps of 4 and 7 that can be attributed to α-helices. In particular, a favoured gap of 7 for Leu is found in coiled-coils. Our method contributes to the characterisation of whole sequences by extracting and quantifying stable stochastic features.

Keywords: Protein sequences, amino acids, information theory, stochastic processes, random, clustering

1 Introduction

Our fundamental approach is to analyse the statistics of the separation between the consecutive occurrences of each amino acids X within the well-characterised *Saccharomyces cerevisiae* genome [(Goffeau et al, 1996)]. For example, in the sequence AKLAATWPFDA, we observe for Ala gaps of 1, 3 and 6 since the successive Ala residues are spaced i,i+3, i,i+1 and i,i+6, respectively.

The basic random model for spatial stochastic processes representing the distribution of events along a line is that arising from a Poisson process; then the lengths of gaps between successive events follows an exponential distribution with just one parameter, the mean τ, which is a positive number. A natural generalization of the exponential distribution is the gamma distribution, which has two positive parameters, the mean τ as before, and an extra one, ν; this 'dispersion parameter' ν controls the shape and we recover the simple exponential distribution by setting $\nu = 1$ for a random distribution in an infinitely long sequence. The important feature of the gamma distribution is that for $\nu < 1$ it corresponds to gap lengths that have a higher variance than for a random (Poisson) process, so the events giving rise to the gaps exhibit clustering. For $\nu > 1$, the gamma distribution represents the opposite of clustering—smoothing or evening out of the underlying disposition of events; for $\nu = 1$ we have precisely the random case. We note that the real gap length distribution for amino acid spacings is not actually a continuous variable, because of the structure of proteins; however, the approximation by a continuous variable is not believed to be likely to compromise any important statistical features of our predictions.

Now, families of 2-parameter probability density functions, like the gamma distributions for all positive values of ν and τ, can be given the structure of a curved surface whereon the metric or distance structure is controlled by the maximum likelihood function. In fact the gamma family has been applied elsewhere to cryptanalysis [(Dodson and Thompson, 2000)], galactic clustering [(Dodson, 1999a), (Dodson, 1999b)] and communication clustering [(Dodson, 2000)] for example. In our present context, the gaps to model are those along a protein chain, between consecutive occurrences of a given amino acid; we fit the maximum likelihood gamma parameters from the sample mean and variance of observed gap frequencies. Thus, we perform an information-theoretic analysis of the whole sequence and extract stable stochastic features which in particular quantify departures from randomness.

2 Statistics of random spacing sequences

In the perfectly *random* case of haphazard allocation of events along a line, the result is an exponential distribution of inter-event gaps when the line is infinite. However, for finite length processes it is a little more involved and we need to analyse this first in order to provide our reference structure; further details can be found in [(Dodson, 2001)].

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

It follows that, in a sequence of n amino acids, the mean or expected number of occurrences of X 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 consecutive occurrencies of X. 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

$$\underbrace{\cdots ?X\overbrace{?\cdots ?}X?\cdots},$$

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

2.1 Derivation of the distributions

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

The probability distribution function $\mathbb{P}(r,p,n)$ for inter-X space length r reduces to the first expression below (1), which is a geometric distribution and simplifies to (2)

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

The mean $\bar{r}$ and standard deviation σ_r of the distribution (2) are given for $r = 0, 1, \ldots, (n-2)$, by

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

where we make the abbreviation

$$N(r,p,n) \;=\; \left(4\,n\,p-4+(n-6)\ (n-1)\ p^2+(n-2)^2\ (n-1)\ p^3\right) \qquad (5)$$

$$\text{for } r = 0, 1, \ldots, (n-2).$$

Example distributions are shown in Figure 1, for a sequence of $n = 200$ amino acids in which X has 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) $\bar{r} = 34.1$, $\sigma_r = 31.5$.

The coefficient of variation is given by

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

where we make the abbreviations

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

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

3 Non-random amino acid sequences as gamma processes

Now we turn to the more realistic situation in which the allocation is not purely random for the occurrence of amino acids along a protein chain. Here we do not have an underlying analytic process like the binomial so we resort to a family of models for the distribution of spacings between occurrences of amino acids; it includes the random case but it covers also a range of processes around the random case.

The family of gamma distributions has event space $\Omega = \mathbb{R}^+$, parameters $\tau, \nu \in \mathbb{R}^+$ and has probability density functions given by

$$f(t;\tau,\nu) = \left(\frac{\nu}{\tau}\right)^{\nu} \frac{t^{\nu-1}}{\Gamma(\nu)}\, e^{-t\nu/\tau} \qquad (8)$$

where 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. In fact, this latter property actually characterizes gamma distributions as shown recently [(Hwang and Hu, 1999)]. They proved, for $n \geq 3$ independent positive random variables $x_1, x_2, \ldots, x_n$ with a common continuous probability density function h, that having independence of the sample mean $\bar{x}$ and sample coefficient of variation $cv = S/\bar{x}$ is equivalent to h being a gamma distribution.

The special case $\nu = 1$ corresponds to the situation of the random or Poisson process with mean inter-event interval τ. In fact, for *integer* $\nu = 1, 2, \ldots,$ equation (8) models a process that is Poisson but with intermediate events removed to leave only every ν^{th}. Formally, the gamma distribution is the ν-fold convolution of the exponential distribution, called also the Pearson Type III distribution.

Thus, gamma distributions can model a range of stochastic processes corresponding to non-independent clustered events, for $\nu < 1$, and dispersed or smoothed events, for $\nu > 1$, as well as the random case $\nu = 1$. Figure 4 shows sample gamma distributions, all of unit mean, representing clustering, random and dispersed spacing distributions, respectively, with $\nu = \frac{1}{2},\ 1,\ 2$.

Elsewhere we have discussed the differential geometry of spaces of gamma distributions and their application to clustering problems and security testing [(Dodson, 2000), (Dodson, 1999a), (Dodson, 1999b), (Dodson and Matsuzoe, 2002), (Dodson and Thompson, 2000)].

The Riemannian information metric on the parameter space $\mathcal{G} = \{(\tau, \nu) \in \mathbb{R}^+ \times \mathbb{R}^+\}$ for gamma distributions is given by the arc length function

$$ds_{\mathcal{G}}^2 = \frac{\nu}{\tau^2}\, d\tau^2 + \left(\psi'(\nu) - \frac{1}{\nu}\right) d\nu^2 \quad \text{for } \tau, \nu \in \mathbb{R}^+, \tag{9}$$

where $\psi(\nu) = \frac{\Gamma'(\nu)}{\Gamma(\nu)}$ is the logarithmic derivative of the gamma function. The 1-dimensional subspace parametrized by $\nu = 1$ corresponds to all possible 'random' (Poisson) processes, or equivalently, exponential distributions.

4 Results

Table 1 gives the values of the total number, mean, variance and ν for each amino acid. We see a large variation in mean gap size τ, ranging from 10.75 (Ser) to 61.82 (Trp). This can be largely explained on amino acid frequency; the gap to the next amino acids will tend to be smaller if the amino acid is more abundant. Similarly, rare amino acids, such as Cys, His, Met and Trp will be more widely spaced. There is a therefore a smooth inverse correlation between mean gap size and amino acid frequency. Clustering is better revealed by the gamma distribution analysis. In all cases, we see that $\nu < 1$; hence every amino acid tends to cluster with itself. There is some variation, with Cys and Trp most clustered and Ile and Val distributed almost randomly. Figure 9 shows an example of the gap distribution histogram for Leu, compared to its maximum likelihood gamma distribution and the exponential distribution with same mean.

Figure 8 shows a plot of all 20 amino acids as points on a surface over the space of (τ, ν) values from Table 1; the height of the surface represents the information-theoretic distance from the point marked as $\bullet$, the case of randomly distributed amino acids with mean $\tau = 20$ and $\nu = 1$. The calculation of this distance uses the differential geometry of the family of all gamma distributions, for which the arc length function is non-Euclidean and given in parametric form by equation (9). The geodesic mesh method for distance estimation is described in the Appendix and equation (15) there provides the analytic expression.

Note that if an exponential distribution gave the maximum likelihood fit then it would yield $\nu \approx 1$. This is arguably within experimental tolerance for I,N and V, but unlikely in the other cases having maximum likelihood $\nu \leq 0.85$. Our illustration in Figure 9 shows a case with $\nu = 0.85$. Significantly, experimentally *we find no case of* $\nu > 0.97$ but the analytic results for the case of finite truly random sequences, illustrated in Figure 6 and Figure 7, did not yield $\nu < 0.95$ in the regime of interest. Indeed, Figure 7 shows the effect on gamma parameter ν of sequence length n in random amino acid sequences for fixed relative abundance $p = 0.05$, the observed average relative abundance of the 20 amino acids in 6294 proteins with sequence lengths up to $n = 4092$; these yielded values $0.59 \leq \nu \leq 0.95$, indicating that all of the amino acids exhibited self-clustering.

Thus, we conclude that our methods therefore reveal an important qualitative property: universal self-clustering for these amino acids, stable over long sequences. Moreover, the information-theoretic geometry allows us to provide quantitative measurements of departures from randomness, as illustrated graphically in Figure 8; such depictions of the space of gap distributions could prove useful in the representation of trajectories for evolutionary or other structurally modifying processes.

5 Why do amino acids cluster?

One possible explanation relates to secondary structure preferences. α-Helices are typically 4-15 and β-strands 2-8 amino acids in length [(Penel et al., 1999)]. In order for any secondary structural element to form, it is necessary to have most amino acids within its sequence to have a high propensity for that structure. Identical amino acids will therefore cluster over these length ranges as this will favour a sequence with a high preference for forming one particular secondary structure. For example, Ala has a high preference for the α-helix. Hence evolution will select sequences where Alanines are clustered in order to favour α-helix formation. If amino acids were randomly distributed, the probability that a stretch of amino acids would contain a high preference for a secondary structural element would be decreased. A second possible explanation is that amino acids of similar hydrophobicity cluster in order to produce a hydrophobic membrane spanning sequence or water exposed polar loop.

Figure 9, with data for Leu illustrates as expected that there are some deterministic effects arising from the preferred spatial configurations.

We get a feel for the overall extent of the non-stochastic aspects from the histogram of all data, shown in Figure 10 with the maximum likelihood gamma fit. Notably, gap sizes of 1 or 2 are disfavoured, 1 strongly so. The only exceptions to this are Gln and Ser, which strongly favour short gaps of 1, 2 or 3. Poly(Gln) sequences, that give a high frequency of gaps of 1, are a well known feature of a number of proteins, and are implicated in several diseases, including Huntington's chorea [(Kaytor and Warren, 1999)]. Gaps of 3-12 are generally favoured, perhaps because this is the usual length of secondary structure. There are also local preferences for gaps of 4 and 7 that can be attributed to α-helices. Side chains spaced i,i+4 and i,i+7 are on the same side of an α-helix so can bond to one another. Sequences are favoured that have identical side chains close in space in the α-helix. In particular, a favoured gap of 7 for Leu can be attributed to coiled-coils that are characterised by pairs of α-helices held together by hydrophobic faces with Leu spaced i,i+7 [(Landschulz et al., 1988)], [(Lupas, 1996)], [(O'Shea et al., 1989)].

Clearly, the maximum likelihood gamma distributions fit only stochastic features and in that respect view the data as exhibiting transient behaviour at small gap sizes; other methods are available for interpretation of such deterministic features and we concentrate in this article on representation of whole sequences as a stochastic process. Our method contributes to the characterisation of whole sequences by extracting and quantifying stable stochastic features. In conclusion, fitting histograms of gap size frequency distributions to gamma distributions shows that all 20 amino acids tend to self-cluster in protein sequences.

Acknowledgements
This work was funded in part by the BBSRC (Grant number 36/BIO14432) and EPSRC (Grant number GR/N21871).

References

[(Dodson, 1999a)] Dodson, C.T.J. (1999). Evolution of the void probability function. Presented at **Workshop on Statistics of Cosmological Data Sets**, 8-13 August 1999, Isaac Newton Institute, Cambridge.
`http://www.ma.umist.ac.uk/kd/PREPRINTS/vpf.ps`.

[(Dodson, 1999b)] Dodson, C.T.J. (1999). Spatial statistics and information geometry for parametric statistical models of galaxy clustering. *Int. J. Theor. Phys.*, **38**, 10, 2585-2597.

[(Dodson, 2000)] Dodson, C.T.J. (2000). Information geodesics for communication clustering. *J. Statistical Computation and Simulation* **65**, 133-146.

[(Dodson, 2001)] Dodson, C.T.J. On amino acid spacing distributions. (2001). Preprint: `http://www.ma.umist.ac.uk/kd/PREPRINTS/gapdist.pdf`.

[(Dodson and Matsuzoe, 2002)] Dodson, C.T.J. and Matsuzoe, H. (2002). An affine embedding of the gamma manifold. *InterStat* January 2002, 1-6.
`http://www.ma.umist.ac.uk/kd/PREPRINTS/affimm.pdf` .

[(Dodson and Poston, 1991)] Dodson, C.T.J. and Poston, T. (1991). **Tensor Geometry** Graduate Texts in Mathematics 130, Second edition, Springer-Verlag, New York.

[(Dodson and Thompson, 2000)] Dodson, C.T.J. and Thompson, S.M. (2000). A metric space of test distributions for DPA and SZK proofs. *Poster Session*, **Eurocrypt 2000**, Bruges, 14-19 May 2000.
`http://www.ma.umist.ac.uk/kd/PREPRINTS/mstd.pdf`.

[(Goffeau et al, 1996)] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. G. (1996). Life with 6000 genes. *Science* **274**, 546, 563-567.

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

[(Kaytor and Warren, 1999)] Kaytor, M. D. and Warren, S. T. (1999). Aberrant protein deposition and neurological disease. *J. Biological Chemistry* **53**, 37507-37510.

[(Landschulz et al., 1988)] Landschulz, W. H., Johnson, P. F. and McKnight, S. L. (1988). The Leucine Zipper - A hypothetical structure common to a new class of DNA-binding proteins. *Science* **240**, 1759-1764.

[(Lupas, 1996)] Lupas, A. (1996). Coiled coils: New structures and new functions. *Trends Biochem. Sci.* **21** (10), 375-382.

[(O'Shea et al., 1989)] O'Shea, E. K., Rutkowski, R. and Kim, P. S. (1989). Evidence that the leucine zipper is a coiled coil. *Science* **243**, 538-542.

[(Penel et al., 1999)] Penel, S., Morrison, R. G., Mortishire-Smith, R. J. and Doig, A. J. (1999). Periodicity in a-helix lengths and C-capping preferences. *J. Mol. Biol.* **293**, 1211-1219.

E-mail addresses of the corresponding authors:

`dodson@umist.ac.uk`
`andrew.doig@umist.ac.uk`
`olaf.wolkenhauer@umist.ac.uk`

Appendix: Local approximations and a geodesic distance mesh

A path through the parameter space $\mathcal{G}$ of gamma models determines a curve, parametrized by t in some interval $a \leq t \leq b$, given by

$$c : [a, b] \to \mathcal{G} : t \mapsto (c_1(t), c_2(t)) \qquad (10)$$

and its tangent vector $\dot{c}(t) = (\dot{c}_1(t), \dot{c}_2(t))$ has norm $||\dot{c}||$ given via (9) by

$$||\dot{c}(t)||^2 = \frac{c_2(t)}{c_1(t)^2}\,\dot{c}_1(t)^2 + \left(\psi'(c_2(t)) - \frac{1}{c_2(t)}\right)\dot{c}_2(t)^2 \qquad (11)$$

and the information length of the curve is

$$L_c(a,b) = \int_a^b ||\dot{c}(t)||\,dt \quad \text{for } a \leq b. \tag{12}$$

For example, the curve $c(t) = (t,1)$, which passes through random processes with $t = \tau$ and $\nu = 1 = \mathit{constant}$, has information length $\log \frac{b}{a}$. Locally, minimal paths in $\mathcal{G}$ are given by the geodesics [(Dodson and Poston, 1991)] defined by (9).

In a neighbourhood of a given point we can obtain a locally bilinear approximation to distances in the space of gamma models. From (9) for small variations $\Delta\tau, \Delta\nu$, near $(\tau_0,\nu_0) \in \mathcal{G}$; it is approximated by

$$\Delta s_{\mathcal{G}} \approx \sqrt{\frac{\nu_0}{\tau_0^2}\,\Delta\tau^2 + \left(\psi'(\nu_0) - \frac{1}{\nu_0}\right)\,\Delta\nu^2}\ . \tag{13}$$

As ν_0 increases from 1, the factor $(\psi'(\nu_0) - \frac{1}{\nu_0})$ decreases monotonically from $\frac{\pi^2}{6} - 1$. So, in the information metric, the difference $\Delta\tau$ has increasing prominence over $\Delta\nu$ as the standard deviation reduces with increasing ν_0—corresponding to increased smoothing.

In particular, near the exponential distribution, where $(\tau_0,\nu_0) = (1,1)$, (13) is approximated by

$$\Delta s_{\mathcal{G}} \approx \sqrt{\Delta\tau^2 + \left(\frac{\pi^2}{6} - 1\right)\,\Delta\nu^2}\ . \tag{14}$$

For a practical implementation we need to obtain rapid estimates of distances in larger regions than can be represented by quadratics in incremental coordinates. This can be achieved using the result [(Dodson and Matsuzoe, 2002)] that established geodesic foliations. Now, a geodesic curve is locally minimal and so a network of two non-parallel sets of geodesics provides a mesh of upper bounds on distances by using the triangle inequality about any point. Such a geodesic mesh is shown in Figure 8 using the geodesic curves $\tau = \nu$ and $\nu = \mathit{constant}$, which foliate $\mathcal{G}$ [(Dodson and Matsuzoe, 2002)].

Explicitly, the arc length along the geodesic curves $\tau = \nu$ from (τ_0,ν_0) to $(\tau = \nu, \nu)$ is

$$|\frac{d^2 \log\Gamma}{d\nu^2}(\nu) - \frac{d^2 \log\Gamma}{d\nu^2}(\nu_0)|$$

and the distance along curves of constant $\nu = \nu_0$ from (τ_0,ν_0) to (τ,ν_0) is

$$|\nu_0 \log \frac{\tau_0}{\tau}|$$

In Figure 8 with data from Table 1 we use the base point $(\tau_0,\nu_0) = (20,1) \in \mathcal{G}$ and combine the above two arc lengths of the geodesics to obtain an upper bound on distances from (τ_0,ν_0) as

$$\mathit{Distance}[(\tau_0,\nu_0),(\tau,\nu)] \leq |\frac{d^2 \log\Gamma}{d\nu^2}(\nu) - \frac{d^2 \log\Gamma}{d\nu^2}(\nu_0)| + |\nu_0 \log \frac{\tau_0}{\tau}|. \tag{15}$$

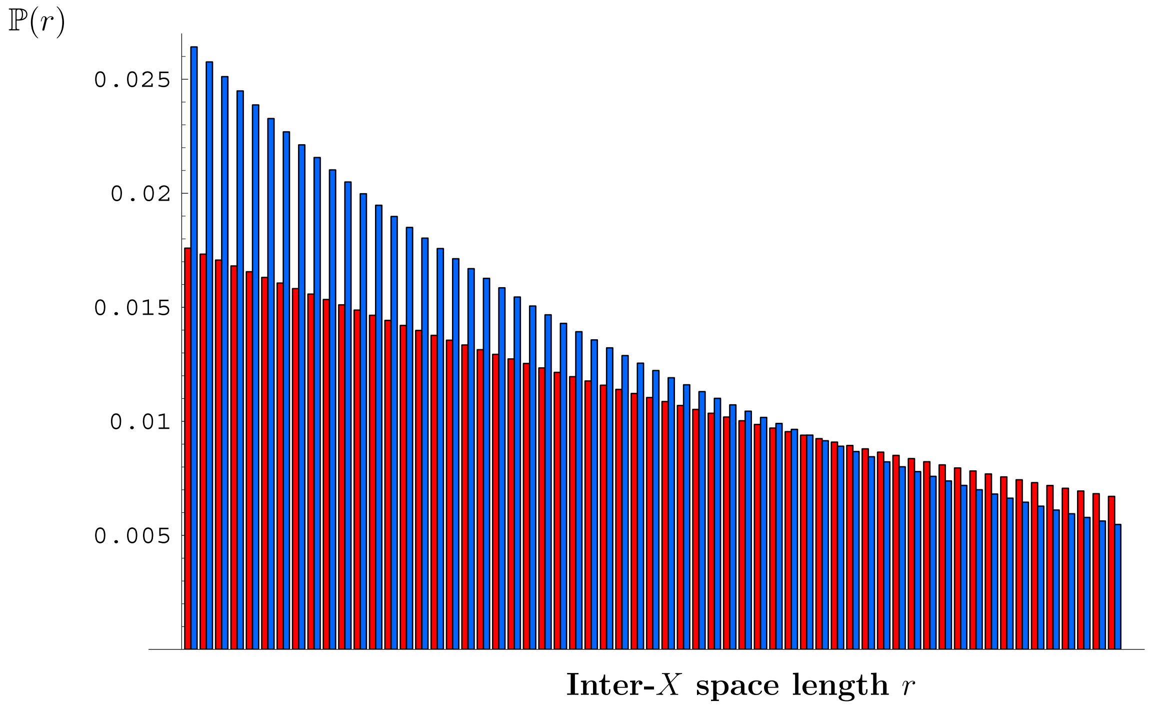

Figure 1: *Probability distribution of space length r between occurrences of amino acid X shown for the range* $0 \leq r \leq 60$ *in a random sequence of* $n = 200$ *amino acids. The mean probability for the occurrence of X is* $p = 0.01$ *(left, red) where we find* $\bar{r} = 46.9$, $\sigma_r = 39.3$,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

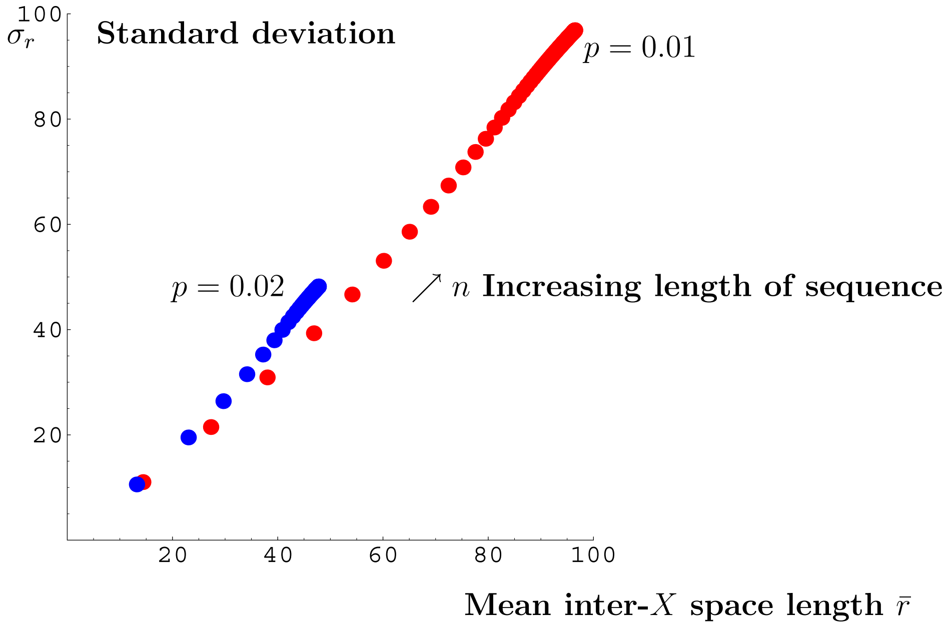

Figure 2: *Effect of sequence length n in random amino acid sequences of length from* $n = 50$ *to* $n = 4000$ *in steps of* 50. *Plot of standard deviation* σ_r *against mean* $\bar{r}$ *for inter-X space length distributions (2). The mean probability for the occurrence of X is* $p = 0.01$ *(right, 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.*

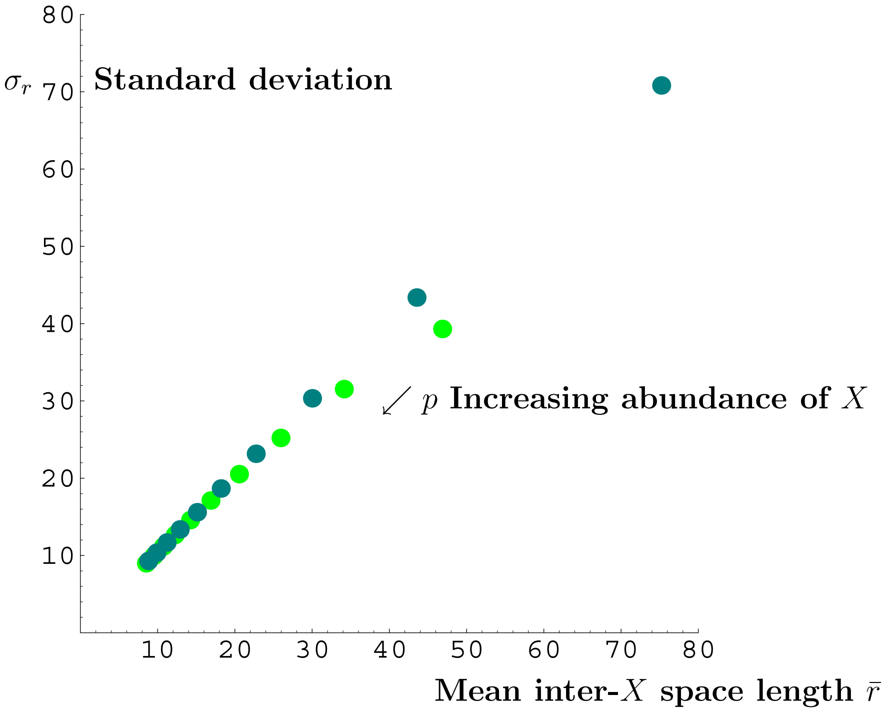

Figure 3: *Effect of relative abundance, probability* p*, over the range* $0.01 \le p \le 0.1$ *in steps of* 0.01*. Plot of standard deviation* σ_r *against mean* $\bar{r}$ *for inter-*X *space length in random sequences of length* $n = 200$ *amino acids (light green) and length* $n = 500$ *amino acids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

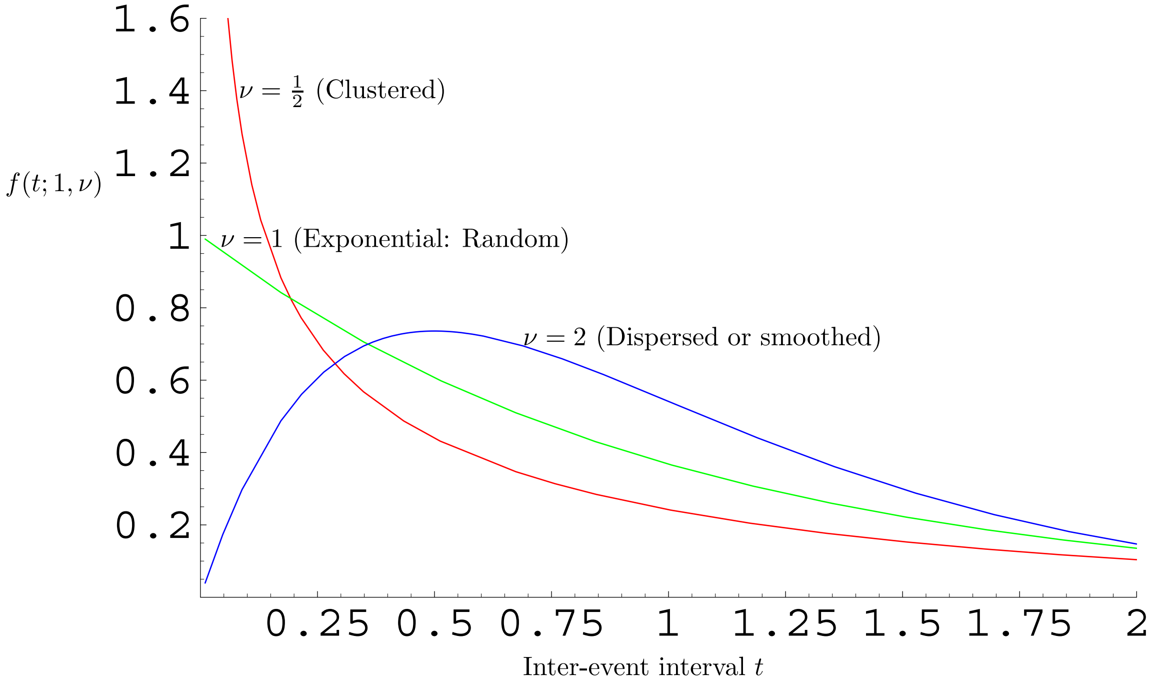

Figure 4: *Probability density functions,* $f(t;\tau,\nu)$*, for gamma distributions of inter-event intervals* t *with unit mean* $\tau = 1$*, and* $\nu = \frac{1}{2},\ 1,\ 2$*. The case* $\nu = 1$ *corresponds to an exponential distribution from an underlying Poisson process;* $\nu \neq 1$ *represents some organization—clustering or dispersion.*

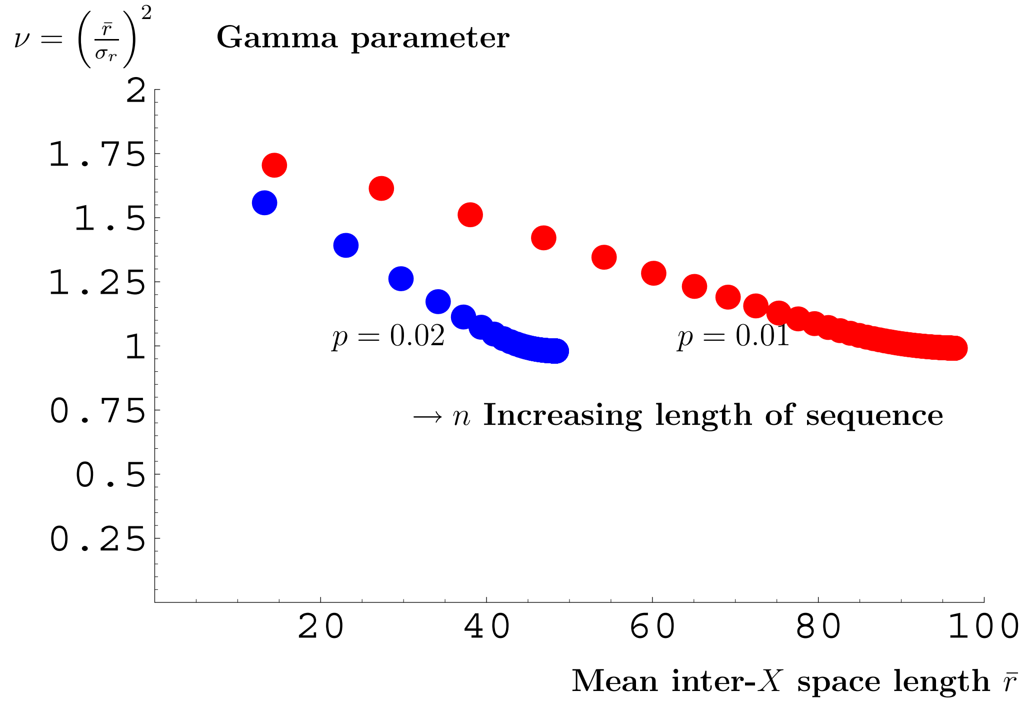

Figure 5: *Effect of sequence length n in random amino acid sequences of length from $n = 50$ to $n = 4000$ in steps of 50. Plot of gamma parameter ν from (8) against mean $\bar{r}$ for inter-X space length distributions (2). The mean probability for the occurrence of X is $p = 0.01$ (right, red) and $p = 0.02$ (left, blue), corresponding to the cases in Figures 1 and 2. We expect that, as $n \to \infty$, so $\nu \to 1$, the random case.*

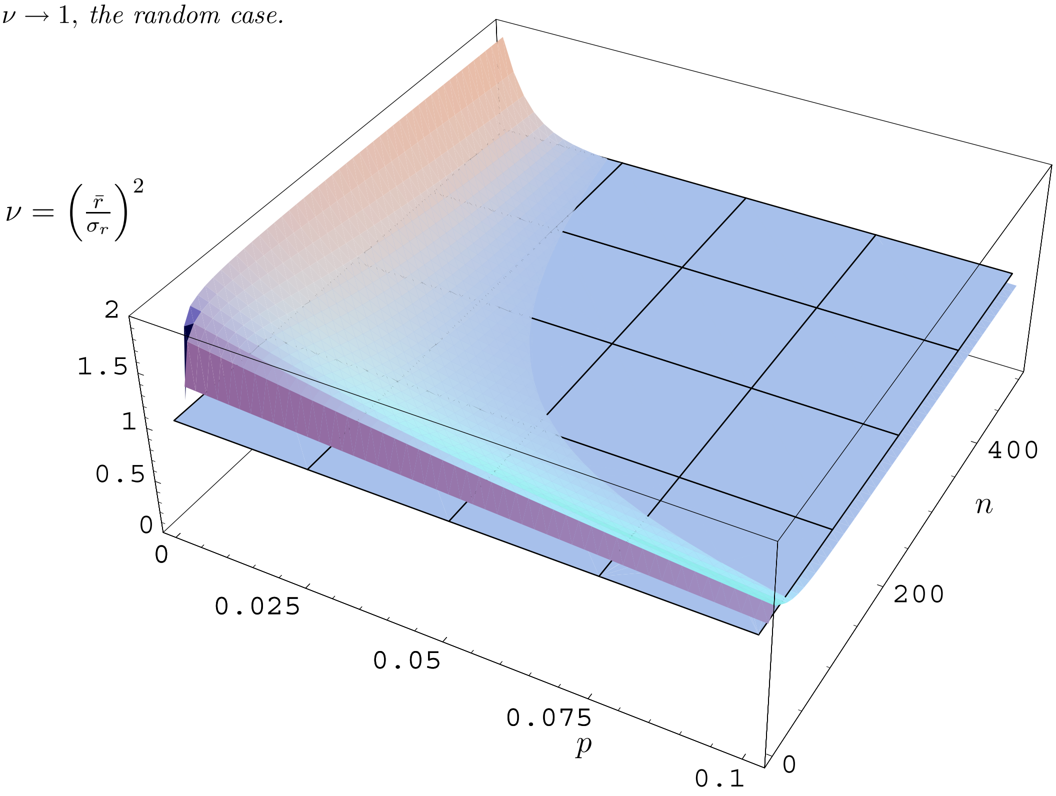

Figure 6: *Effect of relative abundance p, and sequence length n in random amino acid sequences of length from $n = 50$ to $n = 500$. Plot of gamma parameter ν from (8) for inter-X space length distributions (2). This illustrates apparent departures from randomness when we sample the process with finite 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.*

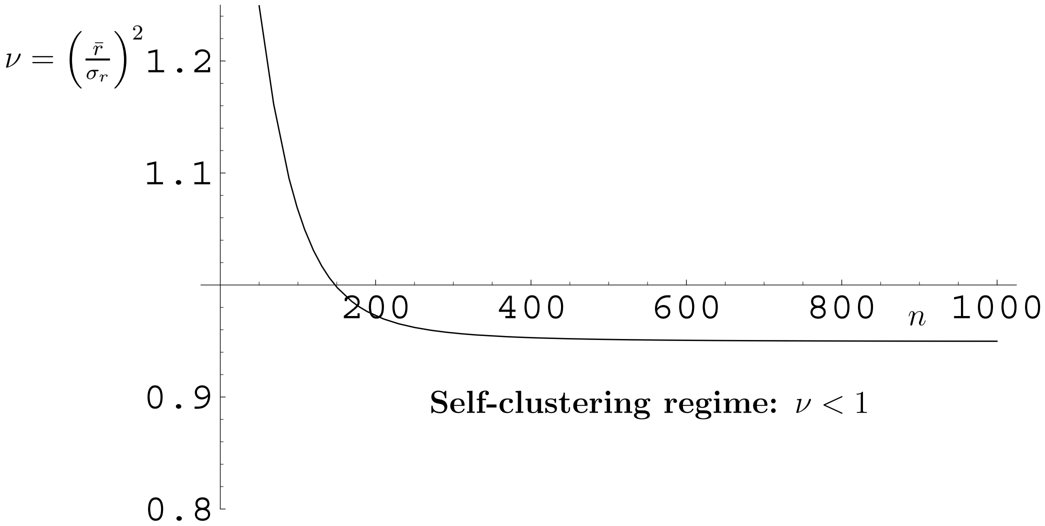

Figure 7: *A section through Figure 6. Effect on gamma parameter ν of sequence length n in random amino acid sequences for fixed relative abundance $p = 0.05$. The value $p = 0.05$ was the observed average relative abundance of the 20 amino acids in 6294 proteins with sequence lengths up to $n = 4092$; these yielded values $0.59 \leq \nu \leq 0.95$, indicating that all of the amino acids exhibited self-clustering.*

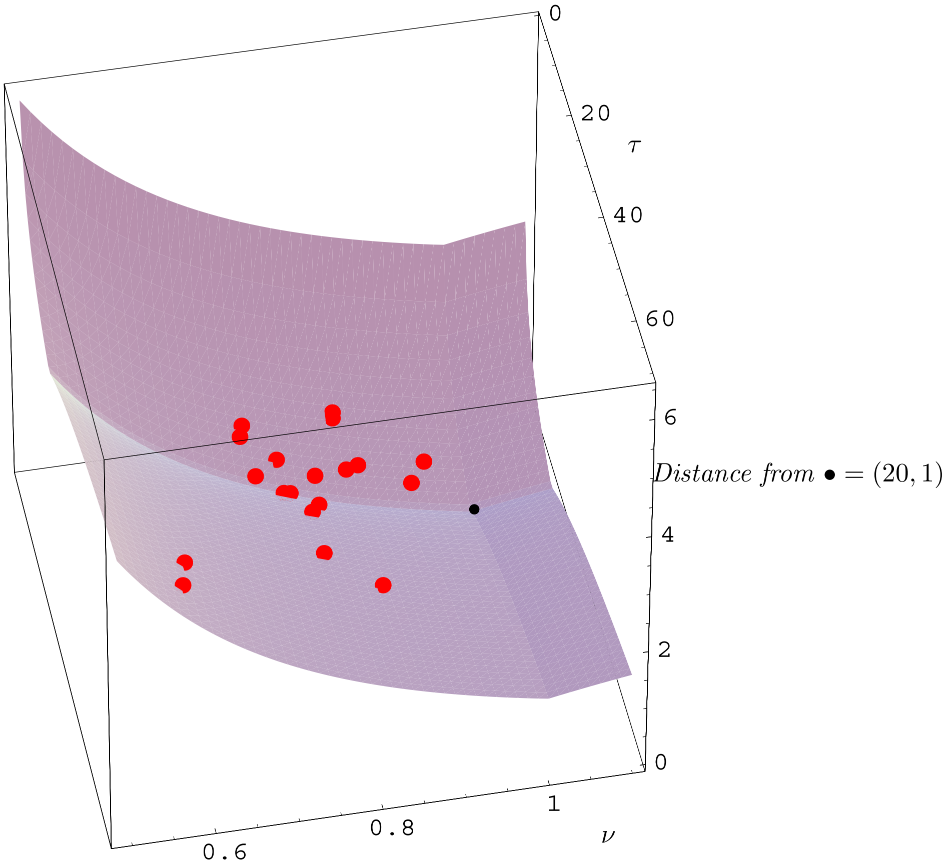

Figure 8: *Distances in the space of gamma models, using a geodesic mesh. The surface height represents upper bounds on distances from $(\tau, \nu) = (20, 1)$, the random case with mean $\tau = 20$, marked with $\bullet$. Depicted also are the 20 data points for the amino acid sequences, all show self-clustering to differing degrees by lying below the line $\nu = 1$, some substantially so.*

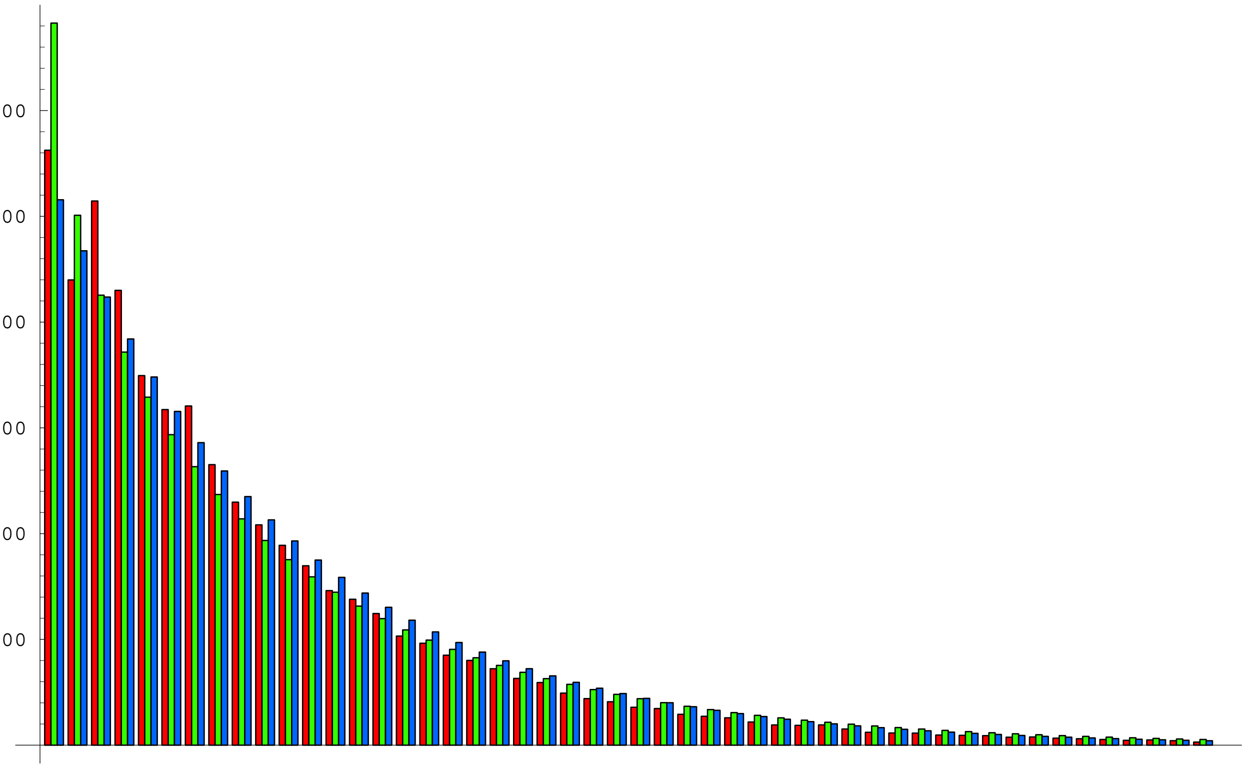

Figure 9: *Histograms of first 50 datapoints for amino acid Leu: observed gaps (left, red) with mean* $\tau = 10$*, maximum likelihood exponential fit (centre, green), and maximum likelihood gamma fit with* $\nu = 0.85$ *(right, blue).*

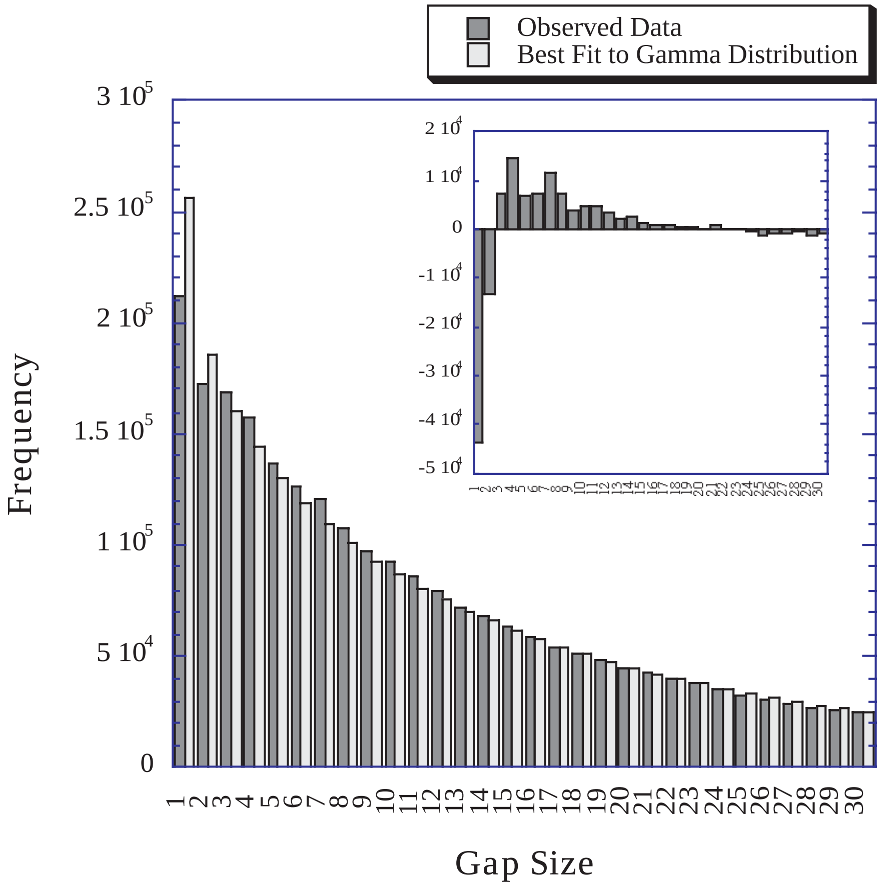

Figure 10: *Histograms of first 30 datapoints for all 20 amino acids; residuals shown inset.*

Amino Acid	Occurrencies	p	τ	Variance	ν
A	163376	0.055	17	374	0.81
C	38955	0.013	55	5103	0.59
D	172519	0.058	16	346	0.77
E	192841	0.065	15	292	0.73
F	133737	0.045	21	554	0.78
G	147416	0.049	19	487	0.74
H	64993	0.022	39	1948	0.79
I	195690	0.066	15	222	0.95
K	217315	0.073	14	240	0.74
L	284652	0.095	10	122	0.85
M	62144	0.021	46	2461	0.85
N	182314	0.061	16	277	0.87
P	130844	0.044	21	587	0.77
Q	116976	0.039	24	691	0.81
R	132789	0.045	21	565	0.78
S	269987	0.091	11	136	0.85
T	176558	0.059	16	307	0.85
V	166092	0.056	17	315	0.93
W	31058	0.010	62	6594	0.58
Y	100748	0.034	27	897	0.79

Table 1: *Number of occurrencies, relative abundance p, mean spacing τ, variance, and maximum likelihood ν for each amino acid, fitting the gap distribution data to the maximum likelihood gamma distribution of equation (8).*