

The Biggest Fragment

Abstract

The situation of one-dimensional fragmentation is examined by several different methods. By considering the number of ways to partition given systems, the probability distributions are calculated as a function of system size, number of fragments, and longest allowable length. These calculations are compared to computer simulations, which are preferred due to their efficiency, and found to agree imperceptibly. Also, one dimensional results are compared to two-dimensional percolation and exponential calculations which give qualitatively different results for all observables. The developed theory of one-dimensional partitioning is compared to the process of chromosomes being partitioned into genes. It is found that no single model accurately predicts this process, and therefore further consideration and modeling of DNA partitioning is necessary.

Exact Solution

In the following sections the situation of one-dimensional fragmentation will be discussed extensively. A one dimensional system may be envisioned as a string of pearls. When partitioning is spoken of it means cutting the string into smaller fragments. Cuts are only allowed between the pearls, resulting in fragments that have integer length. The goal is to find the probability that a given fragment will have a certain rank. The term rank refers to how long the fragment is compared to others; longest – rank 1, second longest – rank 2, etc. These probabilities are calculated with the aid of recurrence relations. They start with a base case and then build up to the full solution. The recurrence relation and calculation of total probabilities is explained in detail below.

Consider a string of length A constituents. This string will be cut into m fragments. So let us define a function, $N(A, m, i)$, which denotes the number of ways to cut a sting of length A into m fragments, with the no fragments bigger than length i . In the following derivation, $i_{\max} = A - m + 1$ is the longest a fragment can be based solely on the system. Here the length of a fragment will be denoted k , while the number of fragments of that length is n_k . Summing over all n_k 's gives the number of fragments m . Summing over the product of n_k and k yields the system length A . Using all the sets of n_k 's that apply to these last two constraints, the value for $N(A, m, i)$ may be determined. If the systems is cut into m fragments the number of ways to arrange those fragments is $m!$. However, redundancies must be accounted for. These redundancies are in the form of n_k . Dividing $m!$ by each $n_k!$ will give the number of ways to arrange the fragments in a given partition. Thus the function takes values as follows:

$$N(A, m, i) = \sum_{\substack{n_k s.t. \\ \sum n_k k = A \\ \sum n_k = m}} m! \prod_k \frac{1}{n_k!}.$$

The equation, however, is not yet useful but it can be manipulated into a recurrence relation. Since $\sum_j^i \frac{n_j j}{A} = 1$ it can be concatenated onto the summation to get,

$$N(A, m, i) = \sum_{\substack{n_k \text{ s.t.} \\ \sum n_k k = A \\ \sum n_k = m}} m! \prod_k^i \frac{1}{n_k!} \sum_j^i \frac{n_j j}{A}.$$

Bringing the second summation to the outside and canceling terms yields,

$$N(A, m, i) = \sum_j^i \frac{mj}{A} \sum_{\substack{n_k \text{ s.t.} \\ \sum n_k k = A \\ \sum n_k = m}} (m-1)! \prod_{k \neq j}^i \frac{1}{n_k! (n_j - 1)!}.$$

The constraints on the second summation can be rewritten so that it is summed over all n_k where $k \neq j$, denoted $\langle n_k, k \neq j \rangle$, and then n_j as well. The constraints implied for $\langle n_k, k \neq j \rangle$ are as follows:

$$\begin{aligned} \sum_{k \neq j} n_k k &= A - n_j j \\ \sum_{k \neq j} n_k &= m - n_j \end{aligned}.$$

Thus the equations is transformed to,

$$N(A, m, i) = \sum_j^i \frac{mj}{A} \sum_{\langle n_k, k \neq j \rangle, n_j} (m-1)! \prod_{k \neq j}^i \frac{1}{n_k! (n_j - 1)!}.$$

Now by making the substitution that $n'_k = n_j - 1$ in both the constraints of the summation and the summation itself produces:

$$\begin{aligned} \sum_{k \neq j} n_k k &= A - n'_k j - j \\ \sum_{k \neq j} n_k &= m - n'_k - 1 \end{aligned},$$

for the constraints and,

$$N(A, m, i) = \sum_j^i \frac{mj}{A} \sum_{\langle n_k, k \neq j \rangle, n'_k} (m-1)! \prod_{k \neq j}^i \frac{1}{n_k! n'_k!},$$

for the equation. After removing the constraint that $k \neq j$, the expression changes to:

$$N(A, m, i) = \sum_j^i \frac{mj}{A} \sum_{\substack{n_{ki} \\ \sum n_k k = A - j \\ \sum n_k = m - 1}} (m-1)! \prod_k^i \frac{1}{n_k!}.$$

The second summation corresponds to $N(A - j, m - 1, i)$ so the recurrence relation becomes:

$$N(A, m, i) = \sum_j^i \frac{mj}{A} N(A - j, m - 1, i).$$

The base cases in this recurrence are that $N(0,0,i) = 1$, and for $A < m$ or $A < 0$ implies $N(A, m, i) = 0$

The next goal is to find the probability that the l^{th} longest fragment will be of length i . We will denote this probability as $P(l, i)$. This will be determined by finding the number of ways this event will occur and divided by the total number of partitions.

Again, assume a one dimensional string of length A constituents will be partitioned into m fragments. Since $P(l, i)$ needs to be calculated there are three categories into which fragments of a partition may fall. They will either be larger than, equal to, or smaller than i . Let l' be the number of fragments longer than or equal to i , k denote the number of constituents strictly longer than i , and then $l' - k$ is the number of constituents equal to length i . Lastly denote A_{small} as the sum of the lengths of all the fragments whose length is less than or equal to i .

Consider the fragments shorter than or equal to i first; they have total length A_{small} . Thus, the total length of fragments strictly smaller than i is $A_{\text{small}} - (l' - k)i$. This chain needs to be cut into $m - l'$ fragments with each fragment shorter than i . This is easily accomplished by the use of the previous function and has the value

$N(A_{\text{small}} - (l' - k)i, m - l', i - 1)$. A similar procedure is carried out for the larger

fragments. $A - A_{\text{small}}$ is the sum of the lengths of the larger fragments. The number of ways to arrange a string of length $A - A_{\text{small}}$ into k fragments, all with length greater than i , is the same as arranging a string of length $A - A_{\text{small}} - ki$ into k fragments with any allowable length; i.e. up to i_{max} . Thus, the number of way to arrange the bigger

fragments is $N(A - A_{\text{small}} - ki, k, i_{\text{max}})$. Since the number of ways to have bigger fragments, the number of ways to have smaller fragments, and how many fragments of size i there are, is known these fragments must simply be arranged to find the total ways of partitioning the string. If all the fragments were considered separately there would be $m!$ ways of arranging these fragments. However since the fragments are collected into three distinct groups we must account for redundancies. The number of ways then to

arrange the fragments is then $\frac{m!}{k!(l' - k)!(m - l)!}$. Now summing over all allowable values

of k , A_{small} and l' , and dividing by the total possible partitions, $N(A, m, i_{\text{max}})$, gives the following value for $P(l, i)$:

$$P(l, i) = \frac{\sum_{A_{\text{small}}}^A \sum_{k=0}^{l-1} \sum_{l'=1}^A \frac{m!}{k!(l' - k)!(m - l)!} N(A_{\text{small}} - (l' - k)i, m - l', i - 1) N(A - A_{\text{small}} - ik, k, i_{\text{max}})}{N(A, m, i_{\text{max}})}$$

This expression yields the exact probability of having the l^{th} longest fragment being length i .

Despite the fact that exact probabilities may be obtained the calculations become untenable for $A > 500$ do to lack of memory needed for the calculations. In cases where A and m are large it is usually more convenient to compute the values of $P(l, i)$ with a

computer simulation. Thus, a simulation was written in C++ and the probabilities of the simulation were compared to the calculated probabilities. The simulation produces random numbers to determine where the cuts for the partitioning should take place. This is done enough times to get accurate statistics and then the probabilities may be from these runs. The calculated and computer generated probabilities were found to be in close agreement as exhibited in the following figure.

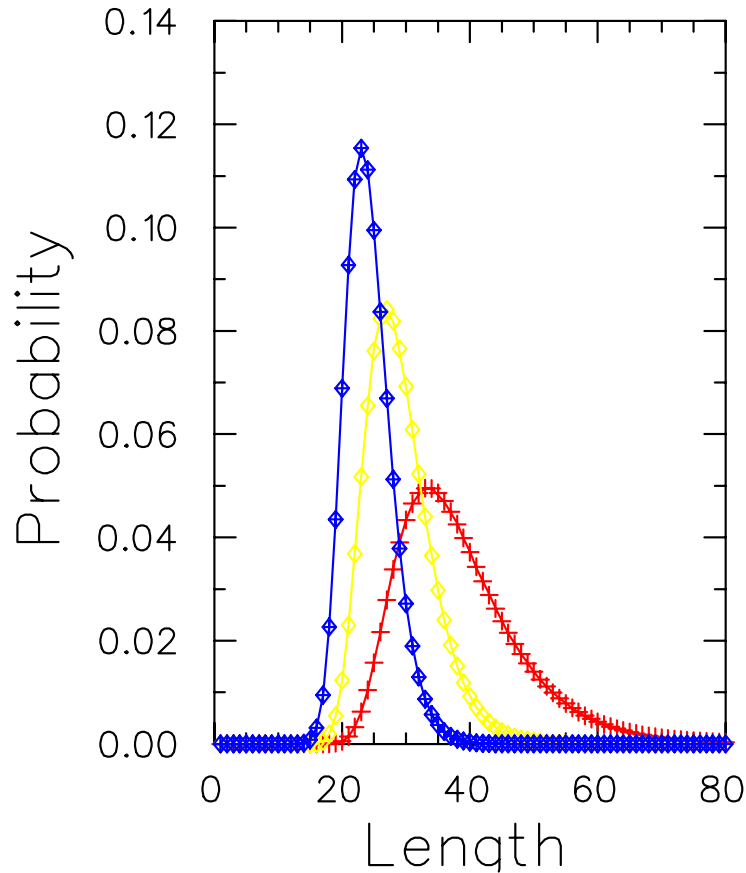


Figure 1: The above figure represents the probabilities compared from simulation and from direct calculations. The solid lines are the simulated probabilities and the symbols are the probabilities from the direct calculation. Both are for a system of length 300 and cut into 30 fragments.

Clearly the simulation and the calculation of exact probabilities yield the same result. For future comparisons the simulation will be used due to efficiency.

Average Length of Longest Fragment

The first relationship examined was between the average fragment size and average size of the longest fragment. The average size of a fragment is A , the length of the chain, divided m , the number of fragments into which the chain is broken. Keeping this value constant and plotting the length of the chain versus the average length of the longest fragment exhibits a logarithmic dependence.

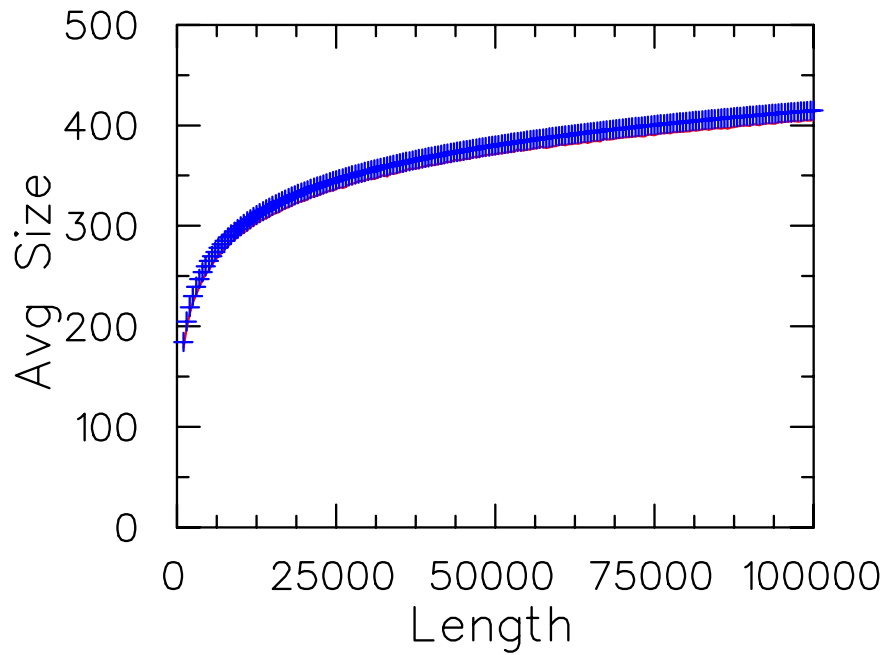


Figure 2: On the x-axis is the total length of the system and on the y-axis is the average length of the longest fragment. The average length of a fragment was held constant at 50.

Similarly, if the length of the system, A , is held constant and the number of fragments varied, the average size of the longest fragment appears to decrease with regularity.

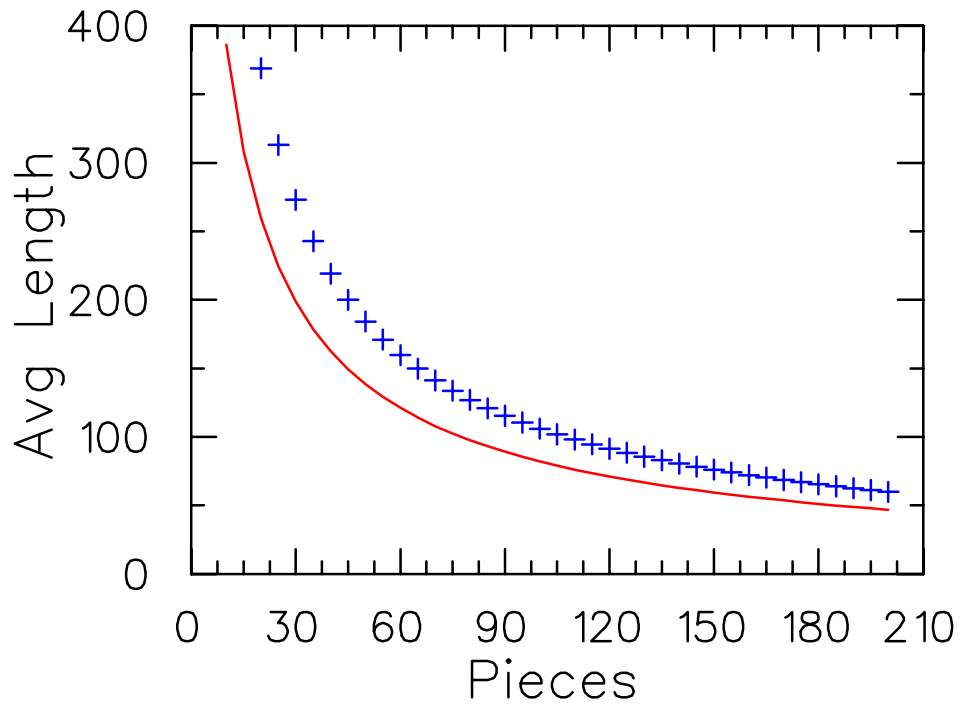


Figure 3: Here the total number of fragments is on the x-axis and the average length of the longest fragment is on the y-axis. The system length is held constant at 2000.

Finally, if the number of cuts is held constant and the size of the system is varied, there appears to be a linear relationship between the length of the longest fragments and the total size of the system.

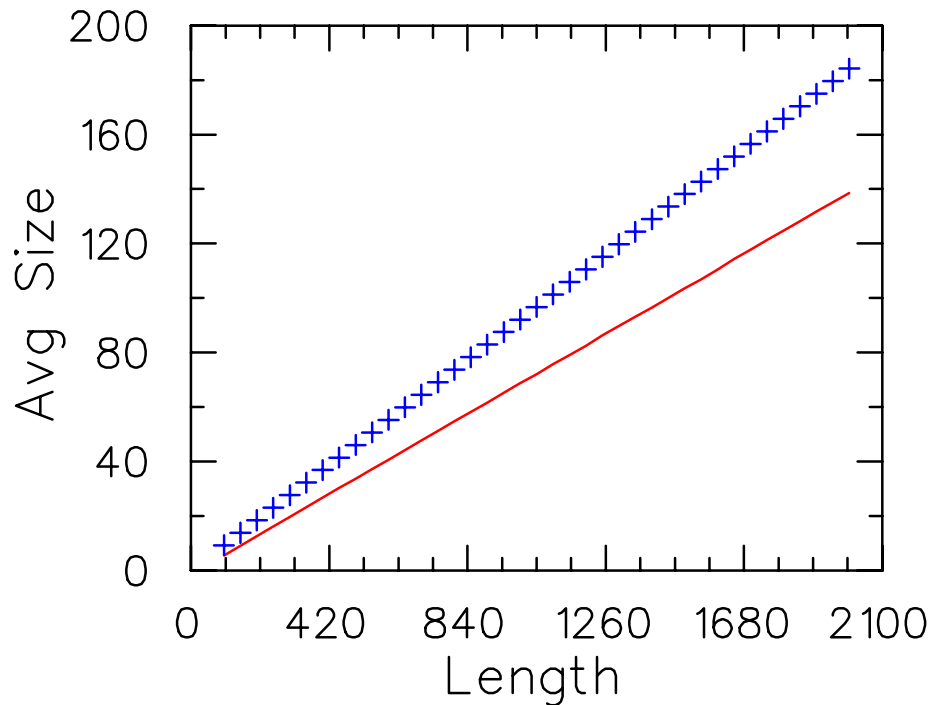


Figure 4: The total length of the system is on the X-axis and the average Size of the longest fragment is on the Y-axis. The number of fragments is held constant at 50.

Although these three relationships are clearly seen they have not yet been accurately explained. Using assumptions and different partitioning models the nature of how these factors affect portioning may be delineated.

Approximation Introduction

Even though the exact probabilities and simulations both give accurate results for random partitioning they can each take large amounts of time to give results. Also, though giving these probabilities they do not explain why the fragments occur in the way that they do. To alleviate this problem, approximations about the partitioning may be made to help gain a better understanding of the situation. A model of the partitioning is created when the exact size of the system, and the number of fragments it will be divided into, is not the initial assumption. Instead the fragments are assumed to occur in a given manner. The three types of model that will be examined here are Poissonian, power law, and exponential models. The theory of each is developed and then compared to the simulation values for random partitioning.

Poissonian Distribution

When dealing with a one dimensional system of sufficiently large size, it may be possible to estimate the probabilities that a fragment will occur using Poissonian statistics. The probability that i is the length of the l^{th} longest fragment approaches a Poissonian distribution. This distribution is based on the average number of fragments of a given size, i , and the average number of fragments bigger than i . Let

$\pi_i = \frac{mN(A-i, m-1, i_{max})}{N(A, m, i_{max})}$, with the exception that $\pi_0 = 0$, be the average number of fragments of length i . It should be noted that $i = 0$ is an allowable length for the

Poissonian Distribution. Then $\gamma_i = \sum_{j=i+1}^{i_{max}} \pi_j$ is the average number of fragments with length

greater than i . When calculating the distribution $\frac{\gamma_i^n}{n!} e^{-\gamma_i}$ is the probability that there are

n with length larger than i . Similarly, $\frac{\pi_i^k}{k!} e^{-\pi_i}$ is the probability that there are k

fragments of length exactly i . Utilizing this information the probability of the l^{th} longest fragment being of length i is,

$$P_{poisson}(l, i) = \sum_{n=0}^{l-1} \frac{\gamma_i^n}{n!} e^{-\gamma_i} \left(1 - \sum_{k=0}^{(l-1)-n} \frac{\pi_i^k}{k!} e^{-\pi_i} \right).$$

The sum is over there being a certain number, n , being at most $l-1$, greater than length i and then multiplies by the probability that there are at least $l-n$ of size i . This distribution can then be compared to the calculated probabilities to test the accuracy of the approximation.

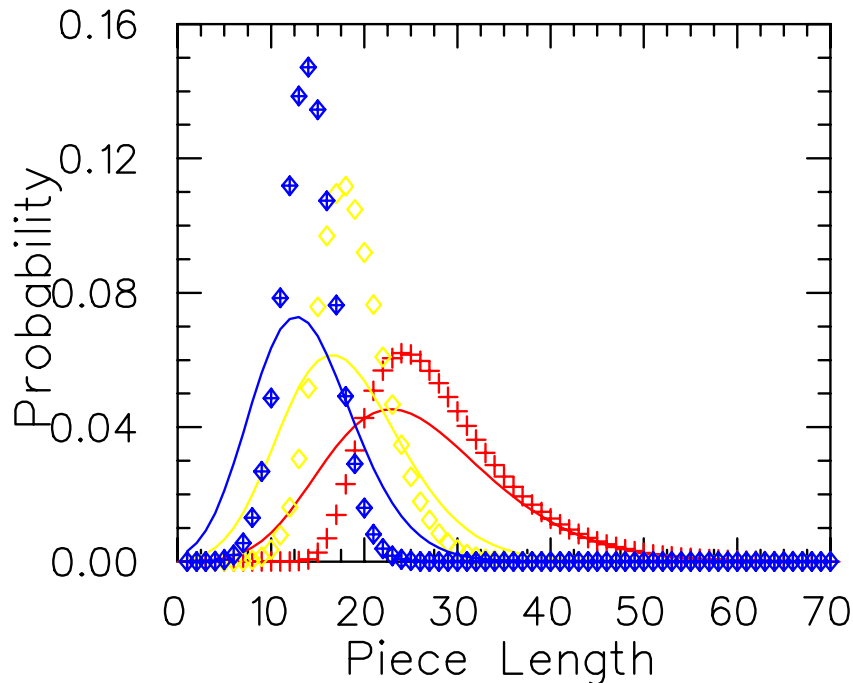


Figure 5: This graph shows the calculated probabilities as compared to those generated by an assumption of a Poissonian distribution. The symbols are calculated and the solid lines are from the Poissonian.

As compared to random partitioning it is clear that the assumption of a Poissonian distribution fails to accurately predict the probabilities of certain fragments appearing.

Exponential Approximation

Another way of looking at the partitioning is with the assumption of an exponential distribution. The probability of having a fragment of length l is equal to the probability of having a fragment of length $l-1$ times the probability of not cutting to the l^{th} power. This suggests that the assumption of an exponential distribution is valid. Also, as the system size tends toward infinity, the exponential solution becomes exact.

Assume that the average number of fragments of a given length i , denote this value $n(i)$, falls off exponentially so that,

$$n(i) = Ce^{-\mu i}.$$

Thus,

$$A = \int_0^{\infty} in(i)di = \frac{C}{\mu^2} \quad \text{and} \quad m = \int_0^{\infty} n(i)di = \frac{C}{\mu}.$$

So, the average length of a given fragment should be $\bar{i} = \frac{1}{\mu}$. Now, the goal is to find an expression for the average size of the longest fragment, denoted A^1 . As a rough estimate say that half the time there are fragments larger than A^1 and half the time it is largest.

This statement is quantified by integrating $n(i)$ from A^1 to infinity and setting it equal to

$\frac{1}{2}$. This will yield an expression for A^1 . So, $\frac{1}{2} = \int_{A^1}^{\infty} Ce^{-\mu i} = \frac{1}{\mu} Ce^{-\mu A^1}$. Solving this

relationship for A^1 yields, $A^1 = \frac{1}{\mu} \ln\left(\frac{2C}{\mu}\right)$. Since, $A = \frac{C}{\mu^2}$ and $\bar{i} = \frac{1}{\mu}$ the longest

fragment can be expressed as a function of only A and \bar{i} . Thus, $A^1 = \bar{i} \ln\left(\frac{2A}{\bar{i}}\right)$.

Percolation Model

A percolation model makes the assumption that the number of fragments of a given size i falls according to a power law. So the assumption is made that $n(i) = Ci^{-\tau}$. Then the average size of a given fragment may be calculated in much the same way as with the exponential model. The result is that an analytic expression may be found for the average length of the l^{th} fragment, denoted $P^{(l)}$, depending only on τ and the total system size. This dependence is as follows:

$$P^{(l)} = \frac{\Gamma\left(\frac{l-1}{\tau-1}\right)}{\Gamma(N)} \cdot \left(\frac{C}{\tau-1}\right)^{\frac{1}{\tau-1}}.$$

The constant C then can be expressed in terms of the length of the system A with the following value:

$$C = (\tau - 2)A \left[\left(\frac{1}{2}\right)^{-(\tau-2)} - A^{-(\tau-2)} \right]^{-1}.$$

This model, along with the exponential model, can now be compared with the average size of the longest fragments to see how well the model explains the random partitioning.

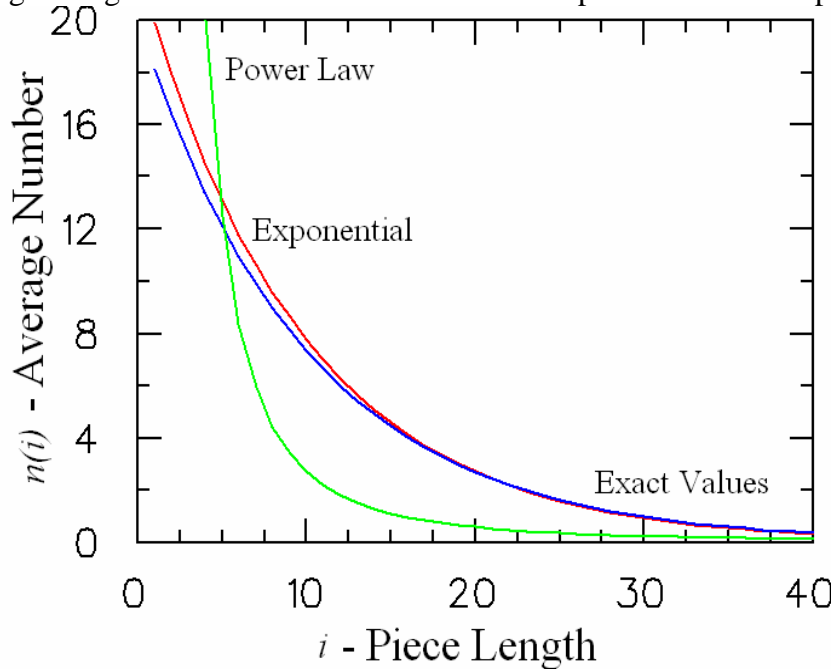


Figure 6: This figure compares the Percolation model and the exponential model with the exact calculated values. Both the models and the exact values were calculated with a system length of 2000 while the number of fragments was 200.

It is clear that the Percolation model yields a separate partitioning form the random case. The model predicts too few fragments of short length and too many fragments of greater length. The Percolation model has the added shortcoming that it always fails for small values of i , because it goes to infinity at zero. Conversely the exponential model explains the random partitioning very well. This accurate prediction then explains the dependence of fragment size upon, length and number of fragments. This is shown on the graphs in the Average Length of the Longest Fragment section. The blue symbols are the points calculated with the formula of the exponential approximation.

Relationship to DNA

Using the information gained about the random partitioning of one-dimensional systems the question arises as to the applicability of these models to physical systems. In particular, to the question of how DNA information, chromosomes, are read into genes. The double-helix of a chromosome can be roughly assumed as a one-dimensional chain that is portioned into exons and introns. The exons are genes, information used to make proteins, while introns are generally considered junk DNA. These genes are composed of codons which give specifications for a particular amino acid to be added to the protein. These codons are the smallest constituent that can be partitioned and determine where exons begin and end. Each codon consists of three base pairs, which can be Cytosine, Adenine, Guanine, Thymine. The underlying question is why the genes are the sizes that they are. Do they obey the random partitioning explored thus far or is there a different system at work. Comparison of the models with actual gene information gives a clear answer.

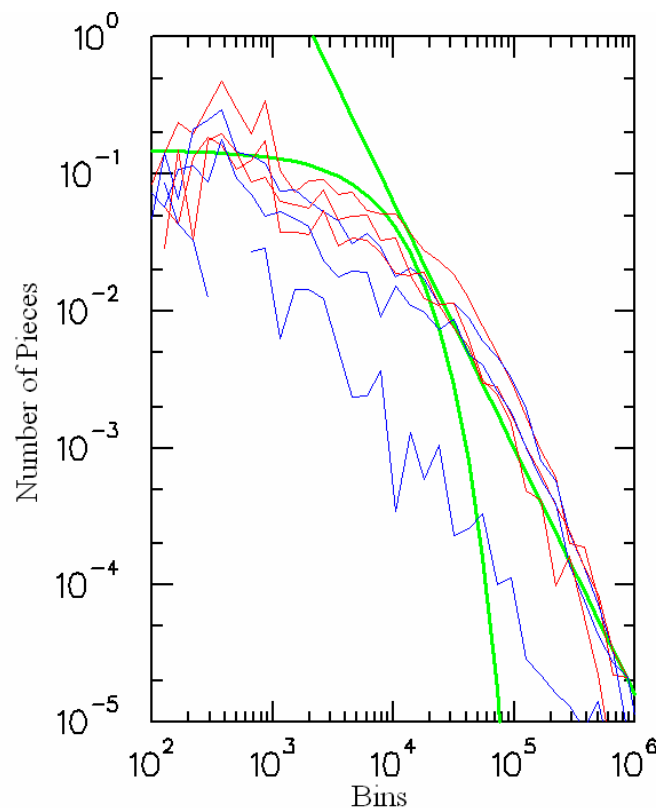


Figure 7: This data is from chromosomes 1, 2, 7, 10, 17, and Y. The jagged lines correspond to the data from the chromosomes. The overlaying bold green lines are attempted fits of an exponential and power law model. The plot is on a log-log scale.

To make sense of the data that comes from the genes they were organized into bins of variable sizes. Larger bins for fragments of larger size, because there are fewer of them and their size differs widely, and smaller bins for the more numerous short genes. These were then divided by the size of each bin to make the graphs. The genes all exhibit similar characteristics: They have a relatively large number of small fragments but still a significant portion of genes of longer length. The plot suggests that neither the

exponential, nor the power law models accurately describe the way in which chromosomes are split into genes.

Conclusions

While the characterization of random partitioning is understood through recursive methods, computer simulations and theoretical assumptions these appear to have little bearing on explaining the way in which chromosomes are partitioned. This could be for several different reasons. Chromosomes partitioning does not appear to have many apparent patterns. The lengths of each chromosome differ widely as do the average length of the genes made from this material. To more accurately explain this process governing principles must be found for the chromosomes themselves. This will lead to better models and reasons as to why chromosomes partitions into genes in a specific way.