Modeling the DNA Space Using Sampling Techniques

Maryam S. Nuser
CIS Dept. Yarmouk University, Jordan.
mnuser@yu.edu.jo

Abstract

DNA oligonucleotides (words) are used for computations as well as other nanoscale applications. These words should hybridize as designed in order to provide correct results. Unfortunately, this is not always the case, and therefore, hybridizations in DNA spaces (all possible DNA words of a given length) were studied to provide reliable DNA-based applications.

In this paper, Hybridizations were modeled as a Gamma distribution, which is a common distribution in reliability theory. Using this model, the consequences of hybridization errors on the reliability of DNA-based applications can be analyzed and understood. Eventually, this should help to select better DNA words and more reliable applications. Words with more A and T bases than G and C bases have fewer cross-hybridizations, and therefore, might be more reliable in applications.

Finally, based on the Gamma distribution, characteristics of DNA spaces are estimated. These characteristics include the mean number of occurrences of cross-hybridizations and the ratio of highly connected words in a DNA space.

Keywords: DNA Computing, Gamma distribution, modeling

1. Introduction

Analyzing large spaces of DNA words might have a great effect on the future of DNA computing. The analysis would provide a large number and a variety of DNA words, which facilitates the selection of best DNA words for a specific application. Large spaces of DNA words are hard to study as a complete system. Comprehensive data collection can be costly, if not completely prohibitive in some instances. Therefore, sampling techniques are used to study such large spaces. These techniques allow decisions to be made on incomplete information. In addition, they summarize large amounts of data.

Using statistical sampling, instead of checking every system component individually, a valid subset of components is tested for the specific property, taking into account any factors that may affect the studied property. An estimate for the complete space will be derived based on the properties found in the subset. The confidence level of these estimates depends on the components in the chosen subset. In contrast to studying the complete space of DNA words, where the cost increases exponentially with the length of the DNA words, statistical sampling of DNA spaces reduces the cost. Thus, when the properties under study remain within a limit, sampling costs may remain bounded, regardless of the system size [1].

Sampling was used in this research to study large spaces of DNA words for the following reasons:

1. Limitations in studying the complete system due to a huge space that grows exponentially with the word length. For example, $4^{20} \approx 10^{12}$ DNA words of length *20* exist in the DNA space of *20*-mers.
2. Enough samples that are chosen using an appropriate sampling method should describe the complete system.
3. Sampling methods had been used for several systems in the literature and provided realistic results [1], [4].
4. The study can be done in a short time and significantly lower cost compared to the study of the complete system. In addition it can provide sufficient results.

The following sections will present two different methods used for generating the samples. This is followed by the methodology used in generating and studying the samples. The results of the computer analysis in conjunction with conclusions based on the analysis will then be discussed.

2. Sampling methods

There are several methods used for sampling. Choosing the inappropriate method will make the conclusions drawn from the samples invalid. Therefore, a sampling method should be chosen taking into account the nature of the system under study and the features of its elements. In addition, the samples should not have any bias to any component of the system. Furthermore, the analysis that will be used and the data that will be analyzed should not be affected by the sampling method. In the analysis of large DNA spaces, two sampling methods were used. These methods were used to generate DNA words from a specific space which will then be checked for their hybridization energies. Following is a description of each of the sampling methods.

2.1 Sampling using random letters

Samples for the DNA analysis were generated using a computer program. The DNA words in the samples were constructed by connecting letters from the DNA alphabet generated randomly such that each letter has a probability P_g of being selected, which is equal for all the genetic alphabet letters. The parameters of the generating system are the sample size n or the number of words, and the DNA word length l. This word length determines the number of letters that should be generated for each word.

Using this method, a sample $S= w_1, w_2, \ldots w_n$ is generated such that $w_i=a_1a_2...a_l$, $i=1,\ldots n$ and l represents the word length, a_j has a probability of being selected $P_g = \frac{1}{|B|}$ where $j=1,2,\ldots|B|$ and B is the set of genetic alphabets, in this case *A, G, C, T* (Adenine, Guanine, Cytosine, Thymine) and $P_g = \frac{1}{4}$.

2.2 Sampling using random words

Random sampling aims to overcome built-in bias by making sure that any member of the population is as likely to be chosen in the sample as any other. The words of the sample, using this method, are generated as above with a probability $P_g = \frac{1}{|B|}$. Unlike the first method, the probability of the word is assumed to follow a multinomial function. The word probability is computed by finding the multinomial factor which is, in this case,

$$\frac{(word\,length)!}{(num\,of\,As)!(num\,of\,Gs)!(num\,of\,Cs)!(num\,of\,Ts)!}$$

This multinomial factor is summed as words are changed. Then each word is checked with probability = $\frac{multinomial\ factor}{space\ size}$. This is done in order to give an equal probability of selection to all words in the DNA space.

In the DNA space, hybridization depends on words. Some distributions of bases in words are more likely than others. For example, there are many more words with equal ratio of all the bases, *A, G, C, T,* than words with just *Gs* and *Cs*. Therefore, the multinomial function is used to give words with mostly one or two bases a good chance of being selected.

3. Samples

To estimate the parameters of DNA spaces, samples of DNA words were constructed ranging from *100* to *9500* words. Samples of larger sizes were time consuming, and huge sizes were hard to achieve. The length of the words used ranged from *6* to *20*. The samples that were obtained from the spaces of DNA words of length *6, 7,* and *8* were chosen such that the summation of the size of the samples is approximately equal to the space size, $\sum_{i=1}^{n} |S_i| \approx N$ where S_i is sample number i. By doing that, the probability of studying all space members increases. Initially, samples of equal size were analyzed, which was followed by analyzing samples of different sizes. The sample sizes were chosen once to be equal ($|S_i|=|S_j|$, $1 \le i,j \le n$), and in another way to be different in order to detect any relation between the sample and space sizes.

The samples were chosen using the above methods and according to the aforementioned strategy for the following reasons:

1. To eliminate any bias that might occur from any other sampling method, and thus having the ability to describe the system correctly.
2. To use sampling methods that are possible and not time consuming.
3. To find a relationship between the sample parameters and the space parameters, and any effects of the sample size.

4. Methods for Comparing Words

The results of analyzing the system depend mainly on the hybridization energy between DNA words. The energy value between two DNA words determines whether the hybridization is favored or not. Therefore, the energy is compared to a threshold value that determines whether the two DNA words hybridized, i.e. whether the corresponding nodes have an edge between themselves.

The energy between words is computed using a dynamic programming algorithm [2] that depends on the Nearest Neighbors energies evaluated in [5]. In reality, the probability of hybridization between two oligonucleotides, P_{ij}, is the Boltzman distribution [3] and can be computed by:

$$P_{ij} = \frac{\sum_k \exp(\frac{-\Delta G_{ij}^k}{RT})}{\sum_i \sum_j \sum_k \exp(\frac{-\Delta G_{ij}^k}{RT})} \text{ , (1),}$$

where k is all possible duplex configurations, R is the gas constant, T is the temperature, and ΔG_{ij} is the change in the free energy. In this research, not all possible configurations of hybridization are considered. The free energy of the configuration with the lowest value is computed. Therefore, the parameter k is ignored and in order to approximate the NCH words the lowest energy configuration is compared to a threshold value such that if $\Delta G_{ij} < \Delta G_{threshold}$ then, the hybrid pair is verified to non cross-hybridize. The threshold value used here is equal to *-6.2* which yields a close approximation to reality.

Using the above information, the node degree (the number of crosshybridizations) for each DNA word in the sample can be calculated and the results can then be analyzed. Two methods were used to determine which words should be involved in the calculation.

4.1 Sample comparison

For a sample $S=\{w_1,...w_n\}$ of size n that is drawn from the DNA space $B^l = \{ W_1, W_2, ... W_N \}$. Let $E(w_i,w_j)$ be the hybridization energy function between the two words $w_i, w_j, 1 \le i,j \le N$.

Define $H(w_i): S \times S \rightarrow S$ by:

$$H(w_i) = \{ w_j \quad | E(w_i,w_j) \le t \}. \quad \textbf{(2)}$$

Then the number of connections C for each DNA word can be defined as:

$C(w_i) = |H(w_i)|$ where $|H(w_i)|$ is the number of words in the set $H(w_i)$. Note that C represents the node degree in the graph representing the sample S. Using this comparison method, the function H is evaluated against all elements of the sample S.

4.2 Space comparison

This method is similar to the previous one. The only difference is that when a DNA word is compared with the other words to check its connectivity, it is compared with the DNA words in the complete space B^l instead of being compared with only words that are in the sample. Therefore, for a sample $S= \{w_1,...w_n\}$ that is drawn from the population $B^l = \{ W_1, W_2, ... W_N \}$, let $E(w_i,w_j)$ be defined as above.

Define $H'(w_i): S \times B^l \rightarrow B^l$ *by* :

$$H'(w_i) = \{w_j \mid E(w_i,w_j) \le t\} \quad \textbf{(3),}$$

where B^l is the space of all words of length l.

Then $C(w_i) = |H'(w_i)|$ where $|H'(w_i)|$ is the number of words in the set $H'(w_i)$.

5. Methodology

Two different sets of samples were generated for the study. The first set was generated using a C++ program, which implemented sampling using random letters, and the inputs were the sample size and the word length. The output of the program are words that were generated one letter at a time. Upon generating each word, it is checked against already generated words in order to make sure that there is no repetition of DNA words in the sample. If word repetition occurs, it is ignored and another word is generated.

The rest of the samples were generated via sampling using random words, by a C++ program which also has the sample size and the word length as inputs. This program divides the sample space into regions such that words in each region have an equal ratio of each DNA base. As for example, *AAAAGG*, and *GAAGAA* belongs to the same region. Samples are generated by selecting words from the space regions with equal probability, therefore, *AAAAGG*, and *GAAGAA* have the same probability of being selected. These samples contain random DNA words with no repetitions.

The previous samples, generated by the above methods, were analyzed using a C++ program which has as its inputs the sample size, the name of the file that contains the sample, and a threshold value. The threshold is compared with the energy produced by each pair of DNA words in order to determine the existence/absence of an edge (crosshybridization) between the two nodes that correspond to these two words. The program evaluates the hybridization energies of each word in the sample either by comparing it with the complete population or by comparing it to only the words in the same sample. The comparison produces the node degree of each node in the graph that represents the sample.

6. Results

Graphs that represent DNA spaces were analyzed with respect to their node degrees. The node degrees were used as an input to two different software applications (ProModel "STAT::FIT", and Arena "Input Analyzer"). The

reason for using two applications is that each one has its own assumptions, different methods of fit, and different data tests.

The following sections will present some of the experimental results of analyzing different samples for different spaces using both of the above software, followed by a Gamma model of DNA spaces. Based on that model, a prediction of the characteristics for the complete DNA space of words of length *20* will be generated.

6.1 Experimental Results

The complete DNA spaces for words of lengths *6, 7*, and *8* (*6, 7*, and *8*-mers) were checked for their node degrees. The degrees were analyzed using the Input Analyzer software, and the results are shown in table 1. The first column lists the distributions that are supported by the Input Analyzer. The cells values in the remaining columns represent the square error generated from fitting the DNA space to the specified distribution where smaller values of square error indicate a better fit. Notice that the Lognormal is the best fit for the *6*-mers space, the Weibull for the *7*-mers, and the Gamma for the 8-mers. Samples from the above spaces and from larger spaces were analyzed for the best fitting distribution.

Table 1
The square error generated from fitting the DNA spaces of length 6,7, and 8 to all distributions supported by "Input Analyzer. Smaller error values indicate a better fit.

Fit distribution	6-mers	7-mers	8-mers
Weibull	0.0306	0.0114	0.271
Gamma	0.0554	0.0776	0.00445
Exponential	0.0705	0.0865	0.0235
Erlang	0.0705	0.0865	0.0235
Lognormal	0.0217	0.0948	0.0132
Beta	0.059	0.0219	0.00479
Normal	0.249	0.358	0.119
Triangular	0.287	0.473	0.149
Uniform	0.314	0.494	0.173

6.2 Modeling DNA Space as a Gamma Distribution

The experimental results suggested different fitting distribution for each complete DNA space. These distributions were Gamma, Weibull, and Lognormal, which are special cases of the general Gamma distribution. Thus, the analysis below will depend on fitting the data to the Gamma distribution, as a parent distribution for the Weibull, and Lognormal.

The Gamma distribution has two parameters: λ that defines the shape of the Gamma, and α is the scale parameter. These parameters differ with the DNA space size, and determine which special case of the Gamma is the right distribution. Therefore, the following section approximates the values for these parameters of a DNA space as a function of the parameter values for several samples.

6.3 Formulas for Predicting Characteristics of DNA Spaces

Two methods were used for predicting Space characteristics. The first method predicts characteristics of DNA spaces based on samples of these spaces. The second method predicts characteristics of DNA spaces depending on available characteristics of small spaces. The formulas for the two methods were generated using the Minitab software.

The Minitab software works by doing a regression analysis on the input data. It takes as an input X-values, Y-value, and several functions that are expected to be in the resulted formula. As for example, to predict the mean formula in terms of the word length, the lengths *6, 7*, and *8* were input as X-values. The Y-values were *2.19556, 24.3*, and *202.542*, which are the means of the associated spaces. Because the Minitab software does a linear regression and the expected relation seems to be exponential, the exponential function was input to the software to try to find an exponential relation between the word length and the space mean. In all of the following relations, the exponential, logarithm, inverse ($\frac{1}{f}$ is the inverse of *f*), and square functions were tried to find formulas with a good fit.

For DNA words of length *6, 7*, and *8*, several samples were drawn such that the summation of the sample sizes is approximately equal to the space size. At first, the sample sizes were equal such that each sample S_i has a size of $|S| = \frac{N}{c}$, where *c* is the number of samples, and *N* is the space size. Then, the sample sizes were different with a summation that is approximately equal to the space size. This increases the probability for each member in the space to be included in one of the samples.

The values of the sample parameters were calculated. Results of these values indicate

that sampling using random letters, with space comparison method, gives the best approximation of all sampling and comparison methods discussed earlier.

In one experiment, *16* samples each of size *1000* were taken for the space of DNA words of length seven. All samples were equal in size. Each sample was input to the "Input Analyzer" to find the best gamma fit. Table 2 shows the average of the parameter values of the *16* samples compared with the values for the whole space of *7*-mers. The parameters include the shape parameter (λ), scale parameter (α), mean, standard deviation, minimum, and maximum.

Table 2
The average values of the characteristic parameters resulted from fitting 16 samples of DNA words of length 7 each sample is of size 1000 with the Gamma distribution, and the values of the parameters for the whole space of 7-mers.

values calculated	Lambda	Alpha	Mean	Std	min	max
Average of samples	74.33	0.33	24.01	40.4	0	279.5
Complete space	73.30	0.33	24.30	40.9	0	314.0

Note that the average sample characteristics are almost the same as the space characteristics which makes it easy to predict the space characteristics from a sample using this method. The characteristics are represented by the above parameters.

In another experiment, different sample sizes were analyzed. The summation of the sample sizes accumulates the space of the *7*-mers. The results again indicate that the characteristics are approximately the same for the sample and the space.

Using the previous comparison method (space comparison) is time consuming for large DNA spaces, therefore, the need to use sample comparison method arises. More than *30* samples were drawn using sampling using random letters technique, and they were compared using sample comparison. The parameters of the distribution were studied and the following relations were generated:

Let N be the space size, where $N=L^l$, L is the number of letters in the alphabet, l is the word length, and n_i is the size of sample number i for $i=1,...k$.

Define the following parameters for each sample s_i:

min_i : the minimum node degree for the sample s_i.

max_i : the maximum node degree for the sample s_i.

λ_i :the λ value when fitting the sample to the Gamma distribution.

α_i :the α value when fitting the sample to the Gamma distribution.

μ_i : the mean of the node degrees for the data in sample s_i.

δ_i : the standard deviation of the node degrees for the data in sample s_i.

Then, the parameters for the whole space can be approximated using the average, minimum, or maximum of the sample parameters as follows:

Space minimum value

$$= \min\{\min_i, \forall i = 1,2,...,k\} \quad (4)$$

Space maximum value

$$= \max\{(\max_i \times \frac{N}{n_i}), \forall i = 1,...,k\} \quad (5)$$

Space mean =average

$$\{(\frac{N}{n_i} \times c1 \times \mu_i), \forall i = 1,...,k\} \quad (6)$$

Space STD= average {

$$(\frac{N}{n_i} \times c2 \times \delta_i), \forall i = 1,...,k\} \quad (7)$$

$$\lambda = \text{average } \{ (\frac{N}{n_i} \times c3 \times \lambda_i), \forall i = 1,...,k\} \quad (8)$$

$$\alpha = \text{average } \{ (\frac{\alpha_i}{c4 \times \frac{N}{n_i}}), \forall i = 1,...,k\} \quad (9)$$

Values for c_1, c_2, c_3 and c_4 are *6.4, 1.52, 12.8*, and *1.125* respectively. These values were estimated using the Minitab software and empirically based on the *6, 7*, and *8*-mers analysis to best represent the correct results. The ratio $\frac{N}{n_i}$ is because the sample words were compared to the sample itself and not with the complete space, therefore, the degree of the word when compared to the complete space should be a ratio of its degree in the sample. One sample can give an approximation to the complete space, but the

more samples the best the approximation, therefore, the larger the *k*, the better the results. Figures 1 and 2 show the relation between the mean (maximum) node degree and the sample size for samples of DNA words of length *7*.

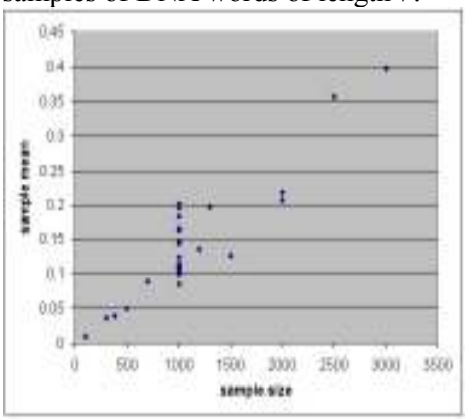

Figure 1
The relation between the sample size and the sample mean for samples of DNA words of length 7.

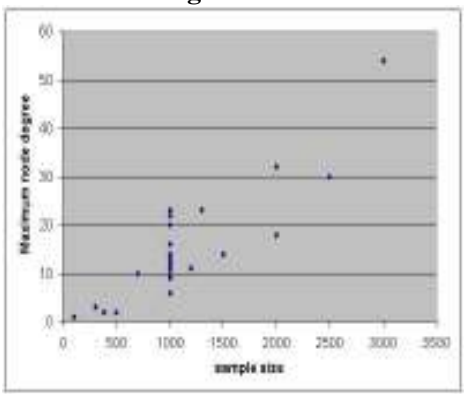

Figure 2
The relation between the sample size and the maximum node degree for samples of DNA words of length 7.

The second type of formulas is also generated using the Minitab software based on the spaces of DNA words of length *6, 7,* and *8*. Formulas for predicting the maximum node degree, the minimum degree, and the lambda value of the gamma distribution are shown below. The resulting formula for predicting the mean was:

$$Y = \exp(a - \frac{b}{l}) \quad (10)$$

where *a=18.8201*, *b=108.368*, and *l* is the word length. The R-sq value which indicates the goodness of the fit (the bigger R-sq value, the better the fit) is 99 %

Figure 3 shows a chart of formula 10 for the *6,7*, and *8*-mers spaces along with the mean node degree resulted from the simulation.

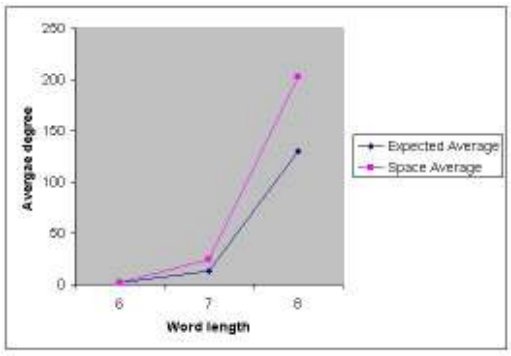

Figure 3
The expected mean of the node degrees predicted using formula 10 along with the values resulted from the simulation

Using the same regression analysis by Minitab, the predicted formula for the maximum node degree was:

$$Y = \exp(a + b \times l^2) \quad (11),$$

where *a=-2.386, b=0.15197*, and *l* is the word length. The R-sq value was *79.8*%. Figure 4 shows a chart of formula 11 for the *6,7*, and *8*-mers spaces along with the maximum node degree resulted from the simulation.

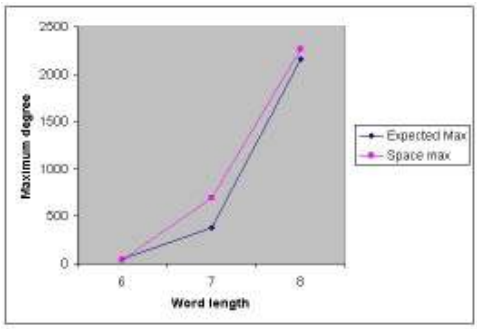

Figure 4
The expected maximum node degrees predicted using formula 11 along with the values resulted from the simulation

The lambda parameter (λ) was predicted to fit the formula:

$$Y = \exp(a - \frac{b}{l}) \quad (12),$$ where *a=20.178*, *b=112.818*, and *l* is the word length.

6.4 Predicting Characteristics of the 20-mers

Results from the previous sections were used to determine the characteristics of the space of DNA words of length *20*. Since it is difficult to compare the sample with all *20*-mers, space comparison was excluded from the analysis. Several samples of different sizes were used to calculate the characteristics of the space of *20*-

mers. Table 3 shows the minimum, maximum, and average node degree of several samples.

Table 3

The maximum, minimum, and average node degree resulting from the computer analysis of several samples from the 20-mer space.

Sample Size	Minimum	Maximum	Mean
1000	0	266	31.5
9500	9	2669	297.1
1000	0	250	32.444
1000	0	254	31.074
1000	0	220	31.786
6500	2	2043	182
6500	5	2082	175
8000	13	2540	311
8000	22	2418	294
2000	6	597	80.2

Based on the previous formulas (4-9) the predicted values of the parameters are shown in table 4. Although the R-sq value for the predicted formulas indicates a good fit, the results of the *20*-mers space are not accurate. This is due to the limited number of spaces studied and the difficulty in studying larger spaces. Figure 5 shows the Gamma distribution that fits the histogram of the data.

Table 4

Predicted parameters for the space of DNA words of length 20. Two methods were used: method 1 predicts the values based on samples from the 20-mer space, and method 2 predicts the parameters based on the 6,7, and 8-mers spaces

Parameter	Type1 (sample based)	Type2 (Space based)
Lambda	2e+12	2057701.21
Max	3.52182E+11	2.31029E+25
Min	2	NA
Mean	2.3132E+11	661126.1845

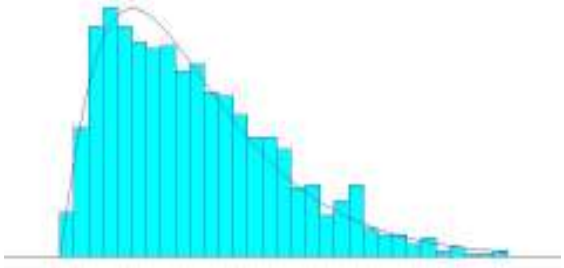

Figure 5

The histogram for a 20-mer sample and the Gamma fitting over the histogram, where the X-axis represents the node degree and the Y-axis represent the frequency of the occurrence of the degree in the population

6.5 Characteristics of a DNA space when modeling it as a Gamma distribution

Based on the previous results of modeling the DNA space as a gamma distribution, and knowing that this distribution is positively skewed, one concludes that words with large number of crosshybridizations occur less in DNA spaces. On the other hand, words with low number of crosshybridizations occur with larger frequency.

In addition, the mean number of crosshybridizations (node degree) can be easily estimated using the parameters of the corresponding Gamma distribution. Furthermore, these parameters which depend on the word length, determines the frequency of words with very low number of crosshybridizations. This frequency is large for some populations and small for others.

Finally, the Gamma distribution measures the reliability of the population which is, in this case, all possible combinations of DNA words of a given length. This helps researches predict the reliability of DNA words that are used to encode problem instances, and therefore gives an indication of the probability of error in the solution.

7. Discussion of the results

The number of cross-hybridizations, which is represented as the node degree, that each DNA word might have with the other nodes in the DNA space was modeled as a Gamma distribution. The Gamma distribution has a maximum value that is followed by a drop in the value until it approaches zero. The simulation results show that the words with the highest frequencies of occurrence are associated with lower node degrees, and the lowest frequencies are associated with the highest node degrees. In other words, the highest frequencies are associated with fewer cross-hybridizations and the lowest frequencies are associated with higher cross-hybridizations.

The occurrence of a higher number of DNA words with low cross-hybridizations and a very low number of high cross-hybridizations is due to the Nearest Neighbor model of the DNA. This model shows that in a DNA word, *G* and *C* bases have a lower free energy than *A* and *T* bases do. In addition, *G* and *C* have a lower free energy as neighbors than if they have *A* or *T* as a

neighbor. The natural tendency of a system is to attain the lowest free energy possible. And since words that have mostly *G*s and *C*s have lower free energies, they tend to hybridize more than other words. This indicates that the number of words that have a large portion of *G*s and *C*s with these bases occurring as neighbors is the portion of the DNA space that have high cross-hybridizations. This proportion is small in the DNA space due to the fact that these words not only have a higher proportion of *G*s and *C*s, but also these *G*s and *C*s are neighbors of each other.

Modeling the number of cross-hybridizations of DNA words as a Gamma or as its special cases appears to be reasonable. This is because cross-hybridization measures the reliability of the DNA words, i.e. whether the hybridization between two DNA words is a perfect W-C complement or with cross-hybridizations. And modeling reliability is one of the uses of the Gamma distribution.

8. Conclusion

Based on the previous analysis the following conclusions can be drawn:

1. Hybridizations between DNA words in complete DNA spaces and in samples of these spaces fit a Gamma distribution.
2. Reliability of DNAC is a function of crosshybridization. The degree of crosshybridization in the model indicates that reliable DNA-based applications can only use a small portion of the whole DNA space.
3. The models for 6,7, and 8-mer spaces were extrapolated to 20-mers with limited success. Though 20-mer estimates were consistent with experiment, the error was large.
4. DNA spaces are positively skewed which indicates a small number of highly connected words in these spaces. The DNA model might be used to understand and better select DNA words that do not crosshybridize. In addition, the model helps to understand the reliability of abiotic DNA applications.
5. The mean number of crosshybridizations can be easily obtained for a DNA space based on the values of its distribution parameters.
6. The frequency of words with low number of crosshybridizations varies with the word length, but in general is more than words with high number of crosshybridizations.

References:

[1] Celso L. Mendes and Daniel A. Reed, "Monitoring Large Systems via Statistical Sampling" International Journal of High Performance Computing Applications, May 2004, vol. 18, iss. 2, pp. 267-277(11) SAGE Publications

[2] Deaton, R. and J. Chen and H. Bi and J. A. Rose, "A software tool for generating non-crosshybridizing libraries of DNA oligonucleotides", DNA Computing: 8th International Workshop on DNA-Based Computers", pp. 252-261

[3] Deaton, R. and J. W. Kim and J. Chen, "Design and test of non-crosshybridizing oligonucleotide building blocks for DNA computers and nanostructures", Appl. Phys. Lett.,2003, vol. 82,pp. 1305-1307.

[4] Gregory D. Speegle, Michael J. Donahoo "Using statistical sampling for query optimization in heterogeneous library information systems" Proceedings of the 1993 ACM conference on Computer science , 1993 pp. 475 - 482.

[5] SantaLucia, Jr., J.,"A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics", pnas,1998, vol. 95,pp.1460-1465