# Visualization of $K$-Tuple Distribution in Procaryote Complete Genomes and Their Randomized Counterparts[1]

Huimin Xie
Department of Mathematics
Suzhou University, Suzhou 215006, China
szhmxie@pub.sz.jsinfo.net

Bailin Hao
T-Life Research Center
Fudan University, Shanghai 200433, China
hao@itp.ac.cn

## Abstract

*A few years ago we developed a simple scheme to visualize the string composition of long DNA sequences in terms of two- and one-dimensional (2D and 1D) histograms. While the patterns in the 2D histograms have been well understood, the structure of the 1D histograms has not been analyzed in details. It turns out that the structure of the 1D histograms of the genomic sequences and their randomized counterparts varies significantly depending on the g+c content of the genomes. In particular, the 1D histograms of some randomized sequences may show rich structure, a seemingly anti-intuitive result.*

*Three approaches are used to explain the phenomenon: (1) Monte Carlo simulation, (2) exact computation by using the Goulden-Jackson cluster method, and (3) a Poisson approximation method. The multi-modal phenomena in $K$-histograms are well elucidated by the last approach.*

## 1. Introduction

The availability of an ever growing number of complete genomes of various organisms enables one to ask many global questions. Perhaps the simplest such question is whether there are short nucleotide strings that are absent in a genome as one could not pose this question before when dealing with pieces of a genome. Take, for example, the *Escherichia coli* complete genome. It is a circular DNA of 4639221 nucleotides/letters. Now count the frequency of appearance of all short strings of length $K = 8$. Since there are $4^8 = 65536$ different types of such strings, each type of strings would on average appear $4639221/65536 \approx 71$ times if the genome is a random sequence. Actually the distribution is very much biased towards smaller counts (see Fig. 1 below for its 2D histogram and Figs. 2 and 3 for the 1D histograms) and there are 173 missing strings. On the other hand, there are also many strings with higher counts with respect to the random sequence. In order to clarify the situation for all available genomes we have developed a simple counting method and a visualization scheme [12, 11]. While the mathematics behind the patterns seen in the 2D histograms has been understood thoroughly, the 1D histograms show a rich variety of structures which has been observed only recently, see Fig. 4. This paper is devoted to the explanation of the structures in the 1D histograms of real genomic sequences and their randomized counterparts.

We call an oligonucleotide of length $K$ a $K$-tuple. The statistical study of $K$-tuple composition is a natural extension of *g+c* content or $CpG$ island analysis and the like. In order to visualize the $K$-tuple composition of a long DNA sequence a total of $4^K$ counters are needed. We display these counters as a $2^K$ by $2^K$ square matrix on a computer screen using a crude color code with 16 colors including black and white. The counters are arranged according to the location of the corresponding string in the direct product of $K$ copies of the 2 by 2 matrix $\left(\begin{smallmatrix} g & c \\ a & t \end{smallmatrix}\right)$. At string length $K = 8$ a portrait shows a matrix of $256 \times 256 = 65536$ counters. These portraits are essentially two-dimensional histograms of the $K$-tuple composition of the genomes. We will simply call them 2D $K$-histograms.

Shown in the two upper figures of Fig. 1 are the 2D $K$-histograms at $K = 8$ for two bacterial genomes, *Escherichia coli* and *Haemophilus influenzae* [4]. In the lower figures the 2D histograms of the corresponding randomized sequences are given. The genome sequences were randomized by using the *shuffleseq* program in the EMBOSS package [3].

In a 2D $K$-histogram white color designates an avoided string and bright colors are allocated to small counts. When

[1]Appeared in *CSB2002 Bioinformatics Conference Proceedings*, IEEE Computer Society, Los Alamitos, Ca., 31–42.

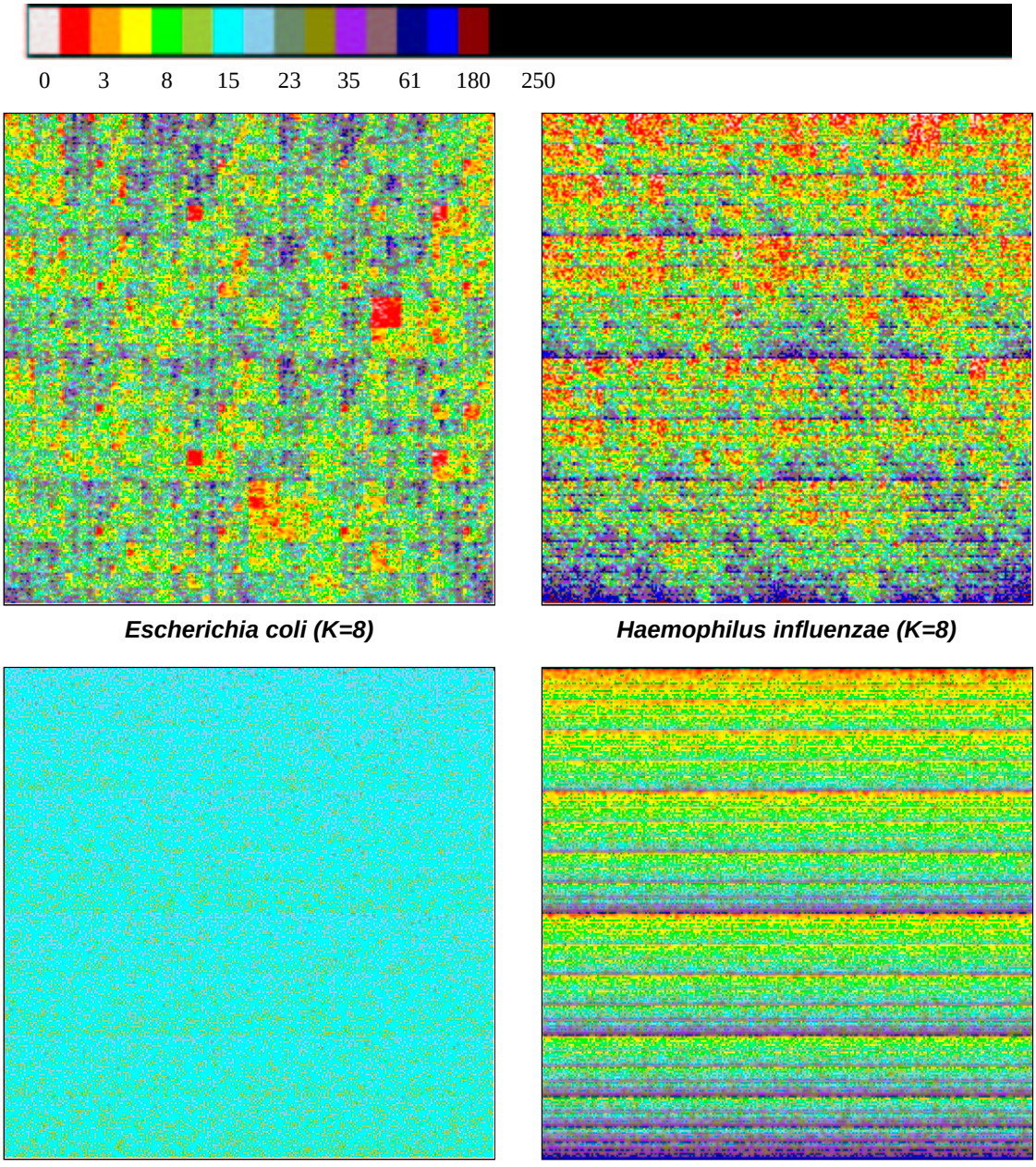



**Figure 1. Two-dimensional histograms of genomic sequences and their randomized counterparts. Each figure shows the 8-tuple counters allocated according to the direct product of 8 copies of the 2 by 2 matrix given in the text. The numbers under the color code give the approximate counts.**

the counts exceed, say, 180, the counters are shown in black. This is a useful "coarse-graining" in order to highlight the under-represented strings. Some mathematical problems associated with those fractal-like patterns in 2D $K$-histograms may be solved exactly by using combinatorial and language theory methods [14, 13, 20].

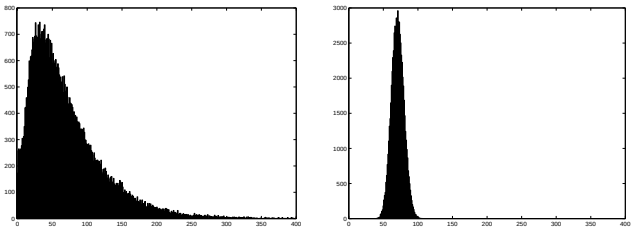
In the 2D $K$-histograms of the randomized sequences all regular patterns disappear. The horizontal contrast in these figures is a consequence of the difference in $g + c$ versus $a + t$ content. In Section 2 we will see that in the 1D $K$-histograms these phenomena become more prominent. In Section 3 three models are proposed to explain the observations. Section 4 is devoted to a study of $K$-histograms by computer simulation, namely by using the Monte Carlo method. A theory for the expectation curve is established by invoking simple mathematical arguments in Section 5. It leads to exact computation of the expectation curve in the simplest model (the eiid model, see Section 3) in Section 6. In Section 7 an approximation by using Poisson distributions is developed, and it clearly elucidates the multimodal phenomenon observed in the $K$-histograms in Section 8. An analysis of short-range correlation in genomes by means of the $K$-histograms for Markov models of different orders makes the content of Section 9. The final Section touches some further problems arising from the exploration of 1D visualization of genomes. We note in passing that most of the figures are calculated for $K = 8$.

## 2. One-Dimensional $K$-Histograms

In order to show the distribution of $K$-tuples as seen in Fig. 1 one regroups the counting results to construct a one-dimensional histogram in the following way. Let the abscissa denotes the counts from a minimal to a maximal number. For a nucleotide sequence of length $L$ the maximal possible count is $L$ or $L - K + 1$ for a circular or a linear genome. This maximal count may be reached only for a sequence made of one and the same type of letter, for example, for an all-$a$ sequence. Usually the actual maximal count is much smaller and will be determined by the drawing program. Along the ordinate we put the number of different $K$-tuple types whose counts fall in the same range on the abscissa.

The first example of a $K$-histogram is given for *Escherichia coli*. The length of the genome is $4639221$ base pairs. The results for the original genome are depicted in Fig. 2 (a). Fig. 2 (b) shows the histogram for a shuffled sequence of the genome.

Two points are worth mentioning. (1) The ranges of counts for these two histograms are 0–774 and 38–112, respectively. For convenience of comparison we take the same range 0–400 in both (a) and (b). (2) Their ranges along the ordinates are quite different. In Figure 3 we superimposed the two histograms one on another. From the area below and above the randomized distribution one calculates that at least $44.4\%$ of the $K$-tuples in *E. coli* are under-represented and $25.4\%$ are over-represented as compared to a random reference sequence.



(a) (b)

**Figure 2. The 8-histograms of *E. coli*: (a) for the original genome, (b) for a shuffled sequence of the genome.**

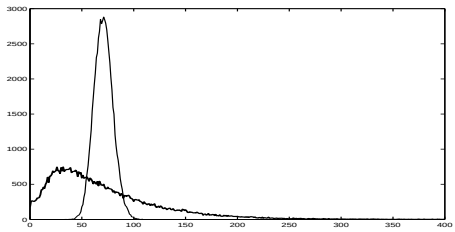


**Figure 3. The comparison of Fig. 2 (a) and Fig. 2 (b).**

In Table 1 we list some statistical characteristics extracted from the calculation that generates Figures 2 (a) and 2 (b). The entries in Table 1 are standard ones in characterizing a statistical distribution (see, e.g., [6]): min and max are the minimal and the maximal count of $K$-tuples. The case min$= 0$ means that there are missing, avoided $K$-tuples. The first quartile, median and third quartile reflect the proportion of the $K$-tuples within the total number $4^K$. The range is the difference between max and min. IQR, the interquartile range, is the difference between the third quartile and the first quartile. It is clear that $50\%$ of $K$-tuples occur in an interval of length IQR. The mode is the most frequently encountered count, and the frequency of mode is the number of distinct $K$-tuples at that count.

The second example is a similar analysis for *Haemophilus influenzae*. The length of the genome is $1830023$. Two $8$-histograms are depicted in Figure $5$: (a) is for the original genome, and (b) is for a shuffled

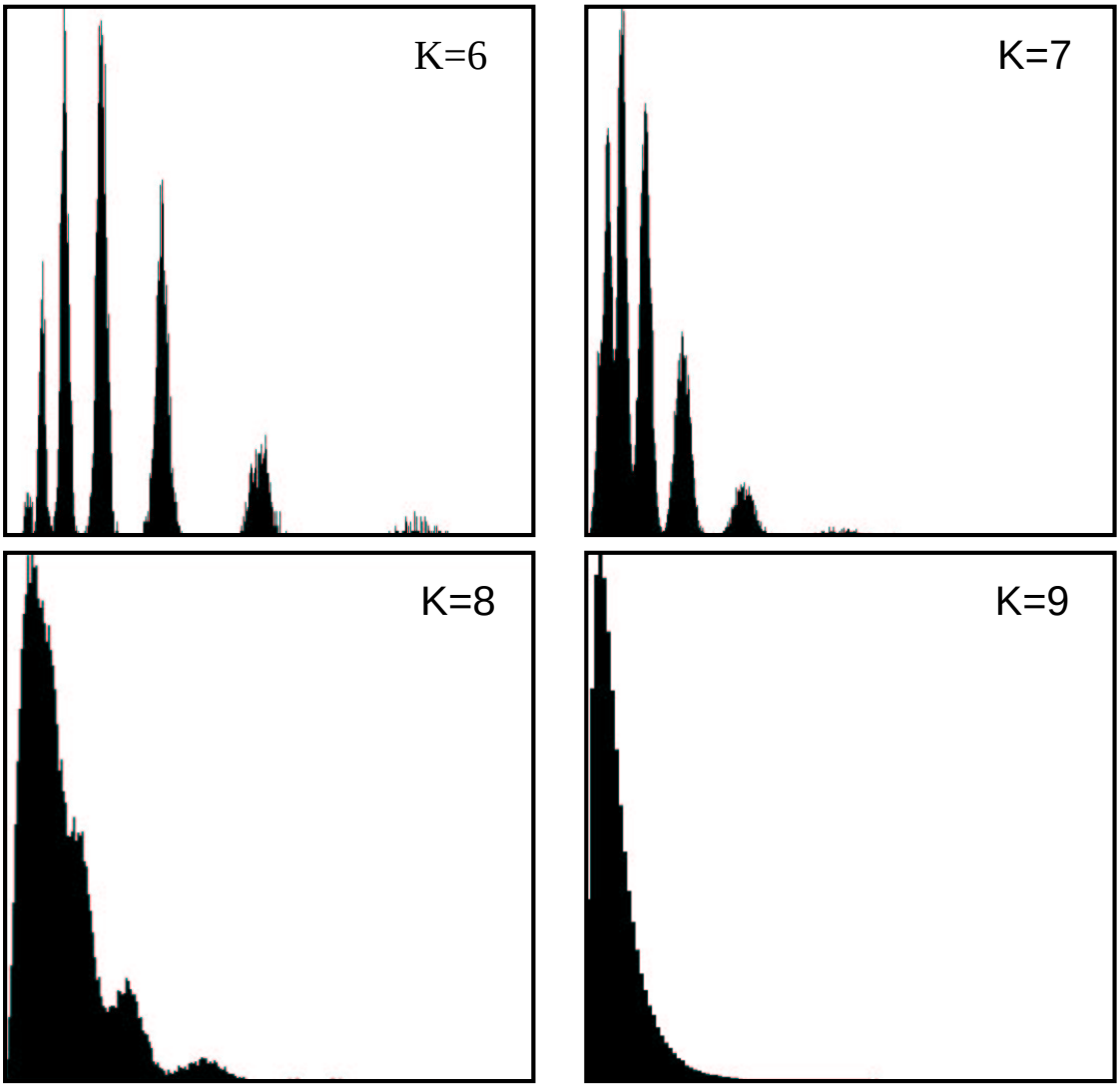


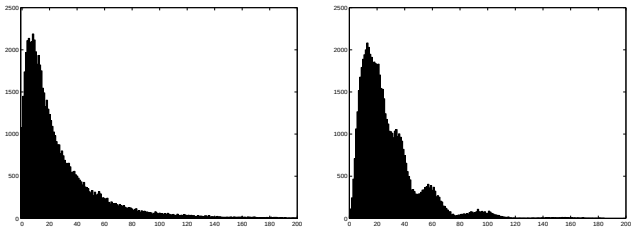

**Figure 4. One-dimensional histograms of the randomized genomic sequence of *H. influenzae* at different string length $K$. Along the abscissa are the counts of $K$-tuples and the ordinate gives the number of string types in each counting range. We did not put numbers on the axes because more important is the multi-modal structure of the histograms.**

| *Escherichia coli* | min | 1st quartile | median | 3rd quartile |
|---|---|---|---|---|
| complete genome | 0 | 32 | 57 | 95 |
| shuffled genome | 38 | 65 | 71 | 77 |

| max | range | IQR | mode | frequency of mode |
|---|---|---|---|---|
| 774 | 774 | 63 | 32 | 746 |
| 112 | 74 | 12 | 70 | 2894 |

**Table 1. Statistical characteristics of Fig. 2 (a) and 2 (b). For meaning of the entries see text.**

sequence of the genome. Here the natural ranges of the two 8-histograms are 0–844 and 0–184, but in Fig. 5 we take the same range 0–200 for comparison. We note that Fig. 5 (b) is the same as the $K = 8$ figure in Fig. 4, which juxtaposes the histograms for the randomized sequences at different $K$ from 6 to 9.



(a) (b)

**Figure 5. The 8-histograms of *H. influenzae*: (a) For the original genome; (b) For its shuffled sequence.**

It is obvious from Fig. 5 (b) and, especially, from Fig. 4 that the histograms of the randomized sequences of *H. influenzae* has much richer structure than its original genome and the randomized genome of *E. coli*. Moreover, we have found that although randomization of one and the same genome produces a different sequence at each realization, the main features of the $K$-histograms, however, remain unchanged. Put in other words, the observed structure is a robust property of the ensemble of randomized sequences obtained from the original complete genome by shuffling.

Figure 6, similar to Figure 3, compares Fig. 5 (a) and Fig. 5 (b).

The statistical characteristics of Fig. 5 are summarized in Table 2 in the same way as in Table 1.

In order to understand the structure of these 1D $K$-histograms we study a few models that generate random sequences as reference in comparison with genomic sequences.

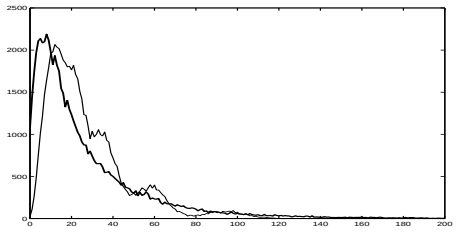


**Figure 6. The comparison of Fig. 5 (a) and Fig. 5 (b) by superimposing the two curves.**

| *Haemophilus influenzae* | min | 1st quartile | median | 3rd quartile |
|---|---|---|---|---|
| complete genome | 0 | 8 | 18 | 36 |
| shuffled genome | 0 | 13 | 22 | 36 |

| max | range | IQR | mode | frequency of mode |
|---|---|---|---|---|
| 844 | 844 | 28 | 8 | 2188 |
| 184 | 184 | 23 | 13 | 2081 |

**Table 2. Statistical characteristics of Fig. 5 (a) and 5 (b).**

## 3. Three Sequence Models

In this paper we consider three classes of stochastic models: (1) equal-probability independently and identically distributed model (abbreviated as eiid), (2) non-equal-probability independently and identically distributed model (abbreviated as niid), and (3) Markov models of order $n \geqslant 1$ (abbreviated as MMn).

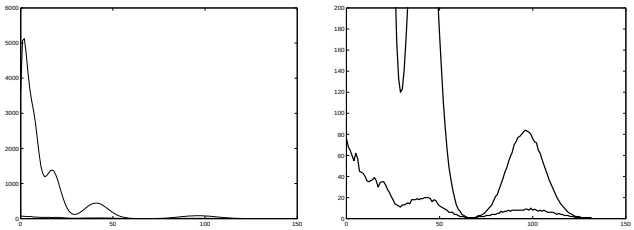
It is natural that the shuffled sequences used in obtaining Fig. 2 (b) and Fig. 5 (b) can be considered as generated from some models of niid, in which the mono-nucleotide compositions are taken from the corresponding genome sequences and kept constant. In the same way we can count the frequencies of $(K + 1)$-tuples in a complete genome to construct a Markov model of order $K$.

The problem whether the stochastic properties of the ensemble of niid models and the shuffled ensemble are the same may be partly elucidated by Monte Carlo simulations. Since we do not know how to shuffle a sequence while keeping its Markov transition probabilities unchanged, a positive answer for niid speaks in favor of treating more complex models such as MMn by inferring the transition matrix from counting $K$-tuples in a real genome.

## 4. Monte Carlo Method

The Monte Carlo method is quite instructive for understanding the robust structure of $K$-histograms of random sequences generated by some stochastic models.

An experiment is carried out for an niid model, in which the compositions of $a, c, g, t$ are $15:35:35:15$, the length of sequences is taken to be $10^6$, and $K = 8$. By running a Visual C++ program 100 times, we obtained the estimated values of the expectation (*ex*) and the standard deviation (*sd*) in the range 0–150 shown in Figure 7.



(a) (b)

**Figure 7. The estimated expectation (the upper curve) and the standard deviation (the lower curve) for an niid model in the range 0–150: (b) is a zoom-in of (a).**

It is clear that either the estimated values of *sd* are much smaller than those of *ex* when *ex* is large or both of them are very small. This provides an explanation for the robustness of the $K$-histograms.

In Fig. 8 we compare the absolute error of *ex* in an experiment with the *sd* obtained for the same range as a reference. We see that the error never goes beyond the limit of $2\times$ *sd*. Let $X_n$ be a random variable for each $n$ in the count range of the $K$-histogram, then under some conditions the distribution of $X_n$ may be asymptotically normal (see related discussion in [21, Ch.12]).

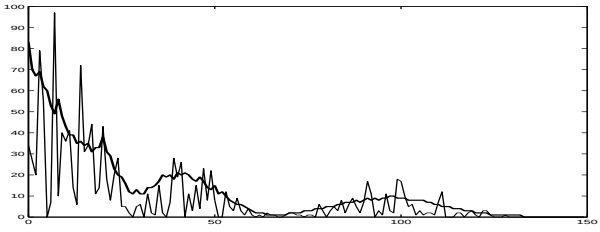
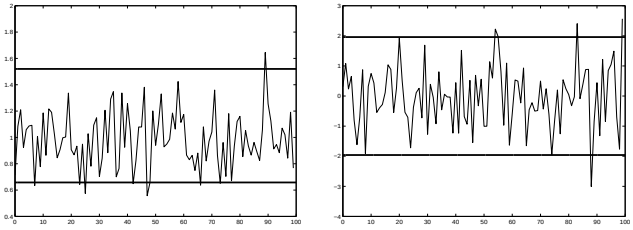


**Figure 8. The comparison of absolute error of *ex* in an experiment with *sd* as a reference.**

Now we are in a position to use Monte Carlo method to answer the question proposed before, namely, whether the ensemble obtained by shuffling a given sequence and the ensemble produced by a stochastic model inferred from the genomic sequence are the same. For the niid model with the same parameters mentioned above, we have done firstly the F test and then the t test for the two ensembles. The results at a confidence level of $0.95$ are shown in Fig. 9



(a) (b)

**Figure 9. Results of (a) F test and (b) t test from Monte Carlo experiments.**

It is clear that the results of F test and t test give positive answers, i.e. to accept the Null hypothesis that the variances and expectations of both ensembles are equal [6].

Hence it is plausible that robustness of the $K$-histograms may be universal for models discussed in this paper. Furthermore, the ensembles obtained from a sequence-generating model and from shuffling a give sequence are stochastically the same (at least for *ex* and *sd*).

Consequently, in the remaining part of this paper we will focus our attention on the study of expectation curve which will be defined in the next Section.

## 5. Two Theorems on the Expectation Curve

In order to reveal the laws which govern the $K$-histograms, we denote all $K$-tuples of 4 letters (i.e. $a, c, g, t$) by $x_i$, $i = 1, 2, \ldots, 4^K$ (sticking to some fixed ordering). The notation $s = s_L$ is used to denote any sequence of length $L$. Since $s$ is randomly generated from a stochastic model, we simply call $s$ a realization of the model.

Now we introduce the following definitions and notations rigorously.

**Definition 1** For each $n \geqslant 0$, the random variable $X_n$ is the number of distinct $K$-tuples which occur in the realization $s$ with the same occurence number of $n$.

**Remark**. The maximal value of $n$ in general varies for different $L$ and $K$. Of course, it is always true that $n \leqslant L - K + 1$. In many cases we will not indicate the maximal range of $n$ explicitly.

**Definition 2** For $n \geqslant 0$ and $i \in \{1, 2, \ldots, 4^K\}$, the notation $p_{i,n}$ is used to represent the probability of occurrence

of the $K$-tuple $x_i$ in a realization $s$ with the exact number of appearance $n$.

If we consider $n$ as discrete time, then each realization $s$ can be thought as a realization of a stochastic process. Hence a natural way of study is to calculate the expectation and variance of random variable $X_n$ for each $n$. In what follows by an *expectation curve* of $K$-histogram we refer to the curve which is obtained by connecting the point set of $\{(n, X_n) \mid n \geqslant 0\}$. The following two Theorems establish the formula for computing the expectation curve.

**Theorem 1** For each $n \geqslant 0$, the mathematical expectation of random variable $X_n$ is given by

$$E(X_n) = \sum_{i=1}^{4^K} p_{i,n} = p_{1,n} + p_{2,n} + \cdots + p_{4^K,n}. \qquad (1)$$

*Proof.* For fixed $n \geqslant 0$, using $A_{i,n}$ to denote the event of occurrence of $K$-tuple $x_i$ in the realization $s$ in exact number of $n$. Define the indicator of $A_{i,n}$ by

$$I_{i,n} = \begin{cases} 1, \text{the event } A_{i,n} \text{ occurs,} \\ 0, \text{otherwise.} \end{cases} \qquad (2)$$

It is obvious that $I_{i,n}$ is a random variable having a two-point distribution. From the definition 2 we know that the expectation of $I_{i,n}$ is

$$E(I_{i,n}) = p_{i,n}.$$

Again using the definition 1 we have

$$X_n = \sum_{i=1}^{4^K} I_{i,n} = I_{1,n} + I_{2,n} + \cdots + I_{4^K,n},$$

and the formula (1) follows by taking expectation on both sides. ∎

We make a further definition as follows.

**Definition 3** For $i, j \in \{1, 2, \ldots, 4^K\}$, $i \neq j$, using the notation $p_{i,j,n}$ to denote the probability of occurrence of both the $K$-tuples $x_i$, $x_j$ in a realization $s$ in the same exact number of $n$.

Using these definitions we can obtain the following result.

**Theorem 2** For $n \geqslant 0$, the variance of random variable $X_n$ is

$$D(X_n) = E(X_n^2) - E(X_n)^2,$$

where the first term can be computed by the formula

$$E(X_n^2) = \sum_{i=1}^{4^K} p_{i,n} + 2 \times \sum_{1 \leqslant i < j \leqslant 4^K} p_{i,j,n}. \qquad (3)$$

The proof is rather the same as that of Theorem 1 and hence omitted.

Note that since the proof of Theorem 1 does not refer to the underlying model, the results of Theorem 1 (and Theorem 2) are valid for all three classes of models, i.e. eiid, niid, and MMn. Of course, the computations of probabilities $p_{i,n}$ and $p_{i,j,n}$ are closely related to the models used in each case.

On the other hand, if we take the number of occurrence of a $K$-tuple in a realization $s$ as a random variable, then Theorem 1 says that the expectation curve is the summation of probability functions of random variables corresponding to all $K$-tuples. To be precise, $p_{i,n}$ is the probability of occurrence of $x_i$, the $i$-th $K$-tuple, in $s$ in exact number of $n$ copies. Fixing $i$, and letting $n$ take all non-negative integers, we obtain the probability function of the number of occurrence of $x_i$ as a random variable.

If we divide the value of $E(X_n)$ by $4^K$, the number of all possible $K$-tuples, then we obtain a probability function in the sense of average. It is evident that for each $n$ the value of $E(X_n)/4^K$ is the expected percentage in all $K$-tuples which will occur repeatedly in exact number of $n$ times.

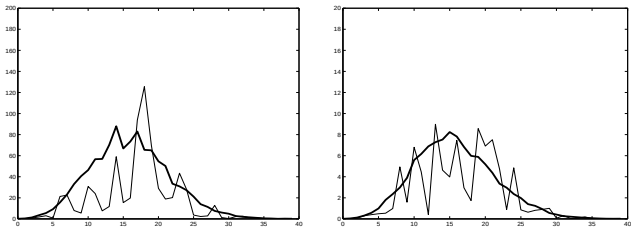
Therefore, we can look at the expectation curve of a $K$-histogram from two points of view. Firstly, this curve can be considered as generated from a stochastic process of $\{X_n \mid n \geqslant 0\}$, and, secondly, it is the summation of all probability functions which are generated from each $K$-tuple individually. Theorem 1 establishes the relationship between these two viewpoints. It may seem that the two terms "$K$-histogram" and "expectation curve" are not in good agreement. However, they do reflect the two lines of thinking on the same object.

## 6. Exact Computation of Expectation Curve

From Theorem 1 it is clear that in order to obtain the expectation curve we need to know the probability distribution of the occurrence number (as a random variable) for every $K$-tuple. As we know from the literature that there have been formulae of expectation and variance for different models (e.g. see [19, 21] and references therein), but it seems that there were very few papers in which the probability distributions of the occurrence number for each and every $K$-tuple were obtained explicitly. The only reference to our knowledge is [5] in which the model was restricted to eiid. In [17] the same problem was considered for eiid and $K$-tuples without overlaps. It is easy to understand that in the eiid model the problem of finding probability functions reduces completely to some problems in enumerative combinatorics. Therefore, the computation in [5] was based on the Goulden-Jackson cluster method [7, 8, 16], which were also used in our works [11, 14, 13].

Although at present time the expectation curve can only

be computed in the case of the simplest model, i.e. eiid, we can still use the result to do two types of experiments: (1) to see the error of an experiment with respect to the exact expectation curve, (2) to see the error occurred in Monte Carlo method of Section 4. Since we have no effective algorithm to execute the computation of exact variance curve through the formula of Theorem 2, we use instead the estimation of *sd* obtained by Monte Carlo method in both experiments. The results are depicted in Figure 10.



(a) (b)

**Figure 10. (a) The error of an experiment versus exact expectation curve. (b) The error of Monte Carlo method versus exact expectation curve. In both pictures the thick curves represent the estimated *sd* by Monte Carlo simulation.)**

An open problem is how to compute the exact expectation curve and variance curve for the more complicated models like niid and MMn. In principle, the Goulden-Jackson cluster method can be used to do the job for small values of $L$ and $K$. However, it is not clear whether this method is capable to provide an effective algorithm for sequence length in the order of $10^6$.

## 7. Poisson Approximation method

It was proved in [17] that the probability distribution of the occurrence number of a $K$-tuple in a realization $s = s_L$ obeys asymptotically a Poisson distribution. The conditions of this assertion are: the model is eiid, the $K$-tuple has no self-overlap, the length $L$ of $s$ is big enough, and the probability of occurrence of a $K$-tuple is small enough.

Many authors have pointed out that the phenomenon of overlaps are indispensable for discussing the problem (e.g. see [19, 5, 21, 18]). One should note that there are different types of overlapping $K$-tuples (see, e.g., [9, 10]). For example, taking $K = 8$, the 8-tuples $aaaaaaaa$, $acagtaca$, $acgtacgt$ belong to different overlapping types. Their probability distributions in a realization are different. For all $65536$ types of 8-tuples there are altogether 13 distinct overlapping types, including the non-overlap type. The numbers of 8-tuples in each type are different. For the model of eiid they have the same expectation, but different variances.

From our experiments we put forward a conjecture that when we use Theorem 1 to compute the expectation curve, the Poisson approximation can give satisfying results for all models in this paper, including niid and MMn, and for all $K$-tuples. At least there is no significant error arising from this approximation method.

A typical experiment is as follows. Here the niid model is established from the complete genome of *H. influenzae* as described in Section 2 (see Figures 5 and 6). The frequencies of mono-nucleotides in the genome are given in Table 3

| $a$ | $c$ | $g$ | $t$ |
|---|---|---|---|
| 0.310173 | 0.191650 | 0.189853 | 0.308325 |

**Table 3. The frequencies of mono-nucleotides in the complete genome of *H. influenzae*.**

In order to compute the Poisson approximation for each $K$-tuple (here taking $K = 8$), we need to know the expectation of its occurence number (as a random variable). Observe that in Table 3 the frequencies of $a$ and $t$ are nearly equal, and the same holds for $c$ and $g$. Hence we can use a simplified niid model as the first approximation, in which the frequencies for $c$ and $g$ take the same value $p = 0.19075$ (the mean of the values in Table 3), and the frequencies of $a$ and $t$ are $0.30925 (= 0.5 - p)$. The probability of occurrence of a 8-tuple in a realization $s = s_L$ is determined by the number of appearance of $a, c, g, t$ in this tuple. Denoting these numbers by $n_a, n_c, n_g, n_t$. If $n_c + n_g = k$, $0 \leqslant k \leqslant 8$, then the probability of occurrence of this 8-tuple in any place of the sequence $s$ is

$$p_k = p^k (0.5 - p)^{8-k},\ 0 \leqslant k \leqslant 8.$$

Since the length of $s = s_L$ is known ($L = 1830023$), then the expectation of this 8-tuple in $s$ is $L \times p_k$ which is nothing but the only parameter $\lambda$ in the corresponding Poisson distribution. (The exact formula should be $(L-7) \times p_k$, but the error caused by this modification is negligible.)

After obtaining the expectation $\lambda_i$ for the $i$-th 8-tuple, we can use the formulae

$$p_{i,n} = \frac{\lambda_i^n}{n!} e^{-\lambda_i},\ 1 \leqslant i \leqslant 4^K,\ n \geqslant 0,$$

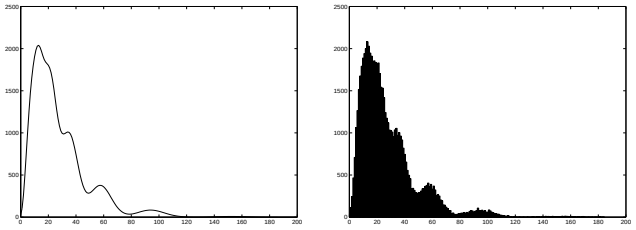
and (1) to perform the calculation requierd in Theorem 1.

A simpler method is to enumerate the number of 8-tuples which have the same expectation. It is clear that for $K = 8$ there are only 9 distinct values of expectations among all $65536$ types of 8-tuples for the simplified niid model mentioned above. If $n_c + n_g = k$, $0 \leqslant k \leqslant 8$, then the number

of 8-tuples having the value $p_k$ as their expectation can be deduced as follows.

$$\begin{aligned} n_k &= \sum_{n_c+n_g=k} \sum_{n_a+n_t=8-k} \frac{8!}{n_c!n_g!n_a!n_t!} \\ &= \sum_{n_c+n_g=k} \sum_{n_a+n_t=8-k} \frac{k!}{n_c!n_g!} \cdot \frac{(8-k)!}{n_a!n_t!} \cdot \frac{8!}{k!(8-k)!} \\ &= \sum_{n_c+n_g=k} \frac{k!}{n_c!n_g!} \sum_{n_a+n_t=8-k} \frac{(8-k)!}{n_a!n_t!} \cdot \frac{8!}{k!(8-k)!} \\ &= 2^k \cdot 2^{8-k} \cdot C_8^k \\ &= 2^8 \cdot C_8^k. \end{aligned} \tag{4}$$

The result is shown in Fig. 11(a). For the convenience of comparison the 8-histogram in Fig. 5(b) is shown again in Fig. 11(b).



(a) (b)

**Figure 11. (a) The Poisson approximation of the expectation curve of an niid model which comes from *H. influenzae*. (b) An 8-histogram of this niid model.**

Although we cannot get the exact expectation curve in the niid model to examine the validity of the Poisson approximation, we can at least use the estimation of *ex* and *sd* obtained by Monte Carlo simulation in Section 4 to do this comparison. The result is given in Fig. 12, in which the thick curve (of the *sd*) is used as a reference for the difference between these two methods. It turns out that the curve obtained from Poisson approximation method is good enough to be used in niid model. We have also done experiments for the model of MMn, the conclusion is similar and thus omitted.

**Remark.** Using a Visual C++ program we have done the experiment with the four exact frequencies listed in Table 3 and carried out the summation of Eq. (1) for *H. influenzae*. The results are very close to those of the simplified version mentioned above and hence also omitted.

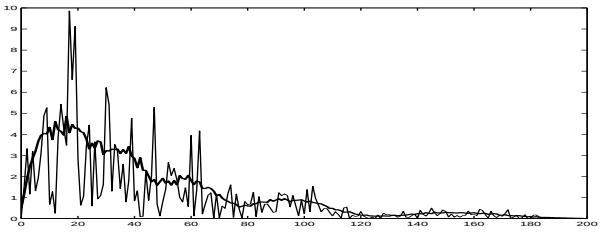


**Figure 12. The comparison of Poisson approximation and Monte Carlo method. The thick curve of estimated *sd* is used as a reference.**

## 8. Analysis of the Mechanism of Multi-Modal $K$-Histograms

The Poisson approximation is useful not only to compute the expectation curve but also to analyze the multi-modal phenomenon as seen in Fig. 11. We explain our discovery keeping using the example of *H. influenzae* in Fig. 11.

As shown in Section 7 that there are only 9 distinct expectation value among all $4^8 = 65536$ types of 8-tuples appearing in the complete genome of *H. influenzae*:

$$\lambda_k = 1830023 \times 0.19075^k \times 0.30925^{8-k},\ k = 0, 1, \ldots, 8.$$

We list the values of these expectations and the corresponding binomial coefficients (which are needed in Eq. (4)) in Table 4.

| $k$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $\lambda_k$ | 153.086 | 94.4254 | 58.243 | 35.9252 |
| $C_8^k$ | 1 | 8 | 28 | 56 |

| 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|
| 22.1592 | 13.6681 | 8.43069 | 5.20018 | 3.20755 |
| 70 | 56 | 28 | 8 | 1 |

**Table 4. The values of $\lambda_k$ and binomial coefficients for the model of niid obtained from *H. influenzae*.**

It is clear that, starting from the largest value of $\lambda_0$, the values of $\lambda_k$ decrease continually in the ratio of

$$0.19075/0.30925 = 0.616815. \tag{5}$$

Hence the distance between two successive $\lambda_k$ becomes smaller and smaller as $k$ increases. Its effect can be seen directly in Fig. 13.

Therefore, it can be said qualitatively that the multi-mode in $K$-histograms is formed from the summation in

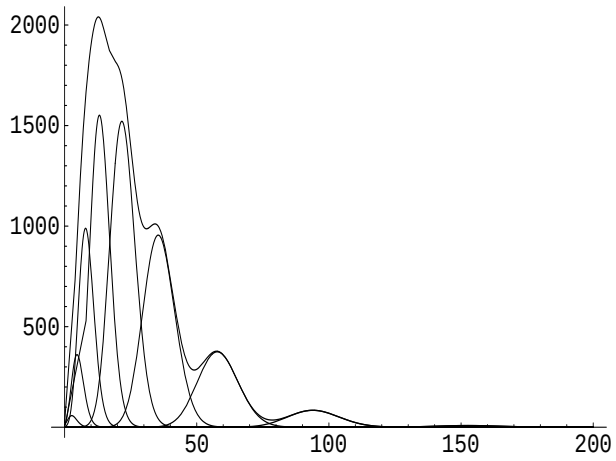



**Figure 13. The individual Poisson probability functions and their summation for an niid model obtained from *H. influenzae*.**

formula (1), where each $p_{i,n}$ forms a unimodal curve (when $n$ runs within a certain range). From Fig. 13 it is clear, however, that if two unimodal curves are very close to each other then their summation may still be an unimodal curve. Here two factors are in effect: both the distance between successive modes and the heights of the modes are important to the final summation.

Returning to the complete genome of *E. coli* whose corresponding 8-histograms were shown in Figures 2 and 3, we can do the same analysis. However, the result is not as drastic as that for *H. influenzae*. From Table 5 we see that the four types of mono-nucleotides are distributed almost evenly in the complete genome of *E. coli*.

| $a$ | $c$ | $g$ | $t$ |
|---|---|---|---|
| 0.246191 | 0.254231 | 0.253658 | 0.245920 |

**Table 5. The frequencies of mono-nucleotides in the complete genome of *E. coli*.**

Based on the experience with the genome of *H. influenzae*, we apply the simplified niid model to *E. coli*. The frequencies of $c$ and $g$ are taken the same as $0.253945$, while those of $a$ and $t$ are $0.246055$. In Table 6 we list the parameters needed for the Poisson approximation, similar to those listed in Table 4 for *H. influenzae*.

The ratio of successive values of expectation for *E. coli* is quite different from that of (5):

$$0.246055/0.253945 = 0.968931. \qquad (6)$$

The values of expectation of every Poisson probability function are rather evenly distributed in the small range from

| $n$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $\lambda_k$ | 62.3308 | 64.3295 | 66.3923 | 68.5212 |
| $C_8^k$ | 1 | 8 | 28 | 56 |

| 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|
| 70.7184 | 72.9861 | 75.3264 | 77.7419 | 80.2347 |
| 70 | 56 | 28 | 8 | 1 |

**Table 6. The values of $\lambda_k$ and binomial coefficients for the model of niid obtained from *E. coli*.**

62.3308 to 80.2347. The individual curves for each Poisson probability function stay close to each other and their summation leads to a unimodal curve as shown in Fig. 14. Therefore, the randomized *E. coli* sequence is well described by an eiid model.

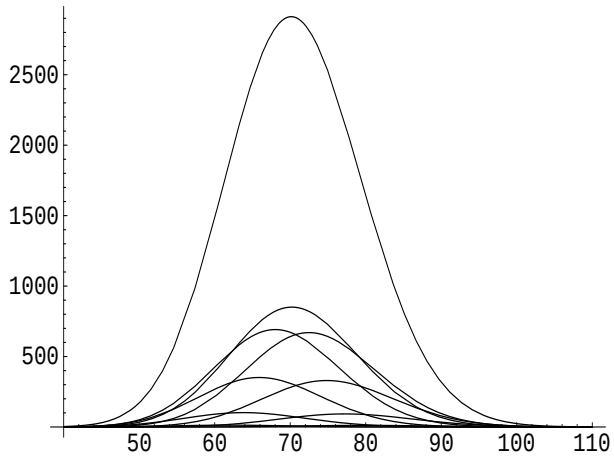


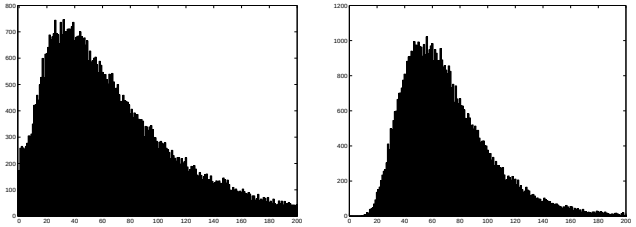

**Figure 14. The individual Poisson probability functions and their summation for an niid model obtained from *E. coli*.**

**Summary**. We have found the underlying mechanism behind the multi-modal phenomena in $K$-histogram. The analysis of the two examples for *E. coli* and *H. influenzae* has shown that the basic reason is due to the content of $g+c$. The exact number of modes in a concrete expectation depends on several factors. Since it is a robust property of a majority of $K$-histograms we can determine the number of modes simply by depicting one $K$-histogram for a randomized sequence.

## 9. Analysis of Short-Range Correlations in Genome Sequences by $K$-Histograms

In this Section we try to analyze the short-range correlations which exist in a complete genome sequence taking *E. coli* as an example.

Our idea is very simple. In Fig. 2 and 3 we see that the 8-histograms of the original genome sequence and its shuffled counterpart are significantly different. Since the shuffled sequence can be seen as generated from a niid model (this has been tested in Section 4, see Fig. 9), there is no correlation between letters in the sequence. What would happen if we consider correlations in the original genome sequence? For example, we can count the number of occurrence of each di-nucleotide and then establish a Markov model of order 1 (as in, e.g., [1]). For clarity of comparison we depict the two 8-histograms in Fig. 15, where Fig. 15 (a) is the same histogram as in Fig. 2 (a), but the range is zoomed to 0–200, and Fig. 15 (b) is the 8-histogram of a random sequence generated from the MM1 obtained above. Note that the ranges of ordinate in two 8-histograms are not the same.

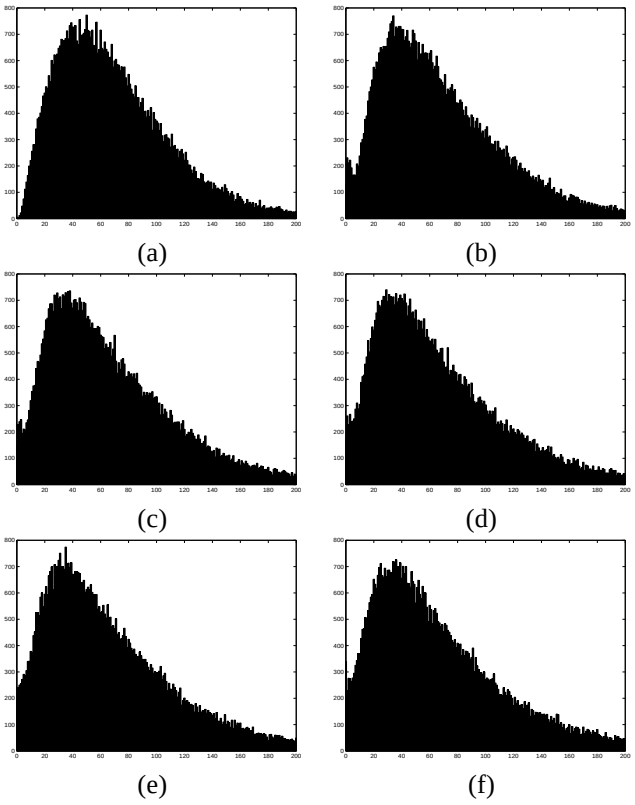


(a) (b)

**Figure 15. (a) The 8-histogram of *E. coli*. (b) An 8-histogram of sequence generated from am MM1 model which has the same di-nucleotides composition as the complete sequence of *E. coli*. Both pictures are restricted to the range 0–200.**

We can pursue these considerations further. However, as long as 8-tuples are used, the order of the corresponding Markov models cannot exceed 7, otherwise the random fluctuations in the sequences will show off. In Figure 16 we list the 8-histograms of sequences generated from Markov models of order 2–7, respectively.

In examining Figures 15 and 16 we can focus on two distinct features in the 8-histograms: (1) The position and height of the mode, (2) The number of avoided or very under-represented 8-tuples. Then we see that starting from the order of about $n = 5$, the 8-histograms of MMn are nearly the same as Fig. 15(a). Therefore, it is enough to use the Markov model of order 5 as the first approximation for the complete genome sequence of *E. coli*. This conclusion has been obtained and used in many previous works for gene-finding (e.g. [2]).



(a) (b) (c) (d) (e) (f)

**Figure 16. The 8-histograms of sequences generated from models of MMn which has the same n+1 nucleotides composition with as the complete sequence of *E. coli* for (a) n=2, (b) n=3, (c) n=4, (d) n=5, (e) n=6, (f) n=7.**

## 10. Discussion

Most of the results in this paper bear an experimental nature except for those in Section 5. Therefore, many problems are left for future study. We list a few.

1. The selection of the value of $K$.

   One fact is plainly true that for $K = 1$ we can only get the frequencies of the 4 mono-nucleotides and nothing more. Hence it is plausible that only one value of $K$ could not grasp too many information contained in the genome sequences. The situation here is different from the 2D visualization scheme that if $K$ is too big, then the corresponding $K$-histogram contain very little useful information since most $K$-tuples would be missing and the $K$-histogram is less interesting. Hence how to

select appropriate values of $K$ is worth discussing in using the 1D visualization presented in this paper.

2. How to use 1D visualization to study protein sequences.

   Here the sequences have 20 distinct letters but are of much shorter length. Hence the difficulty confronted is of quite a different nature. We have been using the counting of $K$-tuples in amino acid sequences already (e.g. see [15]), but a thorough analysis is still lacking.

3. How to study mathematically the exact properties of random variables $X_n$, i.e. the number of distinct $K$-tuples which have the same occurrence number $n$ in a realization $s$.

   Theorems 1 and 2 give only the expectation and variance of $X_n$. Both of them reduce the study of $X_n$ into the study of occurence number of each $K$-tuple. Since the latter problem has its solution only in the eiid model which is the simplest stochastic model, maybe this reduction is not the best approach for studying $X_n$. An easy estimation of $X_n$'s *sd* is useful to provide a true proof of robustness discussed in Section 4. The asymptotic property and correlation of $X_n$ are not known yet.

4. The conjecture that Poisson approximations can be used safely to all three models also needs a proof. We don't know the real reason behind it, and a rigorous proof is welcome.

   Here it is interesting to note that in the early literature of computational biology it was common to use the value of expectation as variance in counting $K$-tuples in DNA sequences (see, e.g., [1]), and later on there appeared many critics on this point (see, e.g., [18]). But our experiments involved in 1D visualization have shown that using the value of expectation as variance will not cause serious error in the sense of average. After all, the 1D visualization scheme is of the nature of averaging over all $K$-tuples and it is well described by Poisson approximations.

**Acknowledgement**

This work is supported by the Special Funds for Major State Basic Research Projects. BLH also received support from the Innovation Project of CAS and the Beijing Municipality "248" Innovation Project.

## References

[1] M. Y. Borodovskii, Y. A. Sprizhistsii, E. I. Golovanov, and A. A. Aleksandrov. Statistical patterns in primary structures of the functional regions of the genome in *escherichia coli*. *Mol. Biol.*, 20:1014–1023, 1024–1033, 1144–1150, 1986.

[2] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94, 1997.

[3] EMBOSS. this package may be fetched by anonymous ftp from: `ftp://ftp.uk.embnet.org/pub/emboss`.

[4] GenBank. Fetched by anonymous ftp from the URL: `ftp://www.ncbi.nlm.nih.gov/genbank`.

[5] J. F. Gentleman and R. C. Mullin. The distribution of the frequency of occurrence of nucleotide subsequences, based on their overlap capability. *Biometrics*, 45:35–52, 1989.

[6] T. Glover and K. Mitchell. *An Introduction to Biostatistics*. McGraw-Hill Companies, Inc., New York, 2001.

[7] I. P. Goulden and D. M. Jackson. An inversion theorem for cluster decompositions of sequences with distinguished subsequences. *J. London Math. Soc.*, 20:567–576, 1979.

[8] I. P. Goulden and D. M. Jackson. *Combinatorial Enumeration*. John Wiley & Sons Inc., New York, 1983.

[9] L. J. Guibas and A. M. Odlyzko. Periods in strings. *J. of Combinatorial Theory, Series A*, 30:19–42, 1981.

[10] L. J. Guibas and A. M. Odlyzko. String overlaps, pattern matching, and nontransitive games. *J. of Combinatorial Theory, Series A*, 30:183–208, 1981.

[11] B.-L. Hao. Fractals from genomes — exact solutions of a biology-inspired problem. 282:225–246, 2000.

[12] B.-L. Hao, H.-C. Lee, and S.-Y. Zhang. Fractals related to long DNA sequences and complete genomes. *Chaos, Solitons and Fractals*, 11:825–836, 2000.

[13] B.-L. Hao, H.-M. Xie, Z.-G. Yu, and G.-Y. Chen. Avoided strings in bacterial complete genomes and a related combinatorial problem. *Annals of Combinatorics*, 4:247–255, 2000.

[14] B.-L. Hao, H.-M. Xie, Z.-G. Yu, and G.-Y. Chen. Fractorisable language: form dynamics to genomes. *Physica A*, 288:10–20, 2000.

[15] B.-L. Hao, H.-M. Xie, and S.-Y. Zhang. Compositonal representation of protein sequences and the number of Eulerian loops, ITP UCSB Preprint NSF-ITP-01-18, Los Alamos National Lab E-archive, physics/0103028.

[16] J. Noonan and D. Zeilberger. The Goulden-Jackson cluster method: extension, applications and implementations, 1998, downloadable from `http://www.math.temple.edu/~zeilberg`.

[17] O. E. Percus and P. A. Whitlock. Theory and application of Marsaglia's monkey test for pseudorandom number generators. *ACM Transactions on Modeling and Computer Simulation*, 5:87–100, 1995.

[18] P. A. Pevzner. *Computational Molecular Biology: An algorithmic Approach*. The MIT Press, Cambridge, Massachusetts, 2000.

[19] P. A. Pevzner, M. Y. Borodovsky, and A. A. Mironov. Linguistics of nucleotide sequences I: The significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *J. of Biomolecular Structure and Dynamics*, 6:1013–1026, 1989.

[20] P. Tiňo. Multifractal properties of Hao's geometrics representations of dna sequences, preprint. 2001.

[21] M. A. Waterman. *Introduction to Computational Biology*. Chapman & Hall, London, 1995.