An Application of the Chen-Stein Method to DNA Sequence Analysis

by
Xianlong Wang

an expository paper submitted to
the Department of Mathematics
at Oregon State University

in partial fulfillment of
the requirement for the degree of
Master of Science

December 2, 2008

Major Professor: Mina Ossiander
Department of Mathematics
Oregon State University

Minor Professor: Xiaoli Fern
School of Electrical Engineering and Computer Science
Oregon State University

ACKNOWLEDGEMENT

I would like to thank my advisor, Dr. Mina Ossiander for her advice and help during my graduate study in Mathematics. She is always available to answer any question in probability and analysis, patient in providing guidance and advice.

I also want to thank Dr. Xiaoli Fern for helping me study in Computer Science, and thank Dr. Yevgeniy Kovchegov for taking valuable time to serve on my committee.

Abstract

This paper explores the application of the Chen-Stein method to DNA sequence analysis. In particular, we are interested in the probability of detecting an unusual congruence between 2 sequences. We compare the results from theoretical prediction according to the Chein-Stein method, real data, and computer simulations. Statistical testing and confidence intervals are also provided to assess the accuracy of the theoretical results. Our results show that the Chen Stein method gives a decent estimate of the exact probability.

Contents

1 The problem **6**

1.1 Sequence matching 6

1.2 The problem 7

2 The Chen-Stein Method **8**

2.1 Introduction to the Chen-Stein method 8

2.2 Birthday Problem 12

3 Application **15**

3.1 Theoretical prediction 15

3.2 Real data result 18

3.3 Computer simulation 18

4 Summary and Discussion **19**

4.1 Simulation vs. prediction 19

4.2 Real data result vs. prediction 21

5 References **23**

6 Appendix **25**

6.1 Multinomial test 25

6.2 The ballot theorem 26

1 The problem

1.1 Sequence matching

DNA consists of 2 long strand of nucleotides, in the shape of a double helix. Mathematically, a strand can be represented by a string composed of 4 letters. Enormous laboratory effort has been made to compile and compare genetic information from living organisms. One of the tasks is to measure the similarity of two strings. Among others, approximate string matching is a popular technique in this context. There are two reasons why approximate matching is important. On one hand, both measurement, e.g. sequencing, errors and fuzzy nature of underlying molecular processes, e.g. hybridization may occur despite mismatches. On the other hand, "redundancy in biology with evolutionary processes resulting in closely related, yet, different sequences that require approximate matching in order to detect their relatedness and identify variable as well as conserved features that may reveal fingerprints of structure and function"(Meller, 2004).

To illustrate the concepts, consider the sequence GCGAT and GGATT. The exact consecutive matching will give us one match as follows.

$$\begin{array}{ccccc} G & C & G & A & T \\ \updownarrow & & & & \\ G & G & A & T & T \end{array}$$

Whereas comparisons allowing mismatches will yield two matches.

$$\begin{array}{ccccc} G & C & G & A & T \\ \updownarrow & & & & \updownarrow \\ G & G & A & T & T \end{array}$$

In biology, the term, indels, is an abbreviation of insertions and deletions, referring to DNA mutations. If indels are allowed, we will have 4 matches, i.e.

$$\begin{array}{cccccc} G & C & G & A & T & - \\ \updownarrow & & \updownarrow & \updownarrow & \updownarrow & \\ G & - & G & A & T & T \end{array}$$

Given 2 strings of length n and m, their comparison is summarized as an $n \times m$ matrix, and

a matching between letters in positions i and j is traditionally represented by a dot in the corresponding position in the matrix.

1.2 The problem

A natural question arising from these comparisons is how likely a comparison detects an unusual congruence shared among the strings. Such statistical problems are naturally cast in the usual hypothesis-testing context in which we need to compute the tail probability (the biologists' $p - value$) for a seemingly unusual event(Arratia, 1990). To facilitate our discussion, we formulate the problem as follows.

Let $A_1, \cdots, A_n$ and $B_1, \cdots, B_n$ be independently chosen from a common alphabet $\{1, 2, \ldots, d\}$ according to a common distribution $\{\mu_l; l = 1, \ldots, d\}$. Choose a test value t and compute,

$$M_n(t) = \max_{1 \le i,j \le n-t+1} \sum_{k=0}^{t-1} \mathbf{1}\{A_{i+k} = B_{j+k}\}, \tag{1}$$

the largest number of matches witnessed by any comparison of length t substrings. What is the distribution of $M_n(t)$? It is possible to answer this question via Bonferroni inequality. But this technique requires computing arbitrarily large moments, which is tedious. In this respect, the reader is referred to Watson (1954), and Karlin and Ost (1987) for relevant investigation. As a tool for asymptotic analysis, the Chen-Stein method provides an promising alternative to the challenge.

Enormous amount of research has been and is being conducted on DNA and protein sequence matching. In the context of applying the Chen-Stein method to investigate distributional properties, we provide a list of relevant literatures in what follows.

For the longest match between 2 random sequences when only mismatches are allowed, see Arratia, Gordon and Waterman (1990). Neuhauser (1994) extends the results to cover indels. To generalize matches and indels to scoring functions, Arratia, Gordon and Waterman (1988)derived an Erdö-Rényi type strong limit theorem for the highest-scoring matching subsequence between 2 sequences. Karlin and Altschul (1990) generalized the results to more general scoring. Waterman and Vingron (1994) extend Poisson approximation techniques using the Aldous clumping heuristic to estimate statistical significance of observed scores. Regarding the longest matching subsequences of fixed length between 2 sequences when a

certain number of mismatches are allowed, Marianne Mansson (2000) used the Chen-Stein method to bound the total variation distance between the distribution of a suitably chosen compound Poisson distribution.

2 The Chen-Stein Method

2.1 Introduction to the Chen-Stein method

Charles Stein(1972) first introduced the method to prove approximation theorems in probability theory in the context of normal distribution. Chen, Barbour and others adapted and developed the method for other probability distributions including Poisson and compound Poisson.

To get an idea what the Chen-Stein method is, consider the following situation. Let $(S, \mathcal{S}, \mu)$ be a probability space, let χ be the set of measurable functions $h : S \to \mathbb{R}$, and let $\chi_0 \subset \chi$ be a set of $\mu-$integrable functions. Suppose we want to compute $\int_S h d\mu$, but μ is so complicated that the calculation is very forbidding. If we allow our solution to be approximately right and trade for a much easier situation, a natural way is to construct another probability measure μ_0 which, ideally, is much simpler than μ while still close to μ. The goal of the Chen-Stein method is to provide a systematic way to construct such a measure. Specifically, choose a probability measure μ_0 on $(S, \mathcal{S})$ such that all $h \in \chi_0$ are μ_0-integrable, and $E_{\mu_0} h$ is easy to compute. Find a set of functions $\mathcal{F}_0$ and a mapping $T_0 : \mathcal{F}_0 \to \chi$, such that, for each $h \in \chi_0$, the equation

$$T_0 f = h - \int_S h d\mu_0 \tag{2}$$

has a solution $f \in \mathcal{F}_0$. Then,

$$\int_S (T_0 f) d\mu = \int_S h d\mu - \int_S h d\mu_0.$$

T_0 is called a Stein operator for the distribution μ_0, and equation (2) is called a Stein equation, with solution f referred to as a Stein transform of h. The key is to choose a Stein operator in such a way that good estimates of $\int_S (T_0 f) d\mu$ can be obtained.

There are 2 standard methods to construct a Stein operator. The first was proposed by Stein (1986), which consists of the following 3 steps.

1. Choose a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ containing an exchangeable pair of random variables (X, Y) (i.e., their joint distribution should be permutation invariant) with marginal distribution μ_0.

2. Choose a mapping: $\alpha : \mathcal{F} \to \mathbb{F}$ where $\mathbb{F}$ is the space of measurable antisymmetric functions $F : S^2 \to \mathbb{R}$ such that $E(|F(X, Y)|) < \infty$.

3. Take $T_0 = T \circ \alpha$, where $T : \mathcal{F} \to \chi$ is defined, for some version of conditional expectation, by
$$(TF)(x) = E(F(X, Y)|X = x) \quad \forall x \in S$$

This procedure implies,

$$\int_S (T_0 f) d\mu_0 = \int_S (TF) d\mu_0 = E(F(X, Y)) \qquad \forall f \in \mathcal{F}_0,$$

where $F = \alpha f$. By the antisymmetry of F,

$$\int_S (T_0 f) d\mu_0 = 0 \quad \forall f \in \mathcal{F}_0. \tag{3}$$

The second approach to construct Stein operators was proposed by Chen (1998), which makes use of adjoint operators. Specifically, choose a linear mapping: $A : \mathcal{F}_0 \to L^2(S, \mu_0)$ so that the constant function 1 on S is in A^*. Then,

$$\int_S (Af) d\mu_0 = \int_S (A^*1) f d\mu_0 \qquad \forall f \in \mathcal{F}_0$$

Set $T_0 = A - (A^*1)I$.

Further, we see that,

$$\int_S (T_0 f) d\mu_0 = \int_S (Af) d\mu_0 - \int_S (A^*1) f d\mu_0 = 0 \qquad \forall f \in \mathcal{F}_0. \tag{4}$$

Equations (3) and (4) are called Stein identity which is a necessary condition for a Stein operator.

For Poisson distribution, the Stein operator is,

$$(T_0 f)(k) = \lambda f(k+1) - k f(k) \tag{5}$$

It has the property that, for any random variable W, $E(T_0 f)(W) = 0$ for all f such that $\sup_k k|f(k)| < \infty$, if and only if $W \sim Po(\lambda)$. Now, let W be a sum of Bernoulli random variables, each with small expectation. Then, it is natural to expect W approximately follows a Poisson distribution if $E(T_0 f)(W)$, i.e., a measure of error, is small. The advantage of the Chen-Stein method is that it translates the properties of the testing function, h, into desirable properties of f through the Stein operator (5).

The research on the Chen-Stein method has been fruitful since it was first proposed by Stein. For theoretical developments, see Barbour and Eagleson, 1983, 1984; Barbour and Hall, 1984a; Barbour, 1987; Arratia, Goldstein and Gordon, 1989; Barbour, Holst and Janson, 1988b. For applications and examples, see Barbour, 1982; Bollobas, 1985; Holst, 1986; Janson, 1986; Stein, 1986; Barbour, Holst and Janson, 1988; Heckman, 1988; Barbour and Holst, 1989; and Holst and Janson, 1990.

To facilitate our discussion, we cite some of the relevant results on approximation error here.

We will use the following notions. There is a finite or countable index set I. For each $\alpha \in I$, let X_α be a Bernoulli random variable with $p_\alpha = P(X_\alpha = 1) > 0$. Let

$$W = \sum_{\alpha \in I} X_\alpha \text{ and } \lambda = EW. \tag{6}$$

We assume $\lambda \in (0, \infty)$. Z will denote a Poisson random variable with the same mean as W. For each $\alpha \in I$, suppose we have chosen $B_\alpha \subset I$ with $\alpha \in B_\alpha$. B_α is a neighborhood of α consisting of the set of indices β such that X_α and X_β are dependent.

Define

$$b_1 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} p_\alpha p_\beta, \tag{7}$$

$$b_2 = \sum_{\alpha \in I} \sum_{\alpha \neq \beta \in B_\alpha} p_{\alpha\beta}, \text{ where } p_{\alpha\beta} = E[X_\alpha X_\beta], \tag{8}$$

and

$$b_3 = \sum_{\alpha \in I} E|E\{X_\alpha - p_\alpha | \sigma(X_\beta : \beta \notin B_\alpha)\}| \tag{9}$$

Loosely, b_1 measures the neighborhood size, b_2 measures the expected number of neighbors of a given occurrence and b_3 measures the dependence between an event and the number of occurrences outside its neighborhood.

Computing b_1 and b_2 usually involves the same work as computing the first and second moments of W. In applications where X_α is independent of the collection $X_\beta : \beta \notin B_\alpha$, the term $b_3 = 0$. When $b_3 = 0$, $b_2 - b_1 = E(W^2) - E(Z^2)$. Thus when $b_3 = 0$ and b_1 is small, the upper bounds on total variation distance given in the theorems below are comparable to the discrepancy between the second moment of W and that of the Poisson.

When b_1, b_2 and b_3 are all small, then we have the following theorems (Arratia, Goldstein and Gordon, 1989).

Theorem 1. *Let $W = \sum_{\alpha \in I} X_\alpha$ be the number of occurrences of dependent events,and let Z be a Poisson random variable with $EZ = EW = \lambda < \infty$. Then*

$$\begin{aligned} &\| \mathcal{L}(W) - \mathcal{L}(Z) \| \\ \leq & 2\left[(b_1 + b_2)\frac{1 - e^{-\lambda}}{\lambda} + b_3(1 \wedge 1.4\lambda^{-0.5})\right] \\ \leq & 2(b_1 + b_2 + b_3), \end{aligned} \tag{10}$$

and

$$\begin{aligned} &|P(W = 0) - e^{-\lambda}| \\ \leq & (b_1 + b_2 + b_3)(1 - e^{-\lambda})/\lambda \\ < & (1 \wedge \lambda^{-1})(b_1 + b_2 + b_3), \end{aligned} \tag{11}$$

Theorem 1 says the total number W of events is approximately Poisson.

The next theorem is a process version of the above theorem.

Theorem 2. *For $\alpha \in \boldsymbol{I}$,let Y_α be a random variable whose distribution is Poisson with mean p_α, with the Y_α mutually independent. The total variation distance between the dependent Bernoulli process $\boldsymbol{X} \equiv (X_\alpha)_{\alpha \in I}$, and the Poisson process $\boldsymbol{Y}$ on $\boldsymbol{I}$ with intensity $p_{(.)}$, $\boldsymbol{Y} \equiv (Y_\alpha)_{\alpha \in I}$, satisfies*

$$\| \mathcal{L}(\mathbf{X} - \mathbf{Y}) \| \leq 2(2b_1 + 2b_2 + b_3). \tag{12}$$

Theorem 2 implies the locations of the dependent events approximately form a Poisson process.

Theorem 3 compares the dependent Bernoulli process $\mathbf{X}$ with an independent Bernoulli process $\mathbf{X}'$. Since $\sum_\alpha p_\alpha^2 \leq b_1$, Theorem 3 implies that if the Chen-Stein method succeeds with b_1, b_2 and b_3 all small, then in the sense of total variation distance the dependent $\mathbf{X}$ process is close to being independent.

Theorem 3. *For $\alpha \in I$, let X'_α have the same distribution as X_α, with the X'_α mutually independent. The total variation distance between the dependent Bernoulli process $\mathbf{X} \equiv (X_\alpha)_{\alpha \in I}$, and the independent Bernoulli process $\mathbf{X}' \equiv (X'_\alpha)_{\alpha \in I}$ having the same marginals, satisfies*

$$\| \mathcal{L}(\mathbf{X}) - (\mathbf{X}') \| \leq 2(2b_1 + 2b_2 + b_3) + 2\sum p_\alpha^2 \tag{13}$$

Theorem 3 The dependent events are almost indistinguishable from a collection of independent events having the same marginal probabilities.

2.2 Birthday Problem

The birthday problem can be found Chen (1975a), Diaconis and Mosteller (1989), Janson (1986), Holst (1986) and Stein (1987).

Assume birthdays of n individuals are independently and uniformly distributed over d days in a year. Our interest is computing the probability that k people share the same birthday ($k-$way coincidence) occurring at once.

Let $\{1, 2, \cdots, n\}$ denote a group of n people, and let $I = \{\alpha \subset \{1, 2, \cdots, n\} : |\alpha| = k\}$. For example, to compute the probability that there is at least one occurrence of 2 people share the same birthday, we set $k = 2$ and I is the set of all pairs of people among whom a two-way coincidence could occur. Let X_α be the indicator of the event that the people indexed by α share the same birthday. The total number of coincidences is, $W = \sum_{\alpha \in I} X_\alpha$.

Because W is the sum of many Bernoulli random variables, each with small success probability $p_\alpha = d^{1-k}$, it is reasonable to approximate W as a Poisson random variable Z with mean $\lambda = EW = \binom{n}{k} d^{1-k}$. Thus, the probability of no birthday coincidence is approximately

$$P(Z = 0) = e^{-\lambda} = exp\left\{-\binom{n}{k} d^{1-k}\right\}.$$

For the special case when $k = 2, d = 365, n = 23$ is the least number of people required to make $Pr(\textit{at least one coincidence}) > 0.5$, which can be verified by the fact that $\binom{23}{2}/365 - ln(2) = 0.6931507 - 0.6931472 = 0.0000028$.

Now we employ Theorem 1 to bound the error the approximation. Since $\alpha \cap \beta$ implies X_α and X_β are independent, we define

$$B_\alpha = \{\beta \in I : \alpha \cap \beta \neq \varnothing\}$$

. With this choice,

$$E|E\{X_\alpha - p_\alpha | \sigma(X_\beta : \beta \notin B_\alpha)\}| = 0$$

by independence; hence $b_3 = 0$. Further,

$$\begin{aligned} b_1 &= |I||B_\alpha| p_\alpha^2 \\ &= \binom{n}{k}\left\{\binom{n}{k} - \binom{n-k}{k}\right\} d^{2-2k}. \end{aligned} \tag{14}$$

In the case $k = 2$, since X_α and X_β are pairwise independent, $p_{\alpha\beta} = p_\alpha p_\beta$. Therefore,

$$b_2 = |I|(|b_\alpha| - 1) p_{\alpha\beta} = \frac{b_1(|b_\alpha| - 1)}{|B_\alpha|}.$$

Consequently, the bound for the error in approximating $P(W = 0)$ by $e^{-\lambda}$ in the case $k = 2$ is,

$$|P(W=0) - e^{-\lambda}| \leq (b_1 + b_2)\frac{1-e^{-\lambda}}{\lambda}$$
$$= \frac{1}{d^2}\binom{n}{2}(4n-7)\frac{1-e^{-\lambda}}{\lambda}.$$

When k=3, it is much more difficult to compute the exact probability of a 3-way coincidence. Nevertheless, we can easily apply Poisson approximation to the problem. Suppose we want to compute the probability that in a group of 50 people, there is at least one triple coincidence. We have $\lambda = \binom{n}{3}/d^2$ and hence the desirable probability is about

$$1 - P(W=0) = 1 - e^{-\lambda} = 1 - 0.863 = 0.137.$$

To obtain a bound on the error, we proceed as follows.

$$b_1 = |I||B_\alpha|p_\alpha^2$$
$$= \binom{n}{3}\left\{\binom{n}{3} - \binom{n-3}{3}\right\}d^{-4},$$

and, for a given α, breaking up $B_\alpha - \{\alpha\}$ into those β such that $|\beta \cap \alpha| = 1$ and those for which $|\beta \cap \alpha| = 2$, we see

$$b_2 = |I|\left\{3\binom{n-3}{2}d^{-4} + 3(n-3)d^{-3}\right\}.$$

This shows the approximation above has an error of no more than

$$(b_1 + b_2)(1 - e^{-\lambda})/\lambda = 0.0597,$$

so that $0.803 \leq P(W=0) \leq 0.923$.

To assess the bound on the Chen-Stein error, we now derive the formula for the exact probability when $k = 3$. In order for there to be no triple coincidence, the d days of the year must be partitioned into h days when there are no birthdays, i days on each of which a single individual was born, and j days on each of which exactly 2 individuals share a

birthday. A factor of $n!/2^j$ is needed to count the number of arrangements of n persons into such a configuration of $i + j$ days. Now we have,

$$P(W = 0) = d^{-n} \sum_{i+2j=n} \binom{d}{h, i, j} \frac{n!}{2^j}.$$

For $n = 50$ and $d = 365$, we have that $P(W = 0) = 0.8736$, for an actual error of $0.8736 - 0.9632 = 0.0104 < 0.0597$, the Chen-Stein bound on the error.

3 Application

In this section, we will compare the results from the theoretical prediction based on the Chen-Stein method, computer simulation and real data. We first describe the setup for each situation, and the results will follow in section 4.

3.1 Theoretical prediction

According to the above introduction to the Chen-Stein method, the main idea of the theoretical prediction is to approximate the actual distribution with a Poisson distribution through proper formulation of the problem.

To help us understand our approach to solve the problem given above, let's first consider a closely related problem (Arratia et al, 1990).

Let $\{Z_i \, i \in \mathbb{Z}\}$ be an independent coin tosses with $p = P(Z_i = 1) = 1 - P(Z_i = 0)$, and let $S_{n;t}$ be the maximum number of heads occurring in a window of length t, starting within the first n tosses, $S_{n;t} \equiv \max_{1 \le i \le n}(Z_i + \cdots + Z_{i+t-1})$. We are interested in the distribution of $S_{n;t}$.

For integer α and positive integers $s \le t$, define indicators

$$Y_\alpha \equiv Y(\alpha, s, t) \equiv 1 \left[s = \sum_{k=0}^{t-1} Z_{\alpha+k} \right],$$

Intuitively, Y_α is the indicator that a window of length t containing s heads starts at α.

It is easy to see that Y_α and $Y_{\alpha+1}$ are dependent, and such windows may overlap. We say these windows occur in clumps, which would make the analysis directly based on Y_α very demanding. Instead, we will work with the number of clumps by using the Poisson clumping heuristic (Aldous, 1989). Define

$$X_\alpha \equiv X(\alpha, s, t) \equiv \prod_{j=1}^{t}(1 - Y_{\alpha-j}).$$

Let $I \equiv \{1, 2, \cdots, n\}$, and define

$$W \equiv W(n, s, t) \equiv \sum_{\alpha \in I} X_\alpha.$$

In another word, X_α is the indicator that a clump starts at α. The random variable W is the number of clumps that begin within the first n tosses.

We now present a result on $\frac{EY_1}{EX_1}$, which can be roughly interpreted as the clump size. The proof which employs the ballot theorem(c.f. Feller (1968)) can be found in Arratia et al (1990).

Lemma 1. *Let s and t be positive integers with $s \leq t$, and let $a \equiv s/t$. Then,*

$$\begin{aligned} a - p \leq \frac{EX_1}{EY_1} &\leq a - p + 2(1-a)P\left(\sum_{j=1}^{t} Z_j > s\right) \\ &\leq a - p + 2(1-a)e^{-tH(a,p)} \end{aligned} \tag{15}$$

where

$$H(a, p) = a log\left(\frac{a}{p}\right) + (1-a) log\left(\frac{1-a}{1-p}\right).$$

Apart from "boundary effect," the event $\{S_{n;t} < s\}$ agrees with the event $\{W = 0\}$. The error in this approximation can be controlled by observing that $W \neq 0 \subset \{S_{n;t} \geq s\}$, and

$$\{S_{n;t} \geq s, W = 0\} \subset \{Y_1 + \cdots, +Y_t > 0\} \cup \{Z_1 + \cdots, +Z_t > s\}.$$

Hence,

$$0 \leq P(W = 0) - P(S_{n;t} < s) \leq tEY_1 + P(Z_1 + \cdots, +Z_t > s). \tag{16}$$

The Chen-Stein method can be applied to establish a Poisson approximation for W. Let $\lambda \equiv EW$. The indicator random variable X_α is measurable with respect to the $2t$ coins Z_j

at $\alpha - t, \ldots, \alpha + t - 1$. Thus, we define the neighborhood,

$$B_\alpha \equiv \{\beta \in I : |\alpha - \beta| < 2t\} \qquad for\ \alpha\ =\ 1\ to\ n,$$

so that $b_3 = 0$ and $b_1 < (4t-1)\lambda EX_1$. If $|\alpha-\beta| \leq t$, then $E(X_\alpha X_\beta) = 0$, but if $t < |\alpha-\beta| < 2t$, then we can only conclude that $E(X_\alpha X_\beta) \leq X_\alpha X_\beta$, so that $b_2 < 2t\lambda EY_1$.

Using Theorem 1, we have

$$\begin{aligned} |P(W = 0) - e^{-EW}| &\leq (b_1 + b_2)(1 \wedge 1/\lambda) \\ &\leq 2t\lambda(2EX_1 + EY_1)(1 \wedge 1/\lambda) \\ &\leq 6tEY_1. \end{aligned} \tag{17}$$

Combining (16) and (17), and rewriting EY_1 in term of Z_i, we have

$$|P(S_{n;t} - e^{-EW})| \leq 7tP(Z_1 + \cdots + Z_t = s) + P(Z_1 + \cdots + Z_t > s). \tag{18}$$

Now, $\lambda \equiv \lambda(n, s, t) \equiv EW \equiv nEX_1$, which, combined with lemma 1, gives,

$$\frac{s}{t} - p \leq \frac{EW}{nP\left(\sum_{j=1}^{t} Z_j = s\right)} \leq \frac{s}{t} - p + 2\left(1 - \frac{s}{t}\right) P\left(\sum_{j=1}^{t} Z_j > s\right). \tag{19}$$

Summarizing these findings, we have the following theorem.

Theorem 4. *Let $\{Z_i\, i \in \mathbb{Z}\}$ be an independent sequence with $p = P(Z_i = 1) = 1 - P(Z_i = 0)$, and let $S_{n;t} \equiv \max_{1 \leq i \leq n}(Z_i + \cdots + Z_{i+t-1})$. For all positive integers n, s, t, with $s \leq t$ and $s/t > p$, $P(S_{n;t} < s)$ is approximately $exp\left(-n\left(\frac{s}{t} - p\right) P(Z_1 + \cdots + Z_t = s)\right)$, with the error in this approximation controlled by* (18) *and* (19).

For the problem of comparing 2 sequences, i.e. the sequence matching problem, the derivation is more involved. We present the net result in Theorem 2 and the interested reader is referred to Arratia et al (1990).

Theorem 5. *Let $M_n(t_n)$ be the maximum number of matches between a word of length t_n taken from $A_1 \ldots A_n$ and a word of length t_n from $B_1 \ldots B_n$, with independent letters from*

alphabet $\{1, \ldots, d\}$ *according to distribution* $\mu_l, l = 1, \ldots, d$. *Then,*

$$P\{M_n(t_n) < s\} - exp\left(-\left(\frac{s}{t_n} - p\right) n^2 P\{binomial(t_n, p) = s\}\right) \to 0, \tag{20}$$

as $n \to \infty$ *whenever* $t_n/ln(n^2) \to c > 1/ln(1/p)$ *and* $c < \infty$, *where* $p = \sum_{l=1}^{d} \mu_l^2$

For the DNA sequence of Indian corn, $p = \sum_{l=1}^{4} \mu_l^2$, where $\mu = (0.3083, 0.1917, 0.1929, 0.3071)$ is the probabilities for the alphabet $\{a, c, g, t\}$ which is calculated from the sequence.

Roughly, this result can be interpreted in the following steps.

1. There are about n^2 blocks of length t requiring comparison.
2. The approximation theory and techniques, e.g. declumping reduces the effective number of comparisons to $(s/t_n - p)n^2$.
3. Multiplying the number of comparisons by the binomial probability gives the expectation.
4. The Poisson probability of seeing zero events gives the desirable approximation.

3.2 Real data result

The complete chloroplast genome of the Indian corn *Zea Mays*, was retrieved from the GenBank database. The genome is given as a sequence of 140,325 letters from the alphabet $\{a, c, g, t\}$. The sequence was cut into 274 blocks of exactly 512 letters with the remaining letters ignored. A simple random sample of 200 pairs from the population of all block pairs was taken, and the probabilities of maximum number of matches were computed.

3.3 Computer simulation

Assuming the letters generated independently from the alphabet $\{a, c, g, t\}$ with probability $\mu = (0.3083, 0.1917, 0.1929, 0.3071)$. Specifically, we first generated a sequence of 140,325 letters, and computed the desirable probabilities as if the sequence is real as above.

4 Summary and Discussion

In this section, we present the results from the above three setups and evaluate the performance of the theoretical prediction. Table 1 summarizes these results and figure 1 visualizes the comparisons.

Table 1: classical table

		$\hat{p}$	99% CI*
14	Real	0.124	(0.079, 0.170)
	Sim.	0.134	(0.088,0.187)
	Pred.	0.092	—
15	Real	0.529	(0.461, 0.598)
	Sim.	0.630	(0.561,0.699)
	Pred.	0.618	—
16	Real	0.290	(0.227, 0.352)
	Sim.	0.193	(0.136,0.249)
	Pred.	0.253	—
17	Real	0.043	(0.015,0.071)
	Sim.	0.040	(0.012,0.068)
	Pred.	0.035	—
18	Real	0.0136	(0.000, 0.029)
	Sim.	0.000	—
	Pred.	0.029	—

*: Bonferroni adjusted for multiple comparisons.

Overall, the main trends of the 3 results are consistent in order of magnitude, which indicates the Poisson approximation gives a good prediction of the big picture of the empirical distribution.

4.1 Simulation vs. prediction

On a gross basis, the simulated distribution is well approximated by the Poisson approximation, but its fine structure is not very well predicted. The shape of the predicted distribution has a thinner left tail and thicker right tail than the simulated distribution. This is not so surprising, given that the binomial distribution has a heavier right tail and lighter left tail than the extreme value.

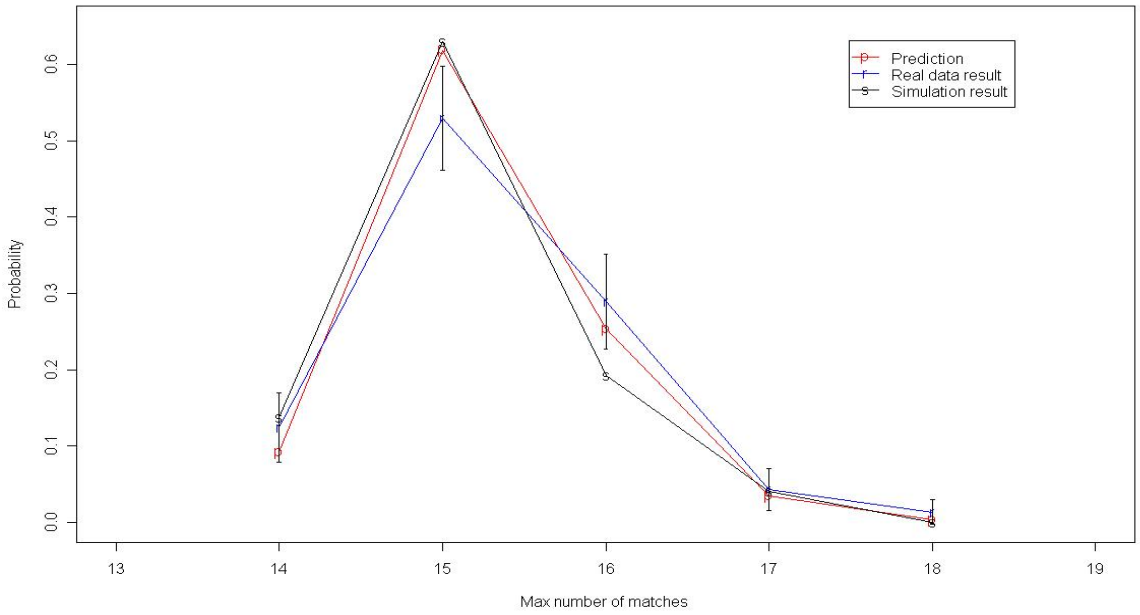

Figure 1: The results from Poisson approximation, real data and computer simulation. The 99% confidence intervals for the real data are also plotted.

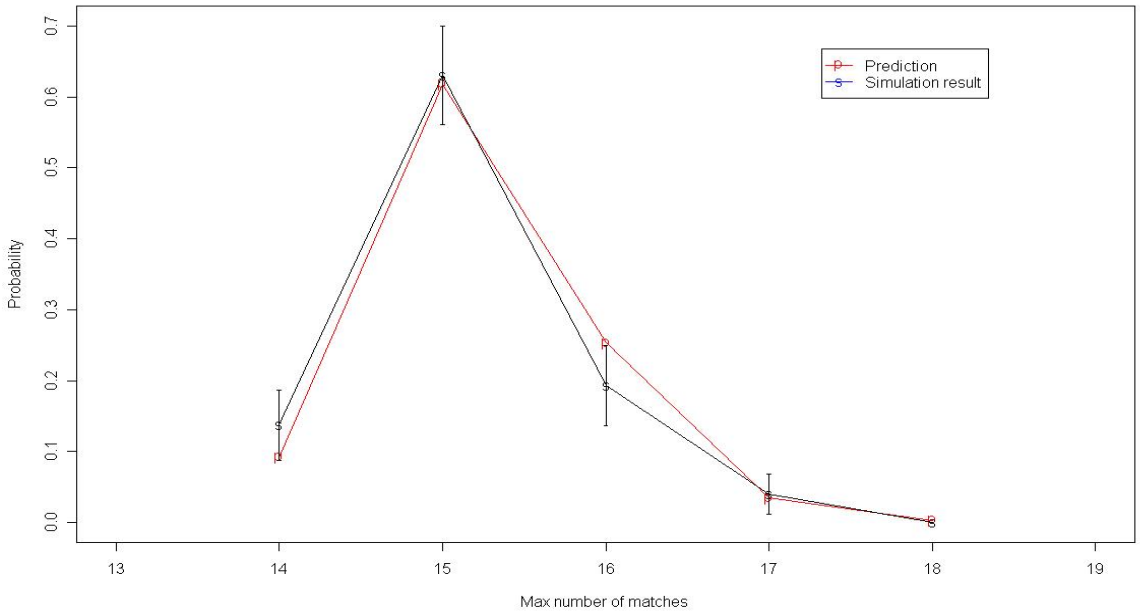

Figure 2: The results from Poisson approximation and computer simulation. The 99% confidence intervals for the simulation are also plotted.

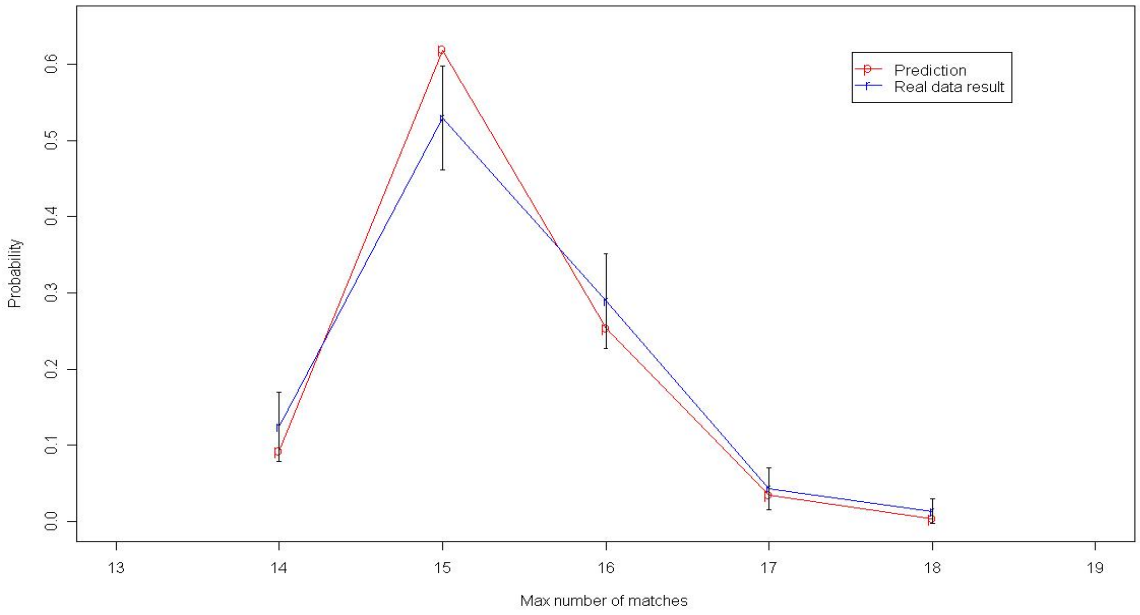

Figure 3: The results from Poisson approximation and real data. The 99% confidence intervals for the real data are also plotted.

Testing the equality of the simulated and predicted distribution using Pearson's chi-square test(see appendix) gives $p - value = 0.022$, which implies the difference between these two distributions is not statistical significant at the 99% confidence level.

4.2 Real data result vs. prediction

Compared with the simulated distribution, the empirical distribution from the real data is less concentrated. The Poisson approximation again roughly captures the shape of the empirical distribution.

The discrepancy between the theoretical prediction and the real data result gets larger when the target probabilities increases. In particular, the theoretical prediction differs the most from the empirical result when the number of matches equals 15, where the probability reaches the maximum.

The likelihood ratio test(see appendix) statistic will be referred to a χ^2_4 distribution and find the $p-value$. Our result shows that $p-value < 0.001$, which indicates that at least one

estimated real probability is different from the predicted probability. To investigate which categories are significantly different, we check table 1. Note that only at the richest matches, 15, is the difference statistically significant where the 99% confidence interval constructed from the real data does not contain the predicted probability. It is worth noticing that we are most likely interested in unusually large number of matches in practice because these matches contains more genetic information than rich matches.

5 References

Aldous, D. (1989). Probability Approximations via the Poisson Clumping Heuristic. Springer, New York.

Meller, J., (2004). Course lecture for *Introduction to Bioinformatics*. http://folding.cchmc.org/intro2bioinfo/intro2bioinfo.html.

Arratia, R., Goldstein, L. and Gordon, L. (1990). Poisson Approximation and the Chen-Stein Method. Statistical Science, Vol. 5, No.4, 403-434.

Arratia, R., Gordon, L. and Waterman, M. S. (1986). An extreme value distribution for sequence matching. Ann. Statist. 14971-993.

Arratia, R., Gordon, L. and Waterman, M. S. (1990). The Erdös-Rényi Law in distribution, for coin tossing and sequence matching. Ann. Statist. 18 539-570.

Arratia, R., Morris, P. and Waterman, M. S. (1988). Stochastic Scrabble: a law of large numbers for sequence matching with scores. J. Appl. Probab. 25 106-119.

Chen, L. H. Y. (1975a). Poisson approximation for dependent trials. Ann. Prob. 3 534-545.

Chen, L. H. Y. (1998). Stein's method: Some perspectives with applications. Probability Towards 2000. Lecture Notes in Statist. 128 97–122. Springer, Berlin.

Diaconis, P. and Holmes, S. (2004). Steins Method: Expository Lectures and Applications Institute of Mathematical Statistics Lecture Notes Monograph Series Vol. 46.

Diaconis, P. and Mosteller, F. (1989). Methods for studying coincidences. J. Amer. Statist. Assoc. 84 853-86l.

Erhardsson, T. (2005) Stein's method for Poisson and compound Poisson approximation, in An introduction to Stein's method, 61–113, Singapore Univ. Press, Singapore.

Feller, W. (1968). An Introduction to Probability Theory and its Applications 1, 3rd ed. Wiley, New York.

Holst, L. (1986). On birthday, collectors', occupancy and other classical urn problems.

Internat. Statist. Rev. 54 15-27.

Janson, S. (1986). Birthday problems, randomly colored graphs, and Poisson limits of dissociated variables. Tech. report 1986 16. Dept. Math., Uppsala Univ.

Karlin, S. and Ost, F. (1987). Counts of long aligned word matches among random letter sequences. Adv. in Appl. Prob. 19.

Månsson, M. (2000). On compound Poisson approximation for sequence matching. Combin. Prob. Comp. 9, 529-548.

Neuhauser, C. (1994). A Poisson approximation for sequence comparisons with insertions and deletions. Ann. Statist. 22.

Stein, C. M. (1972) A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. II . Probability theory, 583-602. Univ. California Press.

Stein, C. M. (1986) Approximate computation of expectattons. IMS lecture Notes Vol. 7. Hayward, California.

Stein, C. M. (1987). The number of monochromatic edges in a graph with randomly colored vertices. Unpublished manuscript.

Waterman, M. S. and Vingron, M. (1994).Sequence Comparison Significance and Poisson Approximation. Statist. Sci. Volume 9, Number 3 (1994), 367-381.

Watson, G. S. (1954). Extreme values in samples from m-dependent stationary stochastic sequences. Ann. Math. Statist. 25 798-800.

6 Appendix

6.1 Multinomial test

Multinomial test tests the null hypothesis that the parameters of a multinomial distribution equal some specified values.

Let $\mathbf{x} = (x_1, \ldots, x_k)$ be a random sample from a multinomial distribution, $Mult(N, p_1, \ldots, p_k)$. Suppose we are interested in testing $H_0 : p_i = p_i^0, \quad \forall i = 1, \ldots, k$ vs. $H_a : p_i \neq p_i^0, \quad for\ at\ least\ one\ i \in \{1, \ldots, k\}$

We first introduce the likelihood ratio test. Under H_0, the exact probability of the observation $\mathbf{x}$ is,

$$P_0(\mathbf{x}) = N! \prod_{i=1}^{k} \frac{(p_i^0)^{x_i}}{x_i!}.$$

Under H_a when $p_i = \hat{p}_i,\ i = 1, \ldots, k$ where $\hat{p}_i = \frac{x_i}{N}$, the maximum likelihood estimate, the exact probability of the observation $\mathbf{x}$ is,

$$P_a(\mathbf{x}) = N! \prod_{i=1}^{k} \frac{(\hat{p}_i)^{x_i}}{x_i!}.$$

The likelihood ratio test statistic is given by,

$$-2ln(LR) = -2 \sum_{i=1}^{k} x_i ln(p_i/\hat{p}_i).$$

This statistic asymptotically follows a chi-square distribution with $k-1$ degrees of freedom.

The hypotheses can be tested using the Pearson's chi-square test statistic,

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - E_i)^2}{E_i}$$

where $E_i = Np_i^0,\ i = 1, \ldots, k$ is the expected count under H_0. As the likelihood ratio statistic, χ^2 also asymptotically follows a chi-square distribution with $k-1$ distribution. For finite sample, the $-2ln(LR)$ converges to the chi-square distribution from above(Lawley,

1956), whereas the chi-square test statistic converges from below. This implies that the likelihood ratio test tends to inflate type I errors, i.e. false positives, whereas the Pearson's chi-square test deflate them.

Lawley, D. N. (1956). "A General Method of Approximating to the Distribution of Likelihood Ratio Criteria". Biometrika 43: 295303.

6.2 The ballot theorem

Suppose that candidates A and B are in an election. A receives a votes and B receives b votes, with $a > b$. Then there are $\frac{a-b}{a+b}\binom{a+b}{a}$ out of $\binom{a+b}{a}$ voting configurations so that A maintains a lead throughout the counting of the ballots.

A more general version of the theorem is as follows,

Suppose $a > kb$ for some positive integer k. There are $\frac{a-kb}{a+b}\binom{a+b}{a}$ out of $\binom{a+b}{a}$ ways so that A maintains a lead throughout the counting of the ballots.