On the distribution of k-tuple matches for sequence homology: a constant time exact calculation of the variance

Gary Benson* Xiaoping Su†

September 18, 1997

Abstract

We study the distribution of a statistic useful in calculating the significance of the number of k-tuple matches detected in biological sequence homology algorithms. The statistic is $R_{n,k}$, the total number of heads in head runs of length k or more in a sequence of iid Bernoulli trials of length n. Calculation of the mean is straightforward. Poisson approximation formulas have been used for the variance because they are simple and powerful. Unfortunately, when $p = P(Head)$ is large, the Poisson approximation no longer works well. In our application, p is large, say .75, and we have turned instead to direct calculation of the variance. Surprisingly, we are able to show that the variance, which is based on the interactions of $O(n^2)$ random variables, can be computed in *constant time*, independent of the length of the sequence and probability p. This result can be used to calculate the mean and variance of a number of other head run statistics in constant time. Additionally, we show how to extend the result to sequences generated by a stationary Markov process where the variance can be calculated in $O(n)$ time.

1 Introduction

Many stochastic processes can be represented by a sequence of independent and identically distributed (iid) Bernoulli trials, *i.e.*, there are two possible outcomes in each trial with p being the probability of the first outcome and $q = (1 - p)$ the probability of the second outcome. Tossing a coin is the obvious analogy, with $p = P(Head)$, $q = P(Tail)$ and frequently, Bernoulli trials are discussed in terms of a sequence of heads and tails. One is often interested in various *statistics* associated with Bernoulli trials, especially with respect to *runs of heads*. Examples are 1) the longest head run, 2) the number of head runs of a particular size k, and 3) the sum of heads in runs of a particular size. Although these statistics are easily stated and conceptually clear, calculating their distributions is not a simple task.

*Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, NY 10029-6574, benson@ecology.biomath.mssm.edu. Partially supported by NSF grant CCR-9623532

†Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, NY 10029-6574, su@ecology.biomath.mssm.edu. Partially supported by NSF grant CCR-9623532

In the case when p is small, we can use the powerful and elegant Poisson approximation to obtain very accurate estimates of the statistics of head runs. But, when p is not small, Poisson approximation no longer works well. In this paper, we present a new method for accurately and efficiently approximating the distributions of a variety of head run statistics. Our method is insensitive to the value of p. Because the distributions for many statistics can be accurately approximated by the Normal distribution, knowing the variance and the mean is sufficient for their explicit use. While calculating the mean is easy, calculating the variance often is not. In this paper, we show how to compute the exact variance of a statistic, $R_{n,k}$, the sum of heads in head runs of length k or longer in a sequence of length n. The remarkable property of our method is that the covariance of many dependent random variables (and therefore the variance of the statistic) can be calculated in *constant time*. For example, for $R_{n,k}$, the number of random variables is $O(n^2)$ and a naive computation of the variance would take $O(n^4)$ time. Our main result is a constructive proof that the exact variance of the distribution of $R_{n,k}$ can be calculated in constant time.

Our work is motivated by the problem of detecting tandem repeats in DNA sequences. DNA is a long linear molecule which in the famous double helix constitutes our chromosomes. For analysis purposes, a single strand of DNA can be viewed as a sequence of letters over the alphabet $\Sigma = \{A, C, G, T\}$. A tandem repeat is any pattern of letters which occurs two or more times in a row. For example the sequence:

$$ACA\underline{CGT}CGTCGTCGTATATCT$$

contains a tandem repeat consisting of four copies of the pattern CGT. Because DNA is a biological molecule subject to random mutations, tandem repeats typically consist of *approximate copies*. Differences between copies consist of *substitutions* where one letter replaces another and *deletions* and *insertions* where some letters are lost or new letters are added. A real (albeit only moderately mutated) example is shown below. The actual DNA sequence appears on the upper line and a consensus sequence is shown on the lower line (only 2 of 8 copies are shown). A $*$ marks a substitution with respect to the consensus and a – represents an insertion or deletion.

```
  *                   *   *                       *  * **
CCCGCCGCCC--CGTCTGGGATGTGGGGAGCGCCTCTGC  CCGGCCACGACCCCGTCTGGGAAGTGAGGAGC-CCTCTGC
CCGGCCGCCCATCGTCTGGGAAGTGAGGAGCGCCTCTGC  CCGGCCGC-CCATCGTCTGGGAAGTGAGGAGCGCCTCTGC
```

Tandem repeats are of great interest in the biological community because they are responsible for genetic diseases [26, 16, 10, 6] and because they are useful for biological studies, such as gene location and phylogeny reconstruction [8, 25, 1, 21]. The development of a practical algorithm for their detection has been a desireable goal [18, 3, 17, 24, 5]. In a new algorithm designed to detect tandem repeats [4] we use the technique of finding matching k-tuples where, in general, $k > 1$ for reasons of algorithm efficiency. Suppose a DNA sequence contains two adjacent, approximate copies of a pattern. If we align one copy with the other, we can convert them to a sequence of heads and tails by writing a head below each column which contains a match, and a tail below each column which contains a mismatch. For example:

```
C C A C G A C C C C G T C T G G C A A G T G T G G G T C
C T G C A C C A T C G T C T G G G A A G T G A G G A G C
H T T H T T H T T H H H H H H H T H H H H H T H H T T H
```

Let us make an assumption about the *average amount of mutation* that has occurred between the copies by fixing $p = P(Heads)$. For example, we could assume that $p = 0.75$. The matching k-tuples approach finds the head runs of length k or longer. So, here is the problem. **How many heads do we see under our assumption about p?**

This problem was already considered important for calculating the significance of matching scores when performing database searches for biological sequence homology. When a new biological sequence is discovered, one of the first tasks is to screen it against a database like Genbank in order to find similar sequences. One popular program for this purpose is FASTA [20] which uses the sum of matches in matching k-tuples as the measure of similarity. Matching k-tuples are, of course, common and the question naturally arises, "How many matches are required for statistical significance?" Goldstein and Waterman [12] estimated the distribution of $R_{n,k}$ for randomly aligned sequences in order to answer this question. In their case, $p = .25$ and they used a *compound Poisson distribution* to approximate the probability $P(R_{n,k} = x)$. In this method, the number, W, of head runs of size k or larger is assumed to be Poisson and the size X_i of the ith head run is geometric with $P(X_i = j) = qp^j$. Note that in reality W and the X_i are *not independent* because each depends on p. Yet, when p is small, the dependence is small and these variables can be treated as independent. In that case, error bounds can be computed for the difference between the real and approximate distributions using the Chen-Stein technique as shown by Arratia, Goldstein and Gordon [2].

In our case, p is too large for the Chen-Stein error bounds to be useful. This is so because 1) for large p, the number of head runs can not be well approximated by a Poisson distribution and 2) the dependence between the number of head runs and the run length is too great to dismiss. The same would be true if we were examining random sequences with only a two letter alphabet, as for example, the *purine–pyrimidine* classification of DNA bases. In our approach, we approximate the distribution of $R_{n,k}$ by the Normal distribution. Therefore, we have to calculate the mean and variance. The mean is straightforward. Only the variance presents difficulty.

(A proof that $R_{n,k}$ can be well approximated by the Normal distribution is beyond the scope of this paper. For our purposes, we have tested the goodness-of-fit of the Normal approximation for representative values of n, k and p. For example, with $n = 100$, $k = 5$ and $p = .75$, we collected simulated data and performed the chi-square goodness-of-fit test for the null hypothesis that this random variable approximately follows the Normal distribution with mean and variance calculated by the method presented in this paper. The asymptotic approximate P-value is much larger than 0.05, so the null hypothesis is supported by these data.)

Besides the distribution of $R_{n,k}$, our method can be applied to a variety of other statistics involving head runs [2, 9], including 1) the length of the longest head run 2) the number of head runs of size exactly k, 3) the number of head runs of size greater than or equal to k, 4) the number of non-overlapping head runs of size exactly k, and 5) the number of overlapping head runs of size

exactly k.

In our presentation, we assume that the sequences are generated by iid Bernoulli trials. We then show how to extend the result to sequences generated by a stationary Markov process. The compound Poisson approach can also be used for stationary Markov processes [23], but with the same limitation that the probability of heads is small. When the sequence is generated by Bernoulli trials that are *not* identically distributed, (*i.e* the probabilities in each trial are not identical), a different method for calculating the distributions based on Markov chains has been described by Fu and Koutras [9]. While that work can be applied to the distribution of $R_{n,k}$ and the other statistics mentioned above, there are several difficulties. The method is based on matrix multiplication, a different matrix is required for each sequence length n, and the matrix must be raised to the nth power. With respect to iid Bernoulli trials, formulas for the exact distributions of non-overlapping and overlapping head runs of size exactly k have been given [22, 13, 11, 19, 15, 14, 7], but the calculations take on the order of $O(n^2)$ time to determine the probability for each individual value which these statistics can assume and are therefore not useful in practice.

The remainder of this paper is organized as follows. In section 2 we state our problem and show how the mean and variance of $R_{n,k}$ can be calculated from a large set of indicator random variables. In section 3 we show how to calculate the covariances of these variables in $O(n^2)$ time. Finally, in section 4 we show how to compute the covariances in $O(n)$ and constant time. In section 5 we show how to extend our result to sequences generated by a stationary Markov process.

2 Problem Statement and Analysis

Our problem is to determine the exact mean and variance of the random variable $R_{n,k}$ where

$$R_{n,k} = \begin{array}{l}\text{the total number of heads in headruns of length } k \text{ or} \\ \text{longer in an iid Bernoulli sequence of length } n.\end{array}$$

Let us first define several important random variables:

1. Let $A = A_1, A_2, \ldots, A_n$ be a sequence of zero-one valued, independent Bernoulli random variables with head probability p, where head means success and is denoted by the value one.

2. Let $I = \{1, 2, \ldots, n\}$ be an index set whose elements denote locations in the sequence A where head runs can begin.

3. For each $\alpha \in I$, let X_α^j be an indicator random variable. $X_\alpha^j = 1$ means that there occurs the following pattern with *exactly* j heads starting at position α (the position of the underlined character):

$$
\begin{array}{ll}
\cdots T \overbrace{\underline{H} \cdots H}^{j} T \cdots & \alpha = 2, 3, \ldots, n-j, \qquad j \leq n-2 \\
\overbrace{\underline{H} \cdots H}^{j} T \cdots & \alpha = 1, \qquad j \leq n-1 \\
\cdots T \overbrace{\underline{H} \cdots H}^{j} & \alpha = n-j+1, \qquad j \leq n-1 \\
\overbrace{\underline{H} \cdots\cdots H}^{n} & \alpha = 1, \qquad j = n.
\end{array}
$$

Given $A_1, A_2, \ldots, A_n$, the value of the variable X_α^j is defined as follows:

$$
X_\alpha^j = \begin{cases}
(1 - A_{\alpha-1})(1 - A_{\alpha+j}) \prod_{i=\alpha}^{\alpha+j-1} A_i & \text{if } \alpha = 2, 3, \ldots, n-j, \quad j \leq n-2 \\
(1 - A_{1+j}) \prod_{i=\alpha}^{\alpha+j-1} A_i & \text{if } \alpha = 1, \quad j \leq n-1 \\
(1 - A_{n-j}) \prod_{i=n-j+1}^{n} A_i & \text{if } \alpha = n-j+1, \quad j \leq n-1 \\
\prod_{i=1}^{n} A_i & \text{if } \alpha = 1, \quad j = n.
\end{cases}
$$

In terms of the indicator random variables, we define $R_{n,k}$ as:

$$
R_{n,k} = \sum_{j=k}^{n} \sum_{\alpha=1}^{n-j+1} j X_\alpha^j.
$$

The probability that each indicator variable equals one is:

$$
P(X_\alpha^j = 1) = \begin{cases}
q^2 p^j & \text{if } \alpha = 2, 3, \ldots, n-j \text{ and } j \leq n-2 \\
q p^j & \text{if } \alpha = 1 \text{ or } \alpha = n-j+1 \text{ and } j \leq n-1 \\
p^n & \text{if } \alpha = 1 \text{ and } j = n.
\end{cases}
$$

The expectation and variance of $R_{n,k}$ are:

$$
\begin{aligned}
E(R_{n,k}) &= E(\sum_{j=k}^{n} \sum_{\alpha=1}^{n-j+1} j X_\alpha^j) = \sum_{j=k}^{n} \sum_{\alpha=1}^{n-j+1} j E X_\alpha^j \qquad (1) \\
Var(R_{n,k}) &= Var(\sum_{j=k}^{n} \sum_{\alpha=1}^{n-j+1} j X_\alpha^j) \\
&= \sum_{j=k}^{n} \sum_{\alpha=1}^{n-j+1} j^2 Var(X_\alpha^j) \\
&\quad + 2 \sum_{j=k}^{n} \sum_{\alpha=1}^{n-j+1} \sum_{\beta=\alpha+1}^{n-j+1} j^2 Cov(X_\alpha^j, X_\beta^j)
\end{aligned}
$$

$$+2\sum_{j=k}^{n}\sum_{h=j+1}^{n}\sum_{\alpha=1}^{n-j+1}\sum_{\beta=1}^{n-h+1} hjCov(X_\alpha^j, X_\beta^h). \tag{2}$$

We know the mean and variance for each variable X_α^j. They are just:

$$EX_\alpha^j = P(X_\alpha^j = 1)$$

$$Var(X_\alpha^j) = P(X_\alpha^j = 1)(1 - P(X_\alpha^j = 1)).$$

Calculating the mean of $R_{n,k}$ is straightforward. As long as we can determine the covariance between the variable pairs, we will be able to calculate the variance of $R_{n,k}$. It is obvious that the variables X_α^j and X_β^h are independent if the patterns denoted by those variables do not overlap. For those variable pairs, the covariance is zero. If the patterns denoted by a pair of variables do overlap, the relationship between those variables is one of the following two types:

Mutually exclusive relationship. Let us first fix j and α. Then for any X_β^h defined below, if $X_\alpha^j = 1$, then X_β^h must equal zero, that is, $X_\alpha^j = 1 \Longrightarrow X_\beta^h = 0$. Let us present an example as illustration:

$$\begin{array}{lc}
 & X_\alpha^j = 1 \\
 & \downarrow \\
 & \overbrace{}^{j} \\
\text{actual sequence:} & \cdots T\ \overbrace{\underline{H}\cdots H}\, T\cdots \\
\text{excluded patterns:} & T\ \underbrace{\underline{H}\cdots H}_{h}\ T \quad T\underline{H}\cdots HT \\
 & \uparrow\leftarrow \quad \{\beta\} \quad \rightarrow\uparrow \\
 & X_{\beta_1}^h \qquad\qquad X_{\beta_{j+h+1}}^h
\end{array}$$

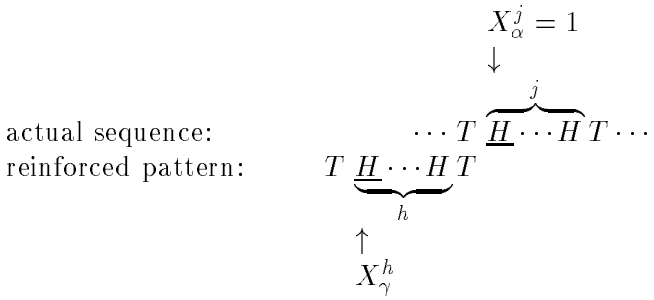
The set $\{\beta\}$ includes all those locations for which $X_\beta^h = 0$. We have shown two excluded patterns. If we ignore the edges of the sequences, the size of the set $\{\beta\}$ is $j + h + 1$. Since the variables X_α^j and X_β^h can *not* take the value one at the same time, we have $EX_\alpha^j X_\beta^h = 0$, and

$$Cov(X_\alpha^j, X_\beta^h) = EX_\alpha^j X_\beta^h - EX_\alpha^j EX_\beta^h = -EX_\alpha^j EX_\beta^h \tag{3}$$

Taking account of the edges of the sequence, equation 3 holds when

$$\begin{aligned}
j, h &= k, k+1, \ldots, n \\
\alpha &= 1, 2, \ldots, n-j+1 \\
\beta &\in \{\beta : \beta \geq 1 \text{ and } \alpha - h \leq \beta \leq \alpha + j \text{ and } \beta \leq n - h + 1\} \\
(\alpha, j) &\neq (\beta, h).
\end{aligned}$$

Reinforcement relationship. Again let us first fix j and α. For any X_γ^h defined below, if $X_\alpha^j = 1$, then the probability that $X_\gamma^h = 1$ increases. For example:



We have shown one reinforced pattern. Ignoring edge effects, γ can take on two values. As illustated, if $X_\alpha^j = 1$, then the probability that $X_\gamma^h = 1$ increases by a factor of $1/q$. To see why, note that the two patterns, denoted by the variables X_α^j and X_γ^h respectively, must share the letter T when $X_\alpha^j = 1$ and $X_\gamma^h = 1$, so one less T is required for these adjacent patterns than for non-adjacent patterns. Meanwhile, the probability of occurrence of letter T is q in Bernoulli trial, so:

$$P\{X_\alpha^j X_\gamma^h = 1\} = (1/q) P\{X_\alpha^j = 1\} P\{X_\gamma^h = 1\}.$$

Therefore $EX_\alpha^j X_\gamma^h = (1/q) EX_\alpha^j EX_\gamma^h$, and

$$\begin{aligned} Cov(X_\alpha^j, X_\gamma^h) &= EX_\alpha^j X_\gamma^h - EX_\alpha^j EX_\gamma^h \\ &= (1/q - 1) EX_\alpha^j EX_\gamma^h \end{aligned} \tag{4}$$

Again, taking into account the edges of the sequence, equation 4 holds for

$$\begin{aligned} j, h &= k, k+1, \ldots, n \\ \alpha &= 1, 2, \ldots, n-j+1 \\ \gamma &\in \{\gamma : (\gamma \geq 1 \text{ and } \gamma = \alpha - h - 1) \text{ or } (\gamma = \alpha + j + 1 \text{ and } \gamma \leq n - h + 1)\}. \end{aligned}$$

Now, if we express the variance of $R_{n,k}$ as the sum of the variances of the individual terms plus the covariances of the reinforced terms plus the covariances of the mutually exclusive terms, we have:

$$\begin{aligned} Var(R_{n,k}) = & \sum_{j=k}^{n} \sum_{\alpha=1}^{n-j+1} j^2 Var(X_\alpha^j) \quad \text{individual variances} \\ & + 2 \sum_{j=k}^{n} \sum_{h=j}^{n} \sum_{\alpha=1}^{n-j+1} \sum_{\{\gamma\}} hj(1/q - 1) EX_\alpha^j EX_\gamma^h \quad \text{reinforced pairs} \\ & - 2 \sum_{j=k}^{n} \sum_{h=j}^{n} \sum_{\alpha=1}^{n-j+1} \sum_{\{\beta\}} hj EX_\alpha^j EX_\beta^h \quad \text{excluded pairs} \end{aligned} \tag{5}$$

where α, β are used for mutually exclusive pairs and α, γ are used for reinforcing pairs.

$$\{\beta\} = \{\beta : \beta \geq 1 \text{ and } \alpha - h \leq \beta \leq \alpha + j \text{ and } \beta \leq n - h + 1 \text{ and } (\alpha, j) \neq (\beta, h)\}$$

$$\{\gamma\} \quad = \quad \{\gamma : (\gamma \geq 1 \text{ and } \gamma = \alpha - h - 1) \text{ or } (\gamma = \alpha + j + 1 \text{ and } \gamma \leq n - h + 1)\}$$

In order to avoid counting the same covariance twice, if $h = j$ then $(\beta > \alpha)$ and $(\gamma > \alpha)$.

3 Quadratic Time Algorithm

By equations 1 and 5, we can compute the mean and variance of $R_{n,k}$. The mean can be easily calculated in constant time. For the variance it is straightforward that the computation can be done in $O(n^3)$ time because almost all EX_β^h have same value for given j, h, α and it is just a matter of finding sums of a constant. Below, we show that the variance can be calculated in $O(n^2)$ time. We show how to handle the covariances of the mutually excluded patterns. The analysis dealing with the reinforced patterns is similar and much simpler. Also, since it is easy to compute the covariances between the variables when $\alpha = 1$ and $\alpha = n - j + 1$, we ignore these two extreme cases in our analysis.

We focus on the covariances between variables denoting the mutually excluded patterns within the range of α from 2 to $n - j$. We first notice that for fixed j and h, many positions α yield the same value. In fact, we can partition the (j, h) pairs into three disjoint sets for which the α's give consistent values. These sets will lead naturally to the linear and constant time algorithms in section 4. Suppose we **fix** $j (k \leq j \leq n - 2)$, then we have three cases which partition the values of $h (j + 1 \leq h \leq n - 1)$. The relevance of this partition will be explained in section 4. The cases in increasing order of h are:

1. $n - j - h \geq h + 2$
2. $n - j - h < h + 2$ and $n - j \geq h + 2$
3. $n - j < h + 2$ and $n - j \geq 2$

We consider each case in turn.

Case 1: $n - h - j \geq h + 2$.

In this case, we partition the range of α into the three intervals (figure 1):

1. $2 \leq \alpha_1 < h + 2$
2. $h + 2 \leq \alpha_2 \leq n - j - h$
3. $n - j - h < \alpha_3 \leq n - j$.

For $\alpha \in \alpha_1$, the number of patterns excluded when $X_\alpha^j = 1$ shrinks by one each time α moves one position to the left from $h + 1$ to 2. When $\alpha = h + 1$, there are $j + h$ patterns of the form TH^hT and one pattern of the form H^hT. Recalling that the covariance for a mutually excluded pair is negative, we have the following covariances for $\alpha \in \alpha_1$:

$$
\begin{array}{ll}
\sum Cov(X_\alpha^j, X_\beta^h) = -((j+h)+1/q)EX^jEX^h & \alpha = h+1 \\
\sum Cov(X_\alpha^j, X_\beta^h) = -((j+h-1)+1/q)EX^jEX^h & \alpha = h \\
\vdots & \vdots \\
\sum Cov(X_\alpha^j, X_\beta^h) = -((j+1)+1/q)EX^jEX^h & \alpha = 2
\end{array}
$$

where, $EX^j = q^2p^j$, and $EX^h = q^2p^h$. If we sum all the above equations, we have the following total covariance in the interval α_1:

$$
\begin{aligned}
\sum_{\alpha \in \alpha_1} Cov(X_\alpha^j, X_\beta^h) &= -\left[h(j+1/q)EX^jEX^h + \frac{(h+1)h}{2}EX^jEX^h\right] \\
&= -h\left((j+1/q) + \frac{h+1}{2}\right)EX^jEX^h \qquad (6)
\end{aligned}
$$

By symmetry, the total covariance for interval α_3 is the same.

For $\alpha \in \alpha_2$, when α varies from $h+2$ to $n-j-h$, each position α has the same number of excluded patterns and yields the covariance $-(j+h+1)EX^jEX^h$. Therefore, since the size of α_2 is $(n-j-2h-1)$, we have the following total covariance:

$$
\sum_{\alpha \in \alpha_2} Cov(X_\alpha^j, X_\beta^h) = -(n-j-2h-1)(j+h+1)EX^jEX^h. \qquad (7)
$$

Case 2: $n-h-j < h+2$ **and** $n-j \geq h+2$.

In this case, we also separate the range of α into three intervals (figure 2):

1. $2 \leq \alpha_1 \leq n-h-j$
2. $n-h-j < \alpha_2 < h+2$
3. $h+2 \leq \alpha_3 \leq n-j$

In the intervals α_1 and α_3, the analysis is the same as in Case 1, except that when $\alpha = n-j-h$, there are $n-h-1$ patterns of the form TH^hT.

We have the following covariances for $\alpha \in \alpha_1$:

$$
\begin{array}{ll}
\sum Cov(X_\alpha^j, X_\beta^h) = -((n-h-1)+1/q)EX^jEX^h & \alpha = n-j-h \\
\sum Cov(X_\alpha^j, X_\beta^h) = -((n-h-2)+1/q)EX^jEX^h & \alpha = n-j-h-1 \\
\vdots & \vdots \\
\sum Cov(X_\alpha^j, X_\beta^h) = -((n-h-w)+1/q)EX^jEX^h & \alpha = 2
\end{array}
$$

where $w = n-h-j-1$. The total covariance in the interval α_1 is:

$$
\sum_{\alpha \in \alpha_1} Cov(X_\alpha^j, X_\beta^h) = -w\left((n-h+1/q) - \frac{w+1}{2}\right)EX^jEX^h. \qquad (8)
$$

In the interval α_2, the number of excluded patterns is constant with $(n-h-1)$ of the form TH^hT and one each of TH^h and H^hT. The size of α_2 is $(2h+j-n+1)$ and the covariance is:

$$\sum_{\alpha \in \alpha_2} Cov(X_\alpha^j, X_\beta^h) = -(2h+j-n+1)((n-h-1)+2/q)EX^jEX^h. \tag{9}$$

Case 3: $n-j < h+2$ **and** $n-j \geq 2$.

This case (figure 3) is similar to interval α_2 of Case 2. Every position α has the same number of excluded patterns. Since there are $n-j-1$ positions, the covariance is:

$$\sum_{\{\alpha\}} Cov(X_\alpha^j, X_\beta^h) = -(n-j-1)\left((n-h-1)+2/q\right)EX^jEX^h. \tag{10}$$

Using formulas $6-10$, we can compute the total variance of $R_{n,k}$ using just the variables j and h, that is, in $O(n^2)$ time.

4 Linear and Constant Time Algorithms

In order to get a linear time algorithm, we need to eliminate the dependence of the formulas on the variable h. We can see how to do this by **fixing** j and examining the change in total variance as we increase from h to $h+1$. If we can determine where the change is consistent, we will be done. Figure 4 illustrates the three cases for the quadratic algorithm. Using it, we can explain the ideas behind the linear time algorithm.

In figure 4, the solid black line is a function $f_j(\alpha)$ which is the number of excluded patterns for position α, when j is fixed. At the smallest h (Case 1), the function is a staircase on either end of a long platform. The minimum of the staircase is $j+2$, and the maximum is $j+h+1$. As h increases, the staircase gets longer because the maximum increases. Simultaneously, the platform shrinks. Eventually, the platform reaches its minimum size (either 1 or 2). This occurs just before the maximum number of patterns excluded (as determined by j) exceeds the maximum number of h patterns (of type $TH\ldots HT$) that fit in n. That is, when $j+h+1 = n-h-1$ or $j+h+1 = n-h-2$. Let

$$h_1 = \lfloor (n-j-2)/2 \rfloor$$

denote this critical value. When $h = h_1 + 1$ (Case 2), the staircases reach their maximum size. Thereafter, as h continues to increase, the maximum number of h patterns (of all types) that fit in n falls, pushing down the staircases, until all that is left is a single step. This occurs just before the minimum number of excluded patterns (as determined by j) exceeds the maximum number of h patterns. That is, when $n-h = j+2$. Let

$$h_2 = (n-j-2)$$

denote this second critical value. As h increases still further (Case 3), h, not j, determines the number of excluded patterns. Only the platform itself remains and it is pushed down until finally there is only one pattern left to be excluded.

Within each α_i region in each Case, the covariances change predictably with increasing h. Therefore (taking symmetry into account) we need five formulas for the total covariance due to the mutually exclusive variables. We show as one example the formulas for Case 1.

Based on the analysis for the quadratic algorithm, we replace, in formula 5, the term for the excluded pairs by:

$$-2\sum_{j=k}^{n}\sum_{h=j+1}^{h_1} hj \cdot Covar(h,j)$$

where $Covar(h,j)$ is one of formulas $6-10$ which are only dependent on the variables j and h. Now, we want to replace the inner sum so that we have

$$-2\sum_{j=k}^{n} j \cdot Covar(j)$$

where

$$Covar(j) = \sum_{h=j+1}^{h_1} h \cdot Covar(h,j)$$

is one of a set of formulas which are only dependent on the variable j. We eliminate h by setting $h = j + i (i = 1, 2 \cdots)$. Since h ranges between $j+1$ and h_1, the range of i is just

$$1 \le i \le (n - 3j - 2)/2.$$

Since $EX^j = q^2p^j$ and $EX^h = E(X^{j+i}) = q^2p^{j+i}$, we set $EX^h = p^iEX^j$.

Recall figure 1 for Case 1. For interval α_1 we have

$$\begin{aligned} Covar(j) &= -\sum_{h=j+1}^{h_1} h\left(h(j+1/q) + \frac{h+1}{2}\right) EX^jEX^h \\ &= -\sum_{i=1}^{(n-3j-2)/2} (j+i)\left((j+i)(j+1/q) + \frac{j+i+1}{2}\right) p^i(EX^j)^2. \end{aligned}$$

If we sum over all i, the covariance for the α_1 region is:

$$\begin{aligned} Covar(j) = -[&(((3j+1)\cdot j \cdot w1 + (4j+1)\cdot w2 + w3)/2)(EX^j)^2 \cdot j + \\ &(((3j+1)\cdot j \cdot w2 + (4j+1)\cdot w3 + w4)/2)(EX^j)^2 + \\ &((j \cdot w1 + w2)/q)(EXj)^2 \cdot j + ((j \cdot w2 + w3)/q)(EX^j)^2], \end{aligned}$$

where $w1 = \sum_{m=1}^{t} p^m$, $w2 = \sum_{m=1}^{t} mp^m$, $w3 = \sum_{m=1}^{t} m^2p^m$, $w4 = \sum_{m=1}^{t} m^3p^m$, and $t = (n - 3j - 2)/2$. We can derive the explicit closed form for each w. The formula for interval α_3 is the same.

For interval α_2, we have:

$$\begin{aligned} Covar(j) &= -\sum_{h=j+1}^{h_1} h(n-j-2h-1)(j+1+h)EX^jEX^h \\ &= -\sum_{i=1}^{(n-3j-2)/2} (j+i)(n-3j-2i-1)(2j+1+i)p^i(EX^j)^2. \end{aligned}$$

Again, summing over all i, we get:

$$Covar(j) = -[((n-3j-1)(2j+1)\cdot w1 + (n-7j-3)\cdot w2 - 2\cdot w3)(EX^j)^2\cdot j + \\ ((n-3j-1)(2j+1)\cdot w2 + (n-7j-3)\cdot w3 - 2\cdot w4)(EX^j)^2]$$

where $w1$, $w2$, $w3$, and $w4$ are same as above.

For the constant time algorithm, we partition the j into three intervals. Those j in the first interval exhibit each of Cases 1, 2 and 3. Those in the second interval exhibit each of Cases 2 and 3 and those in the third interval exhibit only Case 3. Now, *the partitions between these intervals are determined by the critical h values h_1 and h_2 and the minimum value h_{min} which is $j+1$.* All three of these values are determined by first fixing j. Note though, that at some minimum $j = j_1 + 1$, there will be no $h_1 \geq h_{min}$. Then, j_1 is the last j in the first interval. Similarly, at some minimum $j = j_2 + 1$ there will be no $h_2 \geq h_{min}$. Then, j_2 is the last j in the second interval. We determine that

$$j_1 = \lfloor (n-4)/3 \rfloor \text{ and } j_2 = \lfloor (n-3)/2 \rfloor.$$

Again, for each of the α intervals in each of the Cases within each subdivision of the j values, the covariances change predictably for increasing j. Carrying forward the preceding example, we next determine the covariance for the j in the first j interval,

$$-2\sum_{j=k}^{j_1} j\cdot Covar(j),$$

deriving formulas for the α_1 and α_2 regions in a manner similar to that illustrated above. This last step for Case 1 yields 2 formulas which are each longer than one page. For the mutually excluded variables (with α from 2 to $n-j$), there are 5 such formulas.

For our application, instead of programming such complex formulas, we stopped at the linear time algorithm which was fast enough for our purposes. For example, running on a Silicon Graphics O2 R10000, the linear time algorithm computes $R_{500,5}$ for $p = 0.8$ in .033 seconds. The linear time algorithm can be obtained by sending email to benson@ecology.biomath.mssm.edu.

5 Sequences generated by a Markov process

Using the techniques described in the previous sections, we are able to produce a $O(n)$ algorithm for computing the mean and variance of $R_{n,k}$ for sequences generated by a Markov process. We assume the Markov chain is stationary and of order 1. (There is no loss of generality because we can write an order m chain on $\Sigma = \{H, T\}$ as an order 1 chain on Σ^m.)

We use the following notation. $X, Y \in \{H, T\}$. Let $\pi(X)$ be the stationary probability of X. Let P be the transition matrix and $P(X|Y)$ be the transition probability from Y to X. Let $P^{(d)}_{X|Y}$ be the d-step transition probability from Y to X. $P^{(d)}_{X|Y}$ is just an entry in P^d, the dth power of P.

In order to calculate the variance, we must calculate the covariance of $O(n^4)$ pairs of indicator variables. The mutually excluded and reinforced variable pairs are handled in a manner similar to that described in the previous sections. Here, we show how to handle the variables representing patterns that do not overlap. Unlike the situation with iid Bernoulli trials, the covariance of these variables is *not* zero. To simplify the description, we assume that every variable represents a pattern that begins and ends with a T.

First, for the expectation of a single variable we have:

$$E(X_\alpha^j) = \pi(T)P(H|T)P(H|H)^{j-1}P(T|H).$$

For the joint expectation we assume that we have two variables, X_α^j and X_β^h where X_α^j represents the first pattern to occur in the sequence and X_β^h represents the following pattern. We let $\beta = \alpha + j + 1 + d$, that is, d is the difference between the indices of the last T of the pattern represented by X_α^j and the first T of the pattern represented by X_β^h.

$$\begin{aligned} E(X_\alpha^j X_\beta^h) &= \pi(T)P(H|T)P(H|H)^{j-1}P(T|H)P_{T|T}^{(d)}P(H|T)P(H|H)^{h-1}P(T|H) \\ &= Qp^j p^h r(d), \end{aligned}$$

where

$$\begin{aligned} \alpha &\in [2, \ldots, n-j-h-2] \\ \beta &\in [\alpha+j+2, \ldots, n-h] \\ Q &= \pi(T)P(H|T)^2P(T|H)^2P(H|H)^2 \\ p &= P(H|H) \\ r(d) &= P_{T|T}^{(d)} \end{aligned}$$

From equation 2 we can write the part of the variance due to the pairs of non-overlapping patterns as:

$$2\sum_{j=k}^{n-k-4}\sum_{h=k}^{n-j-4}\sum_{\alpha=2}^{n-j-h-2}\sum_{\beta=\alpha+j+2}^{n-h} jhCov(X_\alpha^j, X_\beta^h).$$

We remove the inner summation by replacing β with $\alpha + j + 1 + d$ and let d range from 1 to $n - \alpha - j - h - 1 = n(\alpha, j, h)$. Then

$$\sum_{d=1}^{n(\alpha,j,h)} jhE(X_\alpha^j X_{\alpha+j+1+d}^h) = Qjhp^j p^h \sum_{d=1}^{n(\alpha,j,h)} r(d).$$

Note that all the $r(d) = P_{T|T}^{(d)}$ can be calculated together in $O(n)$ time and once we have the $r(d)$, all the sums, $\sum r(d)$, can also be calculated in $O(n)$ time. Next, we group together those terms that have the same sum, $\sum r(d)$, as a factor. Omitting the constant term Q and setting $n^* = n(2,k,k) = n - 2k - 3$ yields:

$$\begin{aligned}
&\sum_{d=1}^{n^*} r(d) \quad \left(kp^k \sum_{t=k}^{k} tp^t \right) + \\
&\sum_{d=1}^{n^*-1} r(d) \quad \left(kp^k \sum_{t=k}^{k+1} tp^t + (k+1)p^{k+1} \sum_{t=k}^{k} tp^t \right) + \\
&\sum_{d=1}^{n^*-2} r(d) \quad \left(kp^k \sum_{t=k}^{k+2} tp^t + (k+1)p^{k+1} \sum_{t=k}^{k+1} tp^t + (k+2)p^{k+2} \sum_{t=k}^{k} tp^t \right) + \\
&\qquad \vdots \\
&\sum_{d=1}^{1} r(d) \quad \left(kp^k \sum_{t=k}^{k+n^*-1} tp^t + (k+1)p^{k+1} \sum_{t=k}^{k+n^*-2} tp^t + \ldots + (k+n^*-1)p^{k+n^*-1} \sum_{t=k}^{k} tp^t \right).
\end{aligned}$$

Now, looking at just the part of each line in parenthesis, we notice that the difference between lines l and $l+1$ (Δ_l) is just:

$$\begin{aligned}
\Delta_l &= kp^k(k+l)p^{k+l} + (k+1)p^{k+1}(k+l-1)p^{k+l-1} + \ldots + (k+l)p^{k+l}kp^k \\
&= p^{2k+l} \sum_{t=0}^{l} (k+t)(k+l-t) \\
&= p^{2k+l} \left(\frac{(1+l)(6k^2 - l + 6kl + l^2)}{6} \right)
\end{aligned}$$

Each Δ term can be calculated in constant time and thus the entire calculation of the variance takes $O(n)$ time.

6 Acknowledgement

The authors would like to thank Sophie Schbath for pointing out that our technique can also be applied to sequences generated by a Markov process.

References

[1] J. Armour, T. Anttinen, C. May, E. Vega, A. Sajantila, J. Kidd, K. Kidd, J. Bertranpetit, S. Pääbo, and A. Jeffreys. Minisatellite diversity supports a recent African origin for modern humans. *Nature Genetics*, 13:154–160, 1996.

[2] R. Arratia, L. Goldstein, and L. Gordon. Poisson approximation and the Chen-Stein method. *Statistical Science*, 5:403–434, 1990.

[3] G. Benson. A space efficient algorithm for finding the best non-overlapping alignment score. In M. Crochemore and D. Gusfield, editors, *Proc. 5th annual Symp. on Combinatorial Pattern Matching, Lecture Notes in Computer Science*, volume 807, pages 1–14. Springer-Verlag, 1994.

[4] G. Benson. An algorithm for finding tandem repeats of unspecified pattern size. Manuscript, 1996.

[5] G. Benson and M. Waterman. A method for fast database search for all k-nucleotide repeats. *Nucleic Acids Research*, 22:4828–4836, 1994.

[6] V. Campuzano, L. Montermini, M.D. Molto, L. Pianese, and M. Cossee. Friedreich's ataxia: Autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science*, 271:1423–1427, 1996.

[7] O. Chrysaphinou, S. Papastavridis, and T. Tsapels. On the number of overlapping success runs in a sequence of independent Bernoulli trials. *Application of Fibonacci Numbers*, 5:103–112, 1993.

[8] A. Edwards, H. Hammond, L. Jin, C. Caskey, and R. Chakraborty. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics*, 12:241–253, 1992.

[9] J. Fu and V. Koutras. Distribution theory of runs: A Markov chain approach. *Journal of the American Statistical Association*, 89:1050–1058, 1994.

[10] Y.-H. Fu, A. Pizzuti, J. Fenwick, R.G.Jr.and King, S. Rajnarayan, P.W. Dunne, J. Dubel, G.A. Nasser, T. Ashizawa, P. DeJong, B. Wieringa, R. Korneluk, M.B. Perryman, H.F. Epstein, and C.T. Caskey. An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science*, 255:1256–1258, 1992.

[11] A. Godbole. Specific formulas for some success run distributions. *Statistics and Probability Letters*, 10:119–124, 1990.

[12] L. Goldstein and M. Waterman. Poisson, compound Poisson and process approximations for testing statistical significance in sequence comparisons. *Bulletin of Mathematical Biology*, 54:785–812, 1992.

[13] K. Hirano. Some properties of the distributions of order k. In A. Philippou, G. Bergum, and A. Horadam, editors, *Fibonacci Numbers and Their Applications*, pages 43–53. Dordrecht:Reidel, 1986.

[14] K. Hirano and S. Aki. On the number of occurrences of success runs of length k in a two-state Markov chain. *Statistica Sinica*, 3:313–319, 1992.

[15] K. Hirano, S. Aki, N. Kashiwagi, and H. Kuboki. On Ling's binomial and negative binomial distributions of order k. *Statistics and Probability Letters*, 11:503–509, 1991.

[16] Huntington's disease collaborative research group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72:971–983, 1993.

[17] S. Kannan and E. Myers. An algorithm for locating regions of maximum alignment score. In *Proc. 4th Annual Symp. on Combinatorial Pattern Matching, Lecture Notes in Computer Science*, volume 648, pages 74–86. Springer-Verlag, 1993.

[18] G. Landau and J. Schmidt. An algorithm for approximate tandem repeats. In *Proc. 4th Annual Symp. on Combinatorial Pattern Matching, Lecture Notes in Computer Science*, volume 648, pages 120–133. Springer-Verlag, 1993.

[19] K. Ling. On binomial distributions of order k. *Statistics and Probability Letters*, 6:247–250, 1988.

[20] D. Lipman and W. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227:1435–1441, 1985.

[21] W. Messier, S-H. Li, and C-B. Stewart. The birth of mircosatellites. *Nature*, 381:483, 1996.

[22] A. Philippou and F. Makri. Longest success runs and Fibonacci-type polynomials. *Fibonacci Quarterly*, 23:338–346, 1985.

[23] S. Schbath. Compound Poisson approximation of word counts in DNA sequences. *ESAIM*, 1:1–16.(http://www.emath.fr/ps/), 1995.

[24] J.P. Schmidt. All highest scoring paths in weighted grid graphs and its application to finding all approximate repeats in strings. In *Third Israel Symposium on Theory of Computing and Systems*, pages 67–77. IEEE Computer Society Press, 1995.

[25] S.A. Tishkoff, E. Dietzsch, W. Speed, A.J. Pakstis, and J.R. Kidd. Global patterns of linkage disequilibrium at the cd4 locus and modern human origins. *Science*, 271:1380–1387, 1996.

[26] A. Verkerk, M. Pieretti, J. Sutcliffe, Y. Fu, D. Kuhl, A. Pizzuti, O. Reiner, S. Richards, M. Victoria, F. Zhang, B. Eussen, G. van Ommen, A. Blonden, G. Riggins, J. Chastain, C. Kunst, H. Galjaard, C. Caskey, D. Nelson, B. Oostra, and S. Warren. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, 65:905–914, 1991.

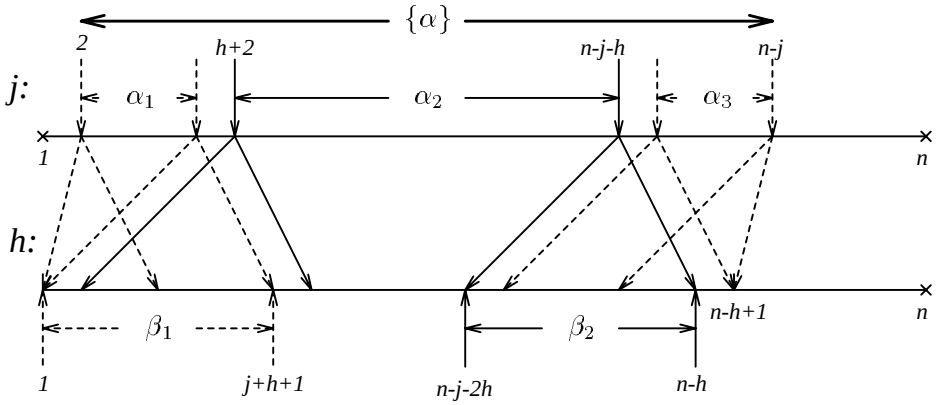


Figure 1:

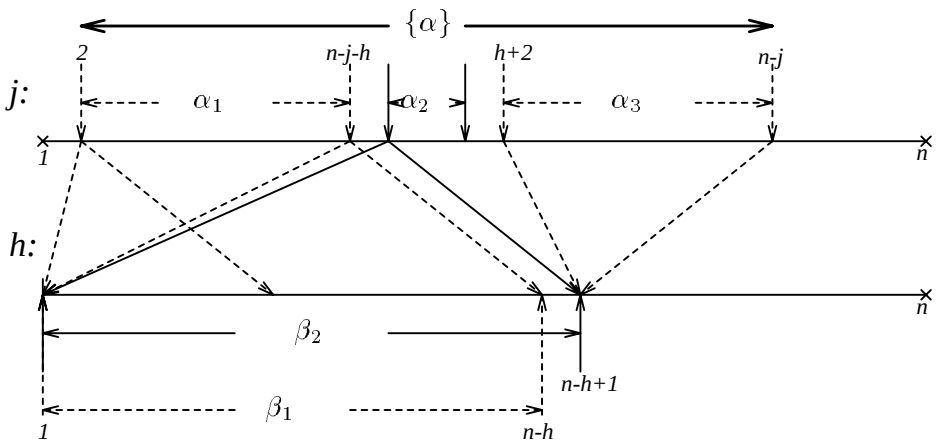


Figure 2:

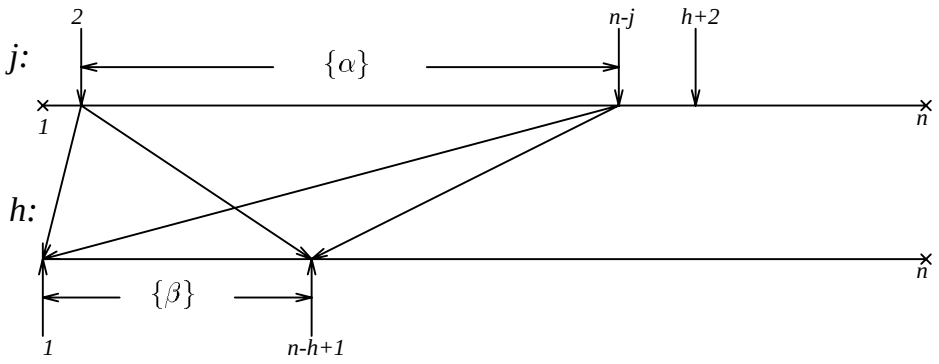
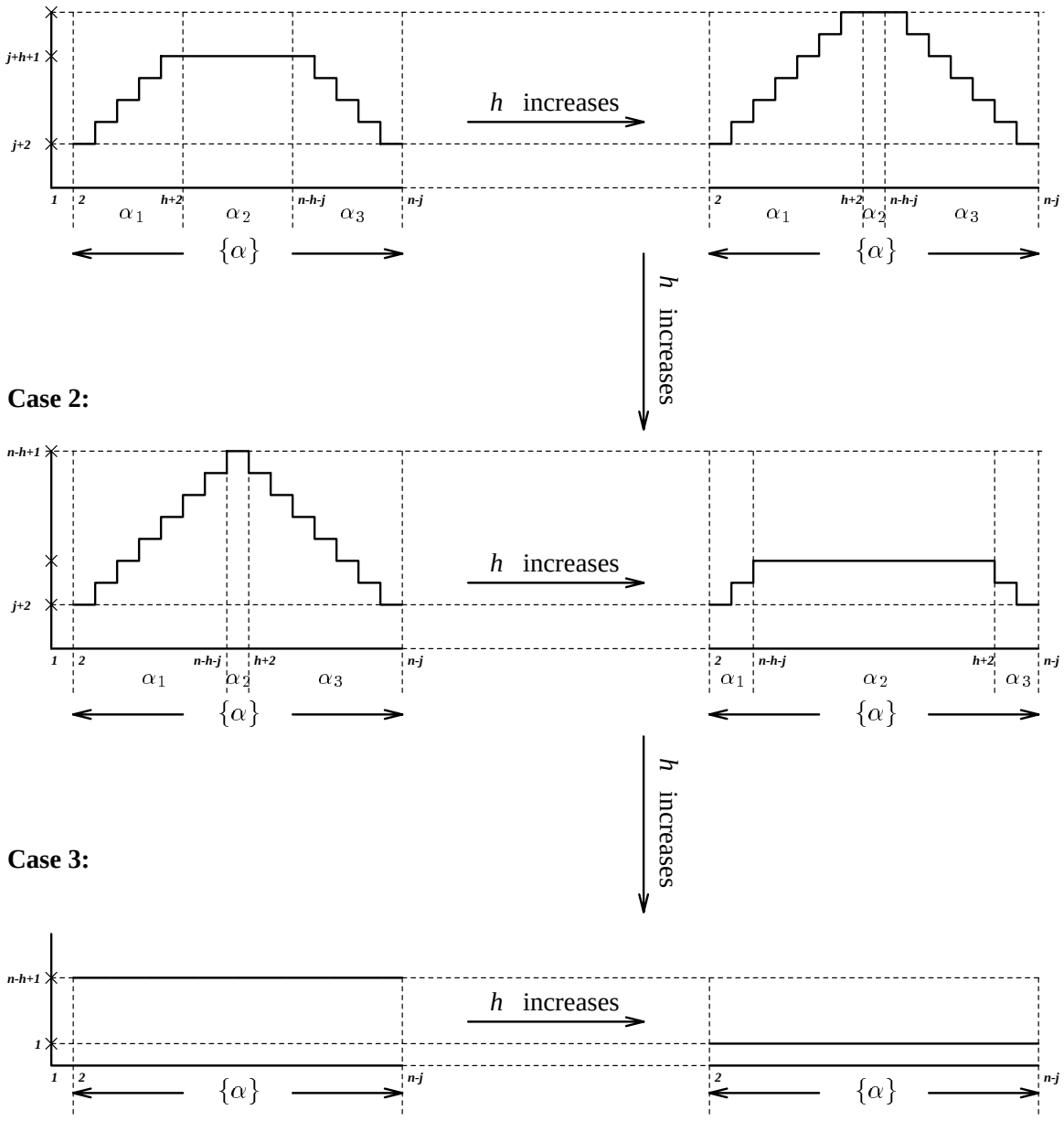


Figure 3:

Case 1:



Case 2:

Case 3:

Figure 4: