Computer Sciences Department

Statistical Analysis of DNA Sequences Using Overlapping Windows

Amy Hauth
Murray K. Clayton

Technical Report #1474

May 2000

UNIVERSITY OF
WISCONSIN
MADISON

Statistical Analysis of DNA Sequences using Overlapping Windows

Amy Hauth[1], Department of Computer Sciences
Murray K. Clayton, Departments of Statistics and Plant Pathology
University of Wisconsin-Madison
1210 W. Dayton Ave.
Madison, WI 53706

May2000

Keywords: Variance, Overlapping k-Length Windows, Count Occurrence Statistics

[1] To whom correspondence should be addressed.

Abstract

Motivation: Our analysis of DNA sequences uses a k-length, sliding window and considers all overlapping windows along the sequence. The k consecutive nucleotides in a window are called a word or k-word. Statistical analysis of this collection of words often assumes independence between words. Since words can overlap, strict independence is not a valid assumption. We derive a statistic to incorporate both the independent and dependent components of overlapping, k-length words.
Results: The expected number of occurrences for a k-word in an N-length sequence is easily calculated given the probabilities of the nucleotides within the word. However, the variance is not straightforward since overlapping occurrences are not independent. We present a derivation of the variance when sequence analysis uses overlapping, k-length windows. The variance can be determined for a word in the entire sequence or at a single position in the sequence. Our analysis assumes that each nucleotide is independent. It does not assume a specific probability of occurrence for each nucleotide.
Contact: Hauth: *kryder@cs.wisc.edu*; Clayton: *clayton@stat.wisc.edu*
Keywords: Variance, Overlapping k-Length Windows, Count Occurrence Statistics

Introduction

Count occurrence statistics typically use sliding overlapping windows to analyze a sequence. Mononucleotide through hexanucleotide occurrence patterns frequently appear in the literature [Reddy and Pandit 1995, Adams *et al.* 1987, Arnold *et al.* 1988, Rogerson 1991, Cuticchia *et al.* 1992, VanLith and VanZutphen 1996]. This includes codon usage patterns (or trinucleotide usage patterns) in a variety of species [Arques and Michel 1996]. Occurrence distributions have been determined for species [Jarret *et al.* 1997, Smutzer and Chamberlin 1994, Primmer *et al.* 1997, Rogerson 1989, VanLith and VanZutphen 1996], non-coding [Rogerson 1989, VanLith and VanZutphen 1996] and coding [Smutzer and Chamberlin 1994, VanLith and VanZutphen 1996, Rogerson 1989, Reddy and Pandit 1995] regions, and various other DNA features [Adams *et al.* 1987, Reddy and Pandit 1995].

The count occurrence distribution for a sequence S of length N is determined using a window of k adjacent nucleotides, a k-word. All overlapping k-words in the sequence are considered, and, for our purposes, the alphabet consists of four nucleotides, A, C, G and T.

To examine the count occurrence distribution, first consider the occurrence of a single word. We assume that each nucleotide has a probability of occurrence p_x (where x is the nucleotide A, C, G or T). To determine p_x in this paper, we use the nucleotide occurrence distribution within the sequence: the number of sequence occurrences of x divided by the sequence length, N. Others have used an equal distribution of probability for all possible nucleotides (i.e. $p_A = p_C = p_G = p_T = 0.25$) as well as the nucleotide occurrence distribution for a particular species.

In our probability model we also assume that a word of k adjacent nucleotides, $x_1x_2 \dots x_k$, has a probability of occurrence, $P(x_1x_2 \dots x_k)$, equal to $p_{x_1} \times p_{x_2} \times \dots \times p_{x_k}$. This is equivalent to assuming that the occurrence of a specific nucleotide in a given position is independent of the nucleotides in other positions. Other possible models exist, of course, including Markov and hidden Markov models [Durbin *et al.* 1998].

Our definitions allow words in a sequence to overlap. Thus, for example, the word *GGG* appears twice in the sequence *GGGG*. With this in mind, we want to calculate the mean and variance of the number of occurrences of a given word in the entire sequence. To calculate the mean, let Y_i be an indicator variable that is 1 if the word of interest appears at position i, and 0 otherwise. So, for example, the word *GGG* appear three times in the sequence *GGGAGGGG* and thus Y_i is 1 for $i = 1, 5$ and 6, and 0 otherwise. Let $\mathcal{N}$ denote the number of occurrences of the word of interest in the sequence.

It is clear that $\mathcal{N} = \sum_{i=1}^{n} Y_i$ where $n = (N - k + 1)$ is the total number of possible words of length k in a sequence of length N. It follows that the expected number of occurrences of the word $x_1x_2 \dots x_k$ is $E(\mathcal{N}) = E(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} E(Y_i) = \sum_{i=1}^{n} p_{x_1}p_{x_2} \cdots p_{x_k} = n(p_{x_1}p_{x_2} \cdots p_{x_k})$. Here we use the fact that $E(Y_i) = P(Y_i = 1)$.

In the above calculation, the Y_i are dependent random variables, a consequence of allowing for the possibility that words can overlap. While this does not affect the calculation of the expected value of $\mathcal{N}$, it does impact the calculation of the variance of $\mathcal{N}$. In the remainder of this paper we will show that the variance associated with the word $x_1x_2 \dots x_k$ is

$$\begin{aligned}&(N - k + 1)\,[P(x_1x_2 \dots x_k) \\ &\qquad\qquad -(2k - 1)P^2(x_1x_2 \dots x_k)] \\ &+ \quad 2\sum_{i=1}^{k-1}(N - k + 1 - i)P(x_1x_2 \dots x_{k+i})\end{aligned} \tag{1}$$

In addition, we derive an upper bound that is easily calculated.

The Variance Derivation

For the random quantity $\mathcal{N}$, we have $Var(\mathcal{N}) = E(\mathcal{N}^2) - E(\mathcal{N})^2$. Because we have already calculated $E(\mathcal{N})$, we now focus on $E(\mathcal{N}^2)$. From the preceding section we may write:

$$\begin{aligned}
&E(\mathcal{N}^2) && (2)\\
&= E\left((\sum_{i=1}^{n} Y_i)^2 \right)\\
&= E\left(\sum_{i=1}^{n} Y_i^2 + 2\sum_{i=1}^{n} \sum_{j=i+1}^{n} Y_iY_j \right)\\
&= E\left(\sum_{i=1}^{n} Y_i + 2\sum_{i=1}^{n} \sum_{j=i+1}^{n} Y_iY_j \right)\\
&= \sum_{i=1}^{n} E(Y_i) + 2\sum_{i=1}^{n} \sum_{j=i+1}^{n} E(Y_iY_j)
\end{aligned}$$

The second last equality follows because $Y_i^2 = Y_i$ (since Y_i is a binary indicator); the last equality follows since the expectation can be passed through the summation. We have already evaluated $\sum_{i=1}^{n} E(Y_i)$; we now focus on the double summation.

Note first that, for a word length k, if $j \geq i+k$ then Y_i and Y_j are independent: in other words, if the positions of interest differ by k or more, then the words do not overlap. Thus we may write:

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{j=i+1}^{n} E(Y_iY_j) && (3)\\
&= \sum_{i=1}^{n} \sum_{j=i+1}^{i+k-1} E(Y_iY_j) + \sum_{i=1}^{n} \sum_{j=i+k}^{n} E(Y_iY_j)\\
&= \sum_{i=1}^{n} \sum_{j=i+1}^{i+k-1} E(Y_iY_j) + \sum_{i=1}^{n} \sum_{j=i+k}^{n} E(Y_i)E(Y_j)
\end{aligned}$$

We will define the "dependent variance component" to be the first double summation in 3 and the "independent variance component" to be the second double summation.

The independent variance component can be handled by again using the fact that $E(Y_i) = P(Y_i = 1) = P(x_1x_2 \dots x_k)$. Thus, we are then left with calculating $\sum_{i=1}^{n} \sum_{j=i+1}^{i+k-1} E(Y_iY_j)$, the component that involves dependent, i.e. overlapping words. To outline our calculations, we will focus on the specific word consisting of 3 adjacent occurrences of the nucleotide G. To expand our notation based on Y_i, let S_i denote the specific nucleotide at location i, and let I_A be an indicator variable which is equal to 1 if A is true, and 0 otherwise. Thus, we may write $Y_i = I_{S_i=G,S_{i+1}=G,S_{i+2}=G}$. With this new notation, and with the specific word GGG,

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{j=i+1}^{i+k-1} E(Y_iY_j) && (4)\\
&= \sum_{i=1}^{N-2} \sum_{j=i+1}^{i+2} E(I_{S_i=G,S_{i+1}=G,S_{i+2}=G}\\
&\qquad \times I_{S_j=G,S_{j+1}=G,S_{j+2}=G})\\
&= \sum_{i=1}^{N-2} \big(E(I_{S_i=G,S_{i+1}=G,S_{i+2}=G}\\
&\qquad \times I_{S_{i+1}=G,S_{i+2}=G,S_{i+3}=G})\\
&\quad + E(I_{S_i=G,S_{i+1}=G,S_{i+2}=G} I_{S_{i+2}=G,S_{i+3}=G,S_{i+4}=G})\big)\\
&= \sum_{i=1}^{N-3} E(I_{S_i=G,S_{i+1}=G,S_{i+2}=G,S_{i+3}=G})\\
&\quad + \sum_{i=1}^{N-4} E(I_{S_i=G,S_{i+1}=G,S_{i+2}=G,S_{i+3}=G,S_{i+4}=G})\\
&= \sum_{i=1}^{N-3} P(GGGG) + \sum_{i=1}^{N-4} P(GGGGG)
\end{aligned}$$

The second to last equality follows from the boolean relation $I_AI_B = I_{A \cap B}$; the last equality follows because, again, $E(Y_i) = P(Y_i = 1) = P(x_ix_{i+1} \dots x_k)$. The above development shows explicitly how dependency results from word overlap. The word, GGG, overlaps at two positions in the word $GGGG$, and at one position in the word $GGGGG$. In general, words have at most $(k-1)$ overlap terms; each overlap term must match exactly in the region of overlap.

By assembling all of the above components and performing the requisite algebra, we have, for the word GGG,

$$\begin{aligned}
&Var(\mathcal{N}) = (N-2)\left[P(GGG) - 5P^2(GGG)\right] && (5)\\
&+2(N-3)P(GGGG) + 2(N-4)P(GGGGG)
\end{aligned}$$

The process and formulas used for GGG can be applied to any word. The variance for the generic word $x_1x_2 \dots x_k$, $Var(\mathcal{N})$, is

$$\begin{aligned}
&(N-k+1)\,[P(x_1x_2 \dots x_k) && (6)\\
&\qquad -(2k-1)P^2(x_1x_2 \dots x_k)]\\
&+ \quad 2\sum_{i=1}^{k-1}(N-k+1-i)P(x_1x_2 \dots x_{k+i})
\end{aligned}$$

Alternative Formulas

We now derive an upper bound on the quantity in 6. First, we focus on the second term, a summation of $(k-1)$

terms. We may write

$$2\sum_{i=1}^{k-1}(N-k+1-i)P(x_1x_2\ldots x_{k+i}) \qquad (7)$$
$$\leq \quad 2(k-1)(N-k)P(x_1x_2\ldots x_{k+1})$$

since $P(x_1x_2\ldots x_{k+i}) \leq P(x_1x_2\ldots x_{k+1})$ and $(N-k+1-i) \leq (N-k)$ for any $i \geq 1$.

Next, we write $P(x_1x_2\ldots x_{x+1}) = P(x_1x_2\ldots x_k) \times p_{x_{k+1}} \leq P(x_1x_2\ldots x_k) \times \max p_x$, and substitute $E(\mathcal{N})/(N-k+1)$ for $P(x_1x_2\ldots x_k)$.

After applying these results, the upper bound for variance becomes:

$$Var(\mathcal{N}) \leq E(\mathcal{N}) - \frac{E(\mathcal{N})}{(N-k+1)}[(2k-1)E(\mathcal{N}) \qquad (8)$$
$$-2(k-1)(N-k)\max p_x]$$

Finally, we note that some applications require the expected occurrence and variance of a word at a single (fixed) position in the sequence rather than for the entire sequence. For example, statistics for subsets of a sequence can use single position statistics. These can be calculated in a straightforward manner given the above results. For a word at a single position we use the standard statistical results:

$$E(\mathcal{N}/n) = E(\mathcal{N})/n \qquad (9)$$

$$Var(\mathcal{N}/n) = Var(\mathcal{N})/n^2 \qquad (10)$$

An Example: Analysis of a DNA sequence

A DNA sequence (see figure 1) which contains a minisatellite tandem repeat is analyzed. A $(GT)_n$ microsatellite occurs within the minisatellite pattern. Three words, GGGGG, TGTGT and TGGGG, are evaluated to illustrate the role of dependency within the variance equation. Alignment of the word to itself dictates when a contribution is made to the dependent component. GGGGG overlaps itself at every position. TGTGT overlaps itself at every other position. TGGGG never overlaps itself and makes no contribution to the dependent variance component. Table 1 details the statistics for GGGGG, TGTGT and TGGGG. As expected, the bound is best when the dependent component contributes to variance for every possible overlap. It is worst when there is no contribution from the dependent component.

The bound on variance (presented in the final column) is probably best interpreted by considering the associated standard deviation. GGGGG has an actual standard deviation of 1.99 and a bound of 2.75. For TGTGT, it is 1.68 bounded by 2.89. For TGGGG, it is 1.46 bounded by 2.79. For most applications, this bound is sufficient.

The sequence in figure 1 contains a $(GT)_n$ microsatellite which is composed of two 5 nucleotide words, GTGTG and TGTGT. The z-scores in Table 1 show that TGTGT occurs many standard deviations above the mean. This supports the notion that TGTGT occurs much more frequently than expected, a common observation with tandem repeats. The z-scores for similar words, GGGGG and TGGGG, which are not contained in a tandem repeat are within one standard deviation of the mean.

References

[Adams *et al.* 1987] Adams, R.L.P., Davis, T., Rinaldi, A., and Eason, R. 1987. CpG deficiency, dinucleotide distributions and nucleosome positioning. *Eur. J. Biochem.*, **165**(1), 107–116.

[Arnold *et al.* 1988] Arnold, J., Cuticchia, A.J., Newsome, D.A., Jennings, W.W. III, and Ivarie, R. 1988. Mononucleotide through hexanucleotide composition of the sense strand of yeast DNA: A Markov-chain analysis. *Nucleic Acids Res.*, **16**(14b), 7145–7158.

[Arques and Michel 1996] Arques, D.G., and Michel, C.J. 1996. A complementary circular code in the protein coding genes. *J. Theor. Biol.*, **182**(1), 45–58.

[Cuticchia *et al.* 1992] Cuticchia, A.J., Ivarie, R., and Arnold, J. 1992. The application of Markov chain analysis to oligonucleotide frequency prediction and physical mapping of *Drosophila melanogaster. Nucleic Acids Res.*, **20**(14), 3651–3657.

[Durbin *et al.* 1998] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge, U.K. : New York, New York: Cambridge University Press.

[Jarret *et al.* 1997] Jarret, R.L., Merrick, L.C., Holms, T., Evans, J., and Aradhya, M.K. 1997. Simple sequence repeats in watermelon (*Citrullus lanatus* (Thunb.) Matsum. and Nakai). *Genome*, **40**(4), 433–441.

[Kashi *et al.* 1990] Kashi, Y., Iraqi, F., Tikochinski, Y., Ruzinzki, B., Nave, A., Beckmann, J.S., Friedmann, A., Soller, M., and Gruenbaum, Y. 1990. (TG)$_n$ Uncovers a sex-specific hybridization pattern in cattle. *Genomics*, **7**, 31–36.

[Primmer *et al.* 1997] Primmer, C.R., Raudsepp, T., Chowdhary, B.P., Moller, A.P., and Ellegren, H. 1997. Low frequency of microsatellites in the avian genome. *Genome Res.*, **7**(5), 471–482.

[Reddy and Pandit 1995] Reddy, B.V.B., and Pandit, M.W. 1995. A statistical analytical approach to decipher information from biological sequences: Application to murine splice-site analysis and prediction. *J. Biomol. Struct. Dyn.*, **12**(4), 785–801.

[Rogerson 1989] Rogerson, A.C. 1989. The sequence asymmetry of the *Escherichia coli* chromosome appears to be independent of strand or function and may be evolutionarily conserved. *Nucleic Acids Res.*, **17**(14), 5547–5564.

[Rogerson 1991] Rogerson, A.C. 1991. There appear to be conserved constraints on the distribution of nucleotide sequences in cellular genomes. *J. Mol. Evol.*, **32**(1), 23–30.

[Smutzer and Chamberlin 1994] Smutzer, G., and Chamberlin, L.L. 1994. Dinucleotide frequencies and codon usage in jawless and cartilaginous fishes. *Mol. Mar. Biol. Biotech.*, **3**(2), 112–119.

[VanLith and VanZutphen 1996] VanLith, H.A., and VanZutphen, L.F.M. 1996. Characterization of rabbit DNA microsatellites extracted from the EMBL nucleotide sequence database. *Anim. Genet.*, **27**(6), 387–395.

```
  1  aagcttcaca  tcccgagaat  tccctcccag  cgctcgtggt  cccacagagg  gctctgctgg
 61  acctgcctcg  ggtcacatgg  caggtctggg  gaggacacac  ctctccccgg  cagagaaatg
121  gccagaagcc  aggtctgctc  cacacgtgcc  ttctcccaat  actctctaac  tttaaaaaaa
181  ctgccaaaga  aaaagcggta  cgtaataaca  agcgcacaga  tacgtaattt  ataatggctg
241  acacggttgg  cagggaaatg  tgttacgcag  gaattatgtt  tttatttatg  tgtgtcctgt
301  tttggagaca  gcataagtaa  tcatgggtgt  gtgtgtgtgt  gtgtgtgtgt  gttgcctgtc
361  tccagcgtaa  gtaatcatgt  gtgtgtgtgt  gtgtgtgtgt  tgcctgtctc  cagcgtaagt
421  aatcgtgtgt  gtgtgtgtgt  gtgtgtgtgt  gtgtgtgtgt  gttgcctgtc  tccagagtaa
481  gtaatcatgg  gtgtgtgtgt  gtgtgtgttg  cctgtctcca  gcataagtaa  tcatgggtgt
541  gtgtgtgtgt  gtgtgttgcc  tgtctccagc  ataagtaatc  atgggtgtgt  gtgtgtgtgt
601  gtgtgtgttg  cctgtctcca  gcataagtaa  tcatgggggg  gtgtgtgtgt  gtgtgtgtgt
661  gtgtgtgtgt  gtgtgtgtgt  gtgtgtgtgt  tgcctgtctc  cagggacttt  tgtacagaga
721  agctt
```

Figure 1: A Bos taurus DNA sequence (GenBank LOCUS:BOVTGN) [Kashi *et al.* 1990] is composed of 134 A, 131 C, 226 G, and 234 T nucleotides and contains a minisatellite at position 311 to position 703. The minisatellite contains a $(GT)_n$ microsatellite within its pattern.

word	variance	independent variance	dependent variance component				z-score	variance bound	z-score based
			overlap	overlap	overlap	overlap			
	(see eq. 6)	component	of 4	of 3	of 2	of 1		(see eq. 8)	on bound
GGGGG	3.97	2.07	1.32	0.411	0.128	0.0398	0.441	7.54	0.320
TGTGT	2.81	2.29	0	0.473	0	0.0474	51.7	8.36	30.0
TGGGG	2.14	2.14	0	0	0	0	-0.135	7.80	-0.0706

Table 1: Variance Statistics for the words GGGGG, TGTGT and TGGGG within the sequence shown in Figure 1. The nucleotide content of the sequence is used to calculate the nucleotide probabilities.