Determining the distribution of word matches between Markovian sequences

Conrad Burden, Sylvain Forêt, *Paul Leopardi

Mathematical Sciences Institute, Australian National University.

For presentation at CTAC, QUT, 2012.

Based on a presentation given by Conrad Burden at COMPSTAT Cyprus.

25 September 2012
Acknowledgements

Sue Wilson (Australian National University, University of New South Wales).

Australian Research Council grant DP120101422.
Definition of $D_2$

Given two sequences from a finite alphabet

\[ A := (A_1, A_2, \ldots, A_m) \text{ and } B := (B_1, B_2, \ldots, B_n), \]

$D_2$ is the number of matches of words (including overlaps) of prespecified length \( k \) between two given sequences.
Definition of $D_2$

Given two sequences from a finite alphabet

$$A := (A_1, A_2, \ldots, A_m) \text{ and } B := (B_1, B_2, \ldots, B_n),$$

$D_2$ is the number of matches of words (including overlaps) of prespecified length $k$ between two given sequences.

Example: consider these two sequences and $k = 7 \ldots$

<table>
<thead>
<tr>
<th>A:</th>
<th>ATGCTTTTGCTAGCGCTATGCTTTTCGCAAACCTCAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>B:</td>
<td>ATGCTTTTTAAAAACCGAGCTGGTCAGCGCTAAGCGCT</td>
</tr>
</tbody>
</table>
Definition of $D_2$

Given two sequences from a finite alphabet

$$A := (A_1, A_2, \ldots, A_m) \text{ and } B := (B_1, B_2, \ldots, B_n),$$

$D_2$ is the number of matches of words (including overlaps) of prespecified length $k$ between two given sequences.

Example: consider these two sequences and $k = 7$...

A: \[\text{ATGCTTTGCTAGCGCTATGCTTTCGCAAAACTCAT}\]

B: \[\text{ATGCTTTTAAAAACCGAGCTGGTCAGCGCTAAGCGCT}\]
Definition of $D_2$

Given two sequences from a finite alphabet

$$A := (A_1, A_2, \ldots, A_m) \text{ and } B := (B_1, B_2, \ldots, B_n),$$

$D_2$ is the number of matches of words (including overlaps) of prespecified length $k$ between two given sequences.

Example: consider these two sequences and $k = 7$...

In this example, for $k = 7$, $D_2 = 3$. 
Our previous results

(with Sue Wilson, Ruth Kantorovitz, and Junmei Jing)

We have exact formulas for $\mathbf{E}(D_2)$ (simple) and $\text{Var}(D_2)$ (complicated), and accurate approximations for the distribution of $D_2$ for the case of sequences composed of i.i.d. letters; i.e. each letter drawn independently from the same distribution.
Periodic boundary conditions

To simplify the calculations (avoiding ‘edge effects’), we imposed periodic boundary conditions:

\[
\begin{align*}
\text{ATGCTTTTGCTAGCGCTATGCTTTTCGCAAACCTCAT} \\
\text{ATGCTTTTTAAAAACCGAGCTGGTCAGCGCTAAGCGCT}
\end{align*}
\]
Periodic boundary conditions

To simplify the calculations (avoiding ‘edge effects’), we imposed periodic boundary conditions:
To simplify the calculations (avoiding ‘edge effects’), we imposed periodic boundary conditions:

Now, for \( k = 7 \) we have \( D_2 = 4 \).
Markovian sequences

The i.i.d. is not realistic. Real DNA sequences are more realistically modelled as Markovian (up to fifth order).

For first order:

\[
\text{Prob}(A_{i+1} = u \mid A_i = v) = M_{u,v},
\]

\[u, v \in \{A, C, G, T\}\]

where

\[0 \leq M_{u,v} \leq 1; \quad \sum_v M_{u,v} = 1.\]
Markovian sequences

The i.i.d. is not realistic. Real DNA sequences are more realistically modelled as Markovian (up to fifth order).

For first order:

\[
\text{Prob}(A_{i+1} = u \mid A_i = v) = M_{u,v}, \quad u, v \in \{A, C, G, T\}
\]

where

\[
0 \leq M_{u,v} \leq 1; \quad \sum_v M_{u,v} = 1.
\]

First problem:
How do we define periodic boundary conditions for a Markovian sequence?
Define a Markov chain

\[ \ldots X_{n-1}, X_n, X_1, X_2, \ldots, X_n, X_1, X_2, \ldots \]

with periodic boundary conditions (PBCs) via the following algorithm:

1. Choose \( X_1 \) from any distribution \( \pi(u), u \in \{1, \ldots, d\} \), where \( 0 \leq \pi(u) \leq 1; \sum_u \pi(u) = 1 \). Thus \( \Pr(X_1 = u) = \pi(u) \).

2. Choose \( X_2, \ldots, X_{n+1} \) via the Markov matrix \( M \),
   \[ \Pr(X_{i+1} = v \mid X_i = u) = M_{u,v}, i = 1, \ldots, n. \]

3. If \( X_{n+1} = X_1 \), accept \( X_1, X_2, \ldots, X_n \), otherwise return to Step 1 and repeat the procedure.
No privileged starting point

We further wish to restrict the definition to repeating Markov chains with no privileged starting point, by which we mean

$$\Pr(X = x) = \Pr (X = (x_{i+1} \ldots x_n, x_1 \ldots x_i)),$$

for all $i = 1, \ldots, n - 1$,

where $X = (X_1X_2 \ldots X_n)$. 
No privileged starting point

We further wish to restrict the definition to repeating Markov chains with no privileged starting point, by which we mean

\[ \Pr(X = x) = \Pr(X = (x_{i+1} \ldots x_n, x_1 \ldots x_i)), \]

for all \( i = 1, \ldots, n - 1, \)

where \( X = (X_1X_2 \ldots X_n). \)

**Theorem 1**

\( X \) has no privileged starting point if and only if \( \pi(u) \) is a uniform distribution: \( \pi(u) = 1/d, u = 1, \ldots, d. \)
Probability of a specific sequence

Corollary 2

If $X$ is a Markov chain with no privileged starting point, the probability of any given sequence $x = (x_1 x_2 \ldots x_n)$ is

$$\Pr(X = x) = \frac{M_{x_1, x_2} M_{x_2, x_3} \cdots M_{x_n, x_1}}{\text{tr}(M^n)}$$
Mean of $D_2$

For two sequences $A$ and $B$ of length $m$ and $n$, both generated using the matrix $M$, and word length $k$, \[
E(D_2) = \frac{mn \text{tr} \left[ \left( M^{m-k+1} \circ M^{n-k+1} \right) \left( M \circ M \right)^{k-1} \right]}{\text{tr}(M^m) \text{tr}(M^n)},
\]
where $\circ$ indicates the Hadamard product of matrices

\[(P \circ Q)_{r,s} = P_{r,s}Q_{r,s}.\]
Mean of $D_2$

Given two sequences

$A = (A_1, A_2, \ldots, A_m)$ and $B = (B_1, B_2, \ldots, B_n)$,
define the word-match indicator

$$I_{i,j} =\begin{cases} 
  1 & \text{if } k\text{-word at position } i \text{ in } A \text{ matches } \\
  0 & \text{otherwise.}
\end{cases}$$

Then

$$D_2 = \sum_{i=1}^{m} \sum_{j=1}^{n} I_{i,j}$$

and

$$E(D_2) = \sum_{i=1}^{m} \sum_{j=1}^{n} E(I_{i,j}) = \sum_{i=1}^{m} \sum_{j=1}^{n} \Pr(I_{i,j} = 1).$$
Variance of $D_2$

The variance of $D_2$ is much harder but can be done, at least for Markov order 1:

$$\text{Var}(D_2) = \text{Var} \left( \sum_{i,j} I_{i,j} \right) = \mathbb{E} \left( \left( \sum_{i,j} I_{i,j} \right)^2 \right) - \left( \mathbb{E} \left( \sum_{i,j} I_{i,j} \right) \right)^2$$

$$= \left( \sum_{i,j,i',j'} \mathbb{E}(I_{i,j}, I_{i',j'}) \right) - \mathbb{E}(D_2)^2.$$

The difficult part is $\mathbb{E}(I_{i,j}, I_{i',j'})$, the probability of word matches like this:
Variance of $D_2$

The formula for $\text{Var}(D_2)$ with periodic boundary conditions and Markov order 1 is complicated ... but is easily evaluated.

\begin{align*}
\text{Var}(D_2) &= \sum_{l=1}^{\infty} \sum_{n=1}^{N} \text{Var}(D_{l,n}) \\
&= \sum_{l=1}^{\infty} \sum_{n=1}^{N} \text{Var}(D_{l,n}) \\
&= 0.1 + 1.2 + 2.3 + 3.4 \\
&= 59.8. \\
\end{align*}

where

\begin{align*}
\text{Var}(D_{l,n}) &= \frac{1}{N} \sum_{k=1}^{N} (D_{l,n} - \mu)^2 \\
&= \frac{1}{N} \sum_{k=1}^{N} \left( (M_k - M_{k+l})^2 + (M_{k+l} - M_{k+2l})^2 + \cdots + (M_{k+(N-l)l} - M_{k})^2 \right). \\
\end{align*}

and $f$ in sequence $Y$.

\begin{align*}
V_f &= \frac{\text{Var}(D_2)}{\text{Var}(D_2)^2} \\
&= \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \text{Var}(D_{i,j})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \text{Var}(D_{i,j})} \\
&= \frac{1}{N} + \frac{1}{N} + \frac{1}{N} + \frac{1}{N} \\
&= \frac{4}{N}. \\
\end{align*}

The positioning reflects the degree of overlap between words in each of the two sequences, and is illustrated in Fig. 2. We assume $m,n \leq N$, which will almost certainly be the case in any biological applications.

We will write a Hadamard product of $p$ factors, $M \odot \cdots \odot M$, using the shorthand notation $M^p$. With this notation, the contributions to the variance are:

\begin{align*}
V_f &= \frac{\text{Var}(D_2)}{\text{Var}(D_2)^2} \\
&= \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \text{Var}(D_{i,j})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \text{Var}(D_{i,j})} \\
&= \frac{1}{N} + \frac{1}{N} + \frac{1}{N} + \frac{1}{N} \\
&= \frac{4}{N}. \\
\end{align*}

The positioning reflects the degree of overlap between words in each of the two sequences, and is illustrated in Fig. 2. We assume $m,n \leq N$, which will almost certainly be the case in any biological applications.

We will write a Hadamard product of $p$ factors, $M \odot \cdots \odot M$, using the shorthand notation $M^p$. With this notation, the contributions to the variance are:
Verification by simulation

1. For a given order 1 Markov matrix, generate 10,000 random pairs of Markovian sequences with periodic boundary conditions (R scripts).

2. Obtain the value of $D_2$ for each pair (SAFT program, written in C).

3. Compare empirical cumulative distribution function of $D_2$ with that of Normal and Pólya-Aeppli (compound Poisson) distributions using theoretical $E(D_2)$ and $\text{Var}(D_2)$ (R scripts).
Generation of sequences

Sequences are generated by using the algorithm shown previously.

Seemingly more efficient algorithms, such as continuing with $X_{n+2}, X_{n+3}, \ldots$ until $X_k$ matches $X_{n+k}$ can yield different distributions.
Calculating $D_2$

The SAFT program was written to compare a given sequence against a database of sequences, by calculating $D_2$ for each pair.

The program was adapted to take two lists of sequences $\mathcal{A}, \mathcal{B}$, and determine $D_2$ for each corresponding pair $\mathcal{A}_k, \mathcal{B}_k$. 
Comparing against known distributions

The R statistical system was used to

▶ Use the `t.test` function in R to compare the empirical and theoretical mean values.

▶ Use the `qchisq` function in R to produce a confidence interval for the variance, and test the variance using this interval.

▶ Use the `ks.test` function in R to compare the empirical distribution of $D_2$ against the Normal and the Pólya-Aeppli distributions, using the theoretical mean and variance. The R function `rPolyaAeppli` was written for this purpose.

▶ Plot cumulative distributions.
Results for one random Markov matrix

Randomly chosen Markov matrix $M$

$n = 100 \ k = 4$

$$M = \begin{pmatrix} 0.344 & 0.087 & 0.283 & 0.286 \\ 0.245 & 0.226 & 0.254 & 0.275 \\ 0.16 & 0.181 & 0.309 & 0.35 \\ 0.187 & 0.34 & 0.267 & 0.207 \end{pmatrix}$$
Results for one random Markov matrix

Randomly chosen Markov matrix