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

AUSTRALIAN RESEARCH COUNCIL
Centre of Excellence for Mathematics and Statistics of Complex Systems

ARC

ANU
THE AUSTRALIAN NATIONAL UNIVERSITY

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 $k = 7 \ldots$

A: ATGCTTTGCTAGCGCTATGCTTTCGCAAACTCAT

B: ATGCTTTTAAAACCGAGCTGGTCAGCGCTAAGCGCT

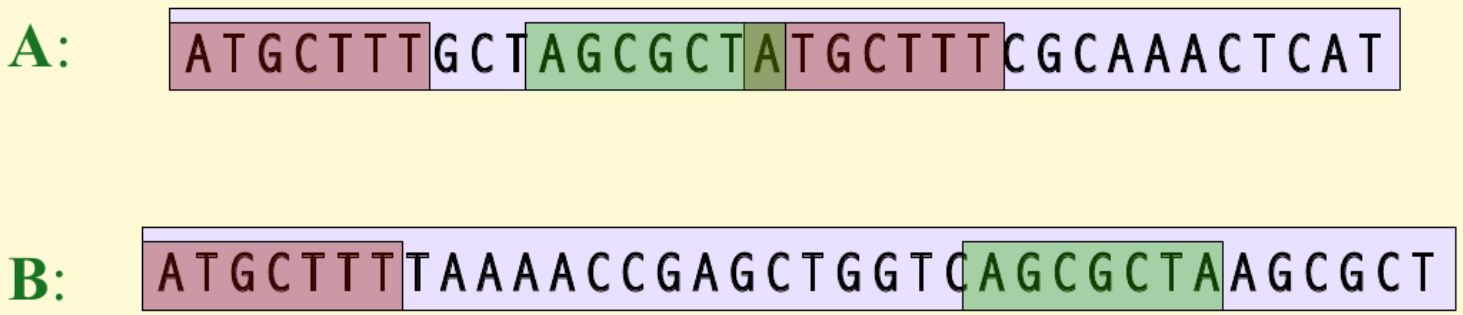
Definition of D_2

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 $k = 7 \ldots$



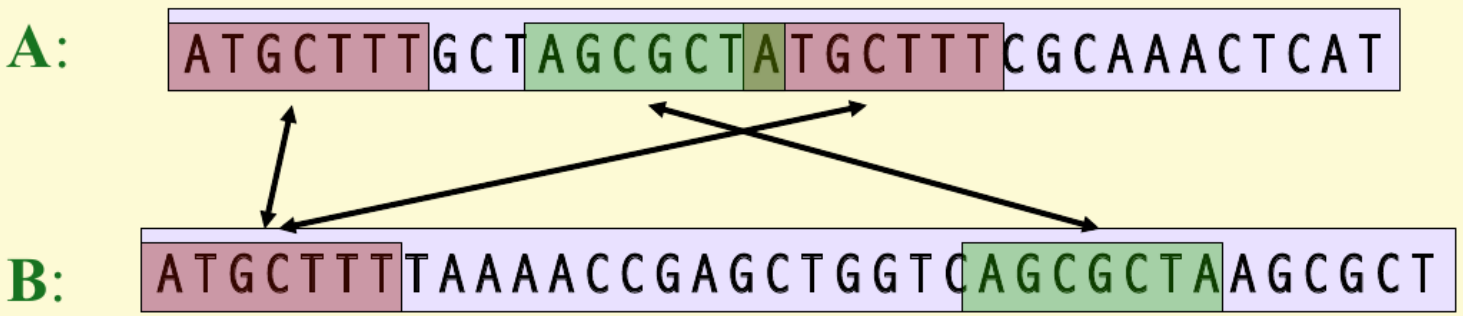
Definition of D_2

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 $k = 7 \ldots$



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

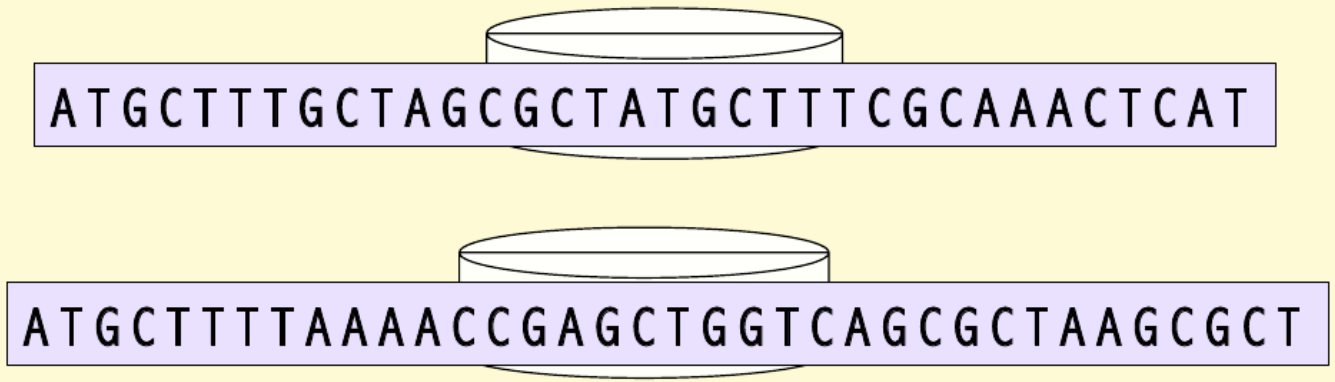
Our previous results

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

We have exact formulas for $\mathbf{E}(\boldsymbol{D_2})$ (simple) and $\mathbf{Var}(\boldsymbol{D_2})$ (complicated), and accurate approximations for the distribution of $\boldsymbol{D_2}$ 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:

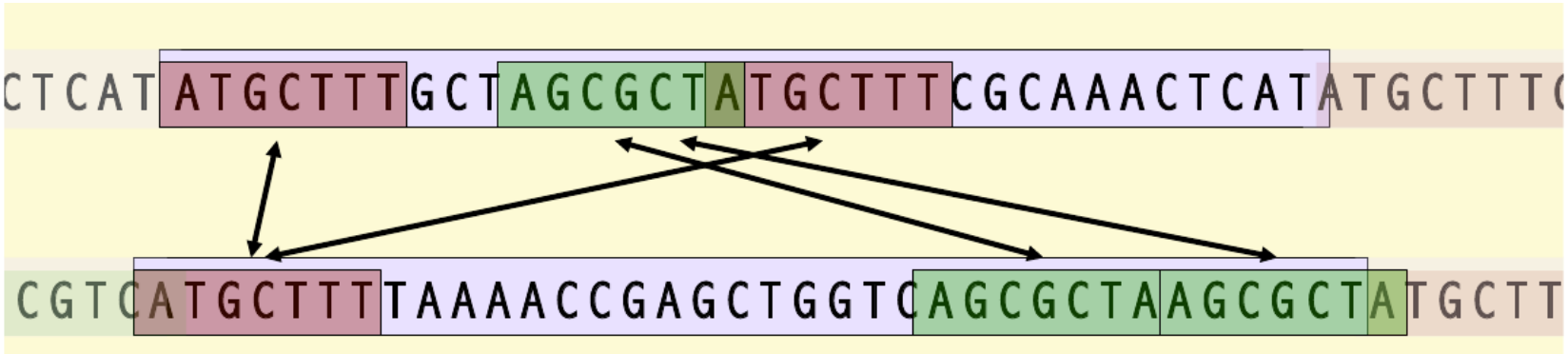


Periodic boundary conditions

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

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:

$$\mathbf{Prob}(A_{i+1} = u \mid A_i = v) = M_{u,v},$$
$$u, v \in \{A, C, G, T\}$$

where

$$0 \leqslant M_{u,v} \leqslant 1; \quad \sum_v M_{u,v} = 1.$$

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:

$$\mathbf{Prob}(A_{i+1} = u \mid A_i = v) = M_{u,v},$$
$$u, v \in \{A, C, G, T\}$$

where

$$0 \leqslant M_{u,v} \leqslant 1; \quad \sum_v M_{u,v} = 1.$$

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

Markov chain with periodic boundary conditions

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 $0 \leqslant \pi(u) \leqslant 1; \sum_u \pi(u) = 1$. Thus $\mathbf{Pr}(X1 = u) = \pi(u)$.

2. Choose $X_2, \ldots, X_{n+1}$ via the Markov matrix M, $\mathbf{Pr}(X_{i+1} = v \mid X_1 = u) = M_{u,v}, i = 1, \ldots, n$.

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

$$\mathbf{Pr}(X = x) = \mathbf{Pr}\left(X = (x_{i+1} \dots x_n, x_1 \dots x_i)\right),$$
$$\text{for all } i = 1, \dots, n-1,$$
$$\text{where } X = (X_1 X_2 \dots X_n).$$

No privileged starting point

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

$$\mathbf{Pr}(X = x) = \mathbf{Pr}\left(X = (x_{i+1} \dots x_n, x_1 \dots x_i)\right),$$
$$\text{for all } i = 1, \dots, n-1,$$
$$\text{where } X = (X_1 X_2 \dots X_n).$$

Theorem 1

X has no privileged starting point if and only if $\pi(u)$ is a uniform distribution: $\pi(u) = 1/d, u = 1, \dots, d$.

Probability of a specific sequence

Corollary 2

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

$$\mathbf{Pr}(\boldsymbol{X} = \boldsymbol{x}) = \frac{\boldsymbol{M_{x_1,x_2} M_{x_2,x_3} \cdots M_{x_n,x_1}}}{\mathbf{tr}(\boldsymbol{M^n})}$$

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,

$$\mathbf{E}(D_2) = \frac{mn \operatorname{tr}\left[(M^{m-k+1} \circ M^{n-k+1})(M \circ M)^{k-1}\right]}{\operatorname{tr}(M^m) \operatorname{tr}(M^n)},$$

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

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

Then

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

and

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

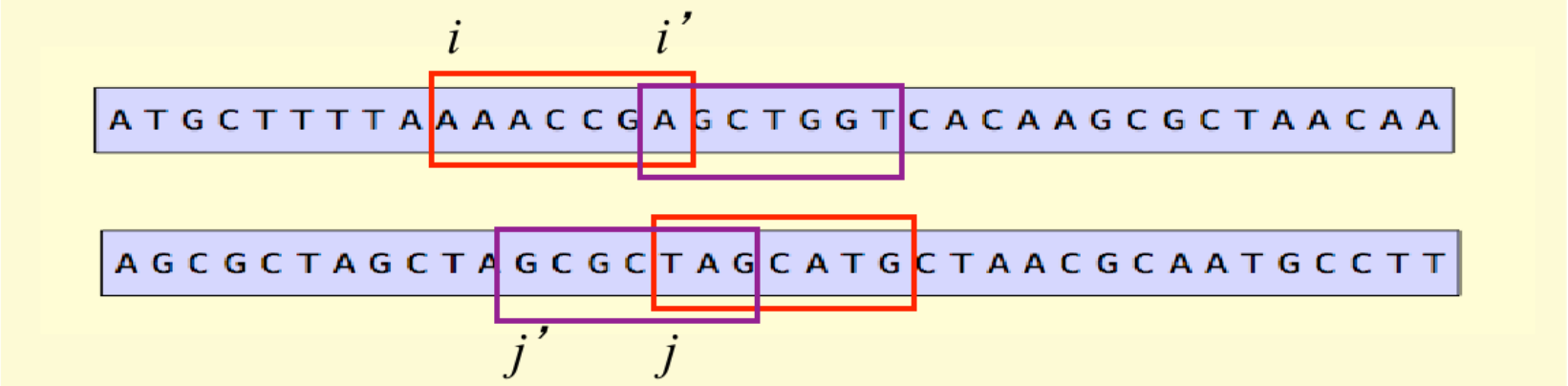
Variance of D_2

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

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

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

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



Variance of D_2

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

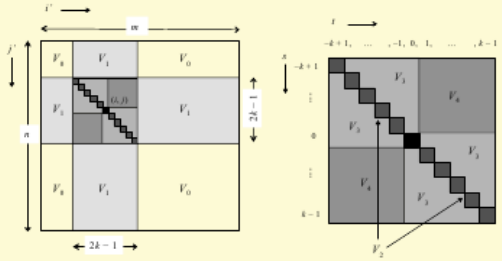


Figure 2: Contributions to $\mathrm{Var}(D_2)$ via the sum in Eq. (24). The left-hand diagram shows the (i', j')-plane for a fixed value of (i, j), shown as the black square. The right-hand diagram is an expanded view of the 'accordion' region $-k+1 \le s, t \le k-1$, where $t = i' - i$ and $s = j' - j$ up to PBCs.

and j' in sequence $\mathbf{Y}$,

$$
\begin{aligned}
E(D_2{}^2) &= \sum_{i,i'=1}^{m} \sum_{j,j'=1}^{n} E(I_{ij} I_{i'j'}) \\
&= \sum_{i,i'=1}^{m} \sum_{j,j'=1}^{n} \mathrm{Prob}\,(I_{ij} = 1, I_{i'j'} = 1) \\
&= V_0 + V_1 + V_2 + V_3 + V_4. \qquad (24)
\end{aligned}
$$

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

We will write a Hadamard product of q factors, $M \circ \ldots \circ M$, using the shorthand notation $M^{\circ q}$. With this notation, the contributions to the variance are:

$$
\begin{aligned}
V_0 = {} & \frac{mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \\
& \sum_{r=0}^{m-2k} \sum_{s=0}^{n-2k} \mathrm{tr}\left[(M^{r+1} \circ M^{s+1})(M \circ M)^{k-1} \times \right. \\
& \left. (M^{m-2k-r+1} \circ M^{n-2k-s+1})(M \circ M)^{k-1} \right], \qquad (25)
\end{aligned}
$$

$$
\begin{aligned}
V_1 = {} & \frac{mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \\
& \left\{ \sum_{r=0}^{n-2k} \left[\mathrm{tr}\left\{ [(M \circ M \circ M)^{k-1} \circ (M^{r+1})^T] \times \right. \right. \right. \\
& \left. \left. \left. (M^{m-k+1} \circ M^{n-2k-r+1}) \right\} \right. \right. \\
& \left. \left. + 2 \sum_{r=1}^{k-1} \mathrm{tr}\left\{ (M \circ M)^r \times \right. \right. \right. \\
& \left. \left. \left. [(M \circ M \circ M)^{k-r-1} \circ (M^{r+1})^T] \times \right. \right. \right. \\
& \left. \left. \left. (M \circ M)^r (M^{m-k-r+1} \circ M^{n-2k-s+1}) \right\} \right] \right. \\
& \left. + \text{the same with } m \text{ and } n \text{ interchanged.} \right\}, \qquad (26)
\end{aligned}
$$

$$
\begin{aligned}
V_2 = {} & \frac{mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \\
& \left\{ \mathrm{tr}\left[(M^{m-k+1} \circ M^{n-k+1})(M \circ M)^{k-1} \right] \right. \\
& \left. + 2 \sum_{t=1}^{k-1} \mathrm{tr}\left[(M^{m-k-t+1} \circ M^{n-k-t+1}) \times \right. \right. \\
& \left. \left. (M \circ M)^{k+t-1} \right] \right\}, \qquad (27)
\end{aligned}
$$

$$
\begin{aligned}
V_3 = {} & \frac{2mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \\
& \sum_{t=1}^{k-1} \sum_{s=0}^{t-1} \mathrm{tr}\left[(M \circ M)^t Q (M \circ M)^s \times \right. \\
& (M^{m-k-t+1} \circ M^{n-k-s+1} + \\
& \left. M^{n-k-t+1} \circ M^{m-k-s+1}) \right], \qquad (28)
\end{aligned}
$$

where

$$
Q = \begin{cases} (M^{\circ(2\nu+3)})^{\rho-1} \circ [(M^{\circ(2\nu+1)})^{t-s-\rho+1}]^T & \text{if } \rho > 0, \\ (M^{\circ(2\nu+1)})^{t-s-1} \circ (M^{\circ(2\nu-1)})^T & \text{if } \rho = 0, \end{cases} \qquad (29)
$$

and

$$
\nu = \left\lfloor \frac{k-s}{t-s} \right\rfloor, \qquad \rho = (k-s) \bmod (t-s). \qquad (30)
$$

Finally,

$$
V_4 = \frac{2mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \sum_{r,t=1}^{k-1} \mathrm{tr}\, U, \qquad (31)
$$

where

$$
U = \begin{cases}
\left\{ (M^{\circ(2\nu+1)})^{t-1} \circ (M^{m-k-r+1})^T \right\} M^{\circ 2\nu} \times \\
\quad \left\{ (M^{\circ(2\nu+1)})^{r-1} \circ (M^{n-k-r+1})^T \right\} M^{\circ 2\nu} \\
\qquad \text{if } \zeta = 0, \\
\left\{ (M^{\circ(2\nu+1)})^{r-\zeta+1} \circ M^{m-k-t+1} \right\} \times \\
(M^{\circ(2\nu+2)})^{\zeta-1} \times \\
\quad \left\{ (M^{\circ(2\nu+1)})^{t-\zeta+1} \circ M^{n-k-r+1} \right\} \times \\
\quad (M^{\circ(2\nu+2)})^{\zeta-1} \\
\qquad \text{if } 0 < \zeta \le r, t, \\
\left\{ (M^{\circ(2\nu+3)})^{\zeta-r-1} \circ (M^{m-k-t+1})^T \right\} \times \\
(M^{\circ(2\nu+2)})^r \left\{ (M^{\circ(2\nu+1)})^{t-\zeta+1} \circ M^{n-k-r+1} \right\} \\
\quad \times (M^{\circ(2\nu+2)})^r \\
\qquad \text{if } r < \zeta \le t, \\
\{\text{as above with } m \text{ and } n \text{ interchanged} \\
\quad \text{and } r \text{ and } t \text{ interchanged}\} \\
\qquad \text{if } t < \zeta \le r, \\
\left\{ (M^{\circ(2\nu+3)})^{\zeta-r-1} \circ (M^{m-k-t+1})^T \right\} \times \\
(M^{\circ(2\nu+2)})^{t+r-\zeta+1} \times \\
\quad \left\{ (M^{\circ(2\nu+3)})^{\zeta-t-1} \circ (M^{n-k-r+1})^T \right\} \times \\
\quad (M^{\circ(2\nu+2)})^{t+r-\zeta+1} \\
\qquad \text{if } r, t < \zeta,
\end{cases}
$$

and

$$
\nu = \left\lfloor \frac{k}{r+t} \right\rfloor, \qquad \zeta = k \bmod (r+t). \qquad (32)
$$

. . . but is easily evaluated.

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 $\boldsymbol{D_2}$ for each pair (SAFT program, written in C).

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

Generation of sequences

Sequences are generated by using the algorithm shown previously.

Seemingly more efficient algorithms, such as continuing with $\boldsymbol{X_{n+2}, X_{n+3}, \ldots}$ until $\boldsymbol{X_k}$ matches $\boldsymbol{X_{n+k}}$ 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 $\mathcal{A}_k, \mathcal{B}_k$.

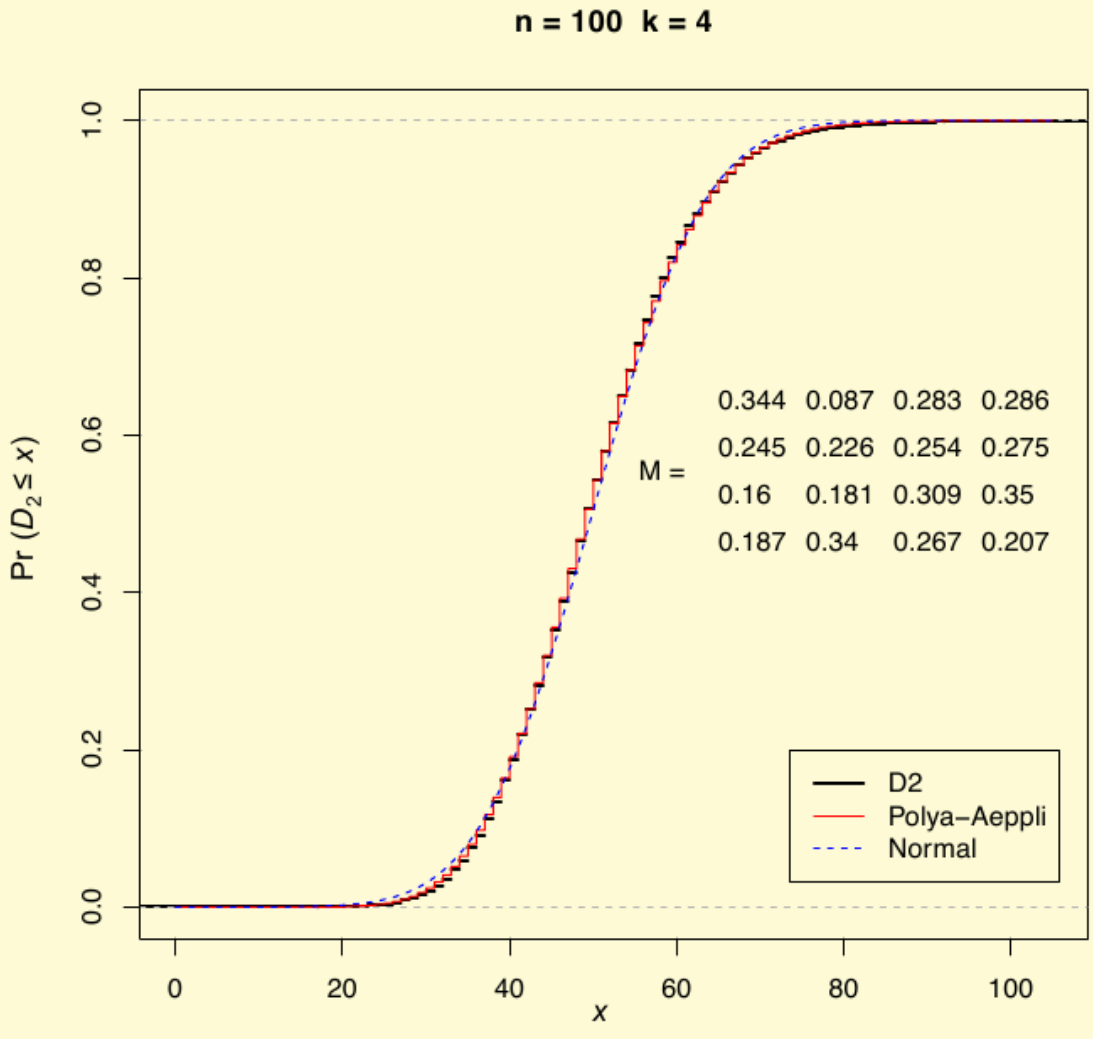
Comparing against known distributions

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 $\boldsymbol{D_2}$ against the Normal and the Pólya-Aeppli distributions, using the theoretical mean and variance. The R function `pPolyaAeppli` was written for this purpose.
- Plot cumulative distributions.

Results for one random Markov matrix

Randomly chosen Markov matrix *M*



Results for one random Markov matrix