Alignment-free comparison of biological sequences

Conrad Burden, Sylvain Forêt, *Paul Leopardi

Mathematical Sciences Institute, Australian National University.

For presentation at ANZIAM, Newcastle, 2013.

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

4 February 2013

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

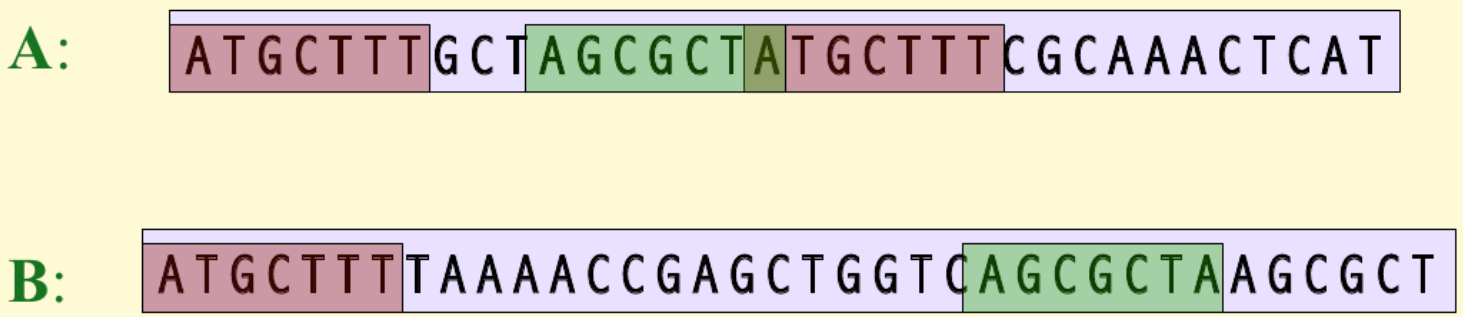
Definition of D_2

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 $k = 7 \ldots$



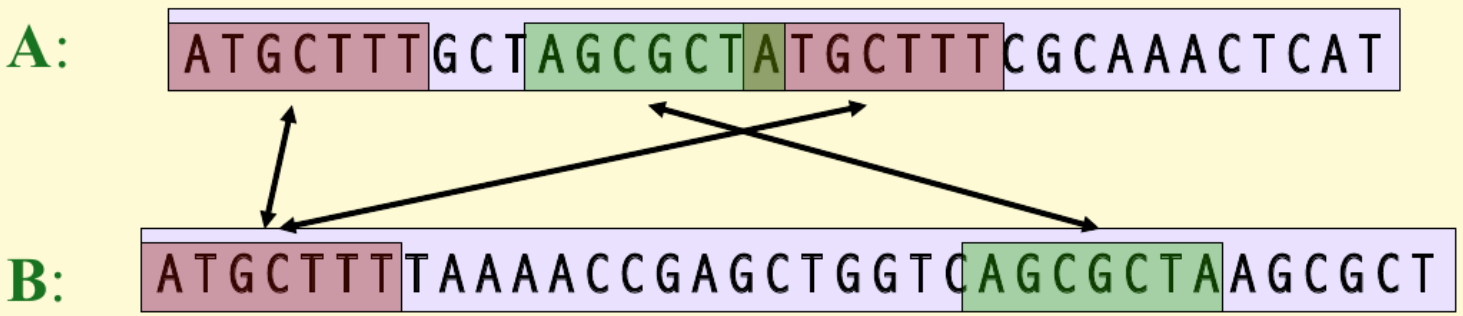
Definition of D_2

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 $k = 7, D_2 = 3$.

Markovian sequences

Real DNA sequences are modelled as Markovian.

For first order:

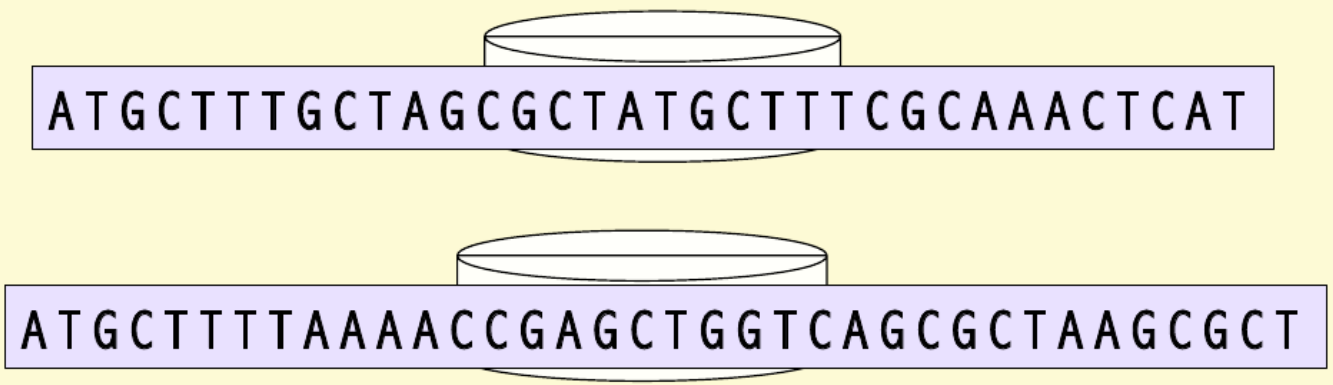
$$\mathbf{Prob}(A_{i+1} = u \mid A_i = v) = M_{u,v},$$
$$u, v \in \{A, C, G, T\}$$

where

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

Periodic boundary conditions

To simplify the calculations of theoretical mean and variance (avoiding 'edge effects'), we impose periodic boundary conditions:

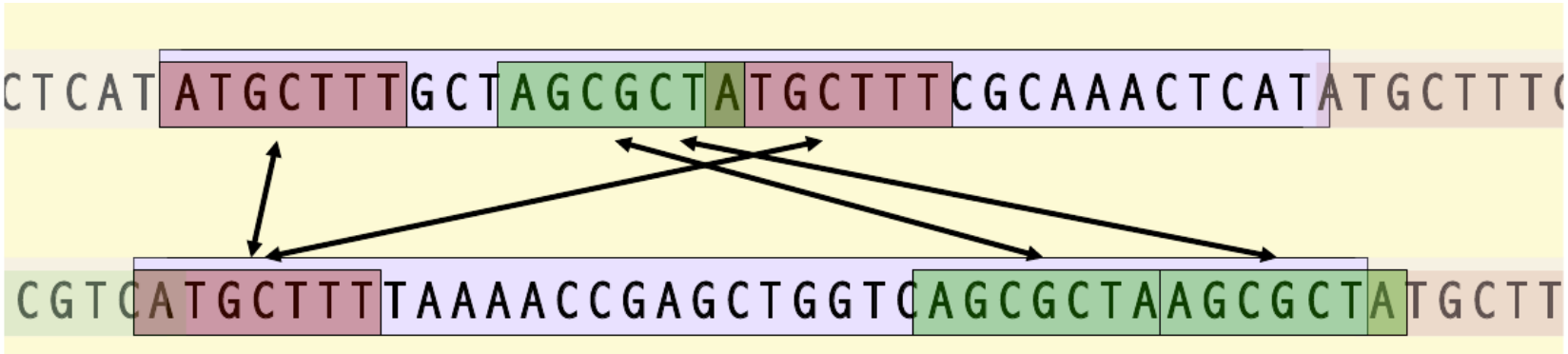


Periodic boundary conditions

To simplify the calculations of theoretical mean and variance (avoiding 'edge effects'), we impose periodic boundary conditions:

Periodic boundary conditions

To simplify the calculations of theoretical mean and variance (avoiding 'edge effects'), we impose periodic boundary conditions:



Now, for $k = 7$ we have $D_2 = 4$.

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 $\boldsymbol{x = (x_1 x_2 \dots x_n)}$ is

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

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

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

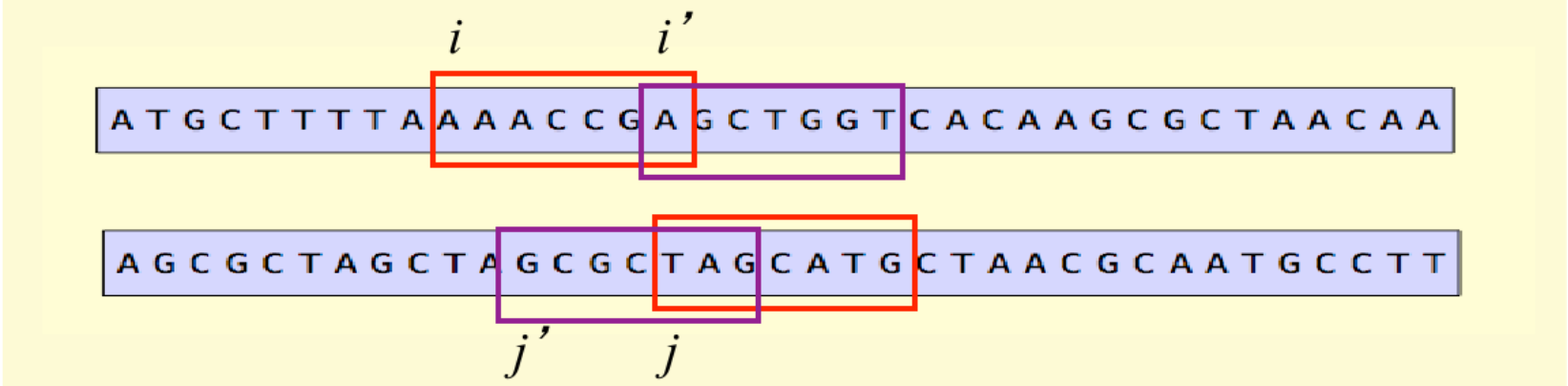
Variance of D_2

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 . . .

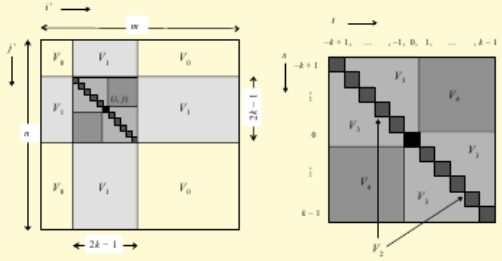


Figure 2: Contributions to $\mathrm{Var}(D_2)$ via the sum in Eq. (24). The left-hand diagram shows the (i', j')-plane for a fixed value of (i, j), shown as the black square. The right-hand diagram is an expanded view of the 'accordion' region $-k+1 \le s,t \le k-1$, where $t = i' - i$ and $s = j' -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

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

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

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:

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

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

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

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

where

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

and

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

Finally,

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

where

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

and

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

. . . but is easily evaluated.

Verification by simulation

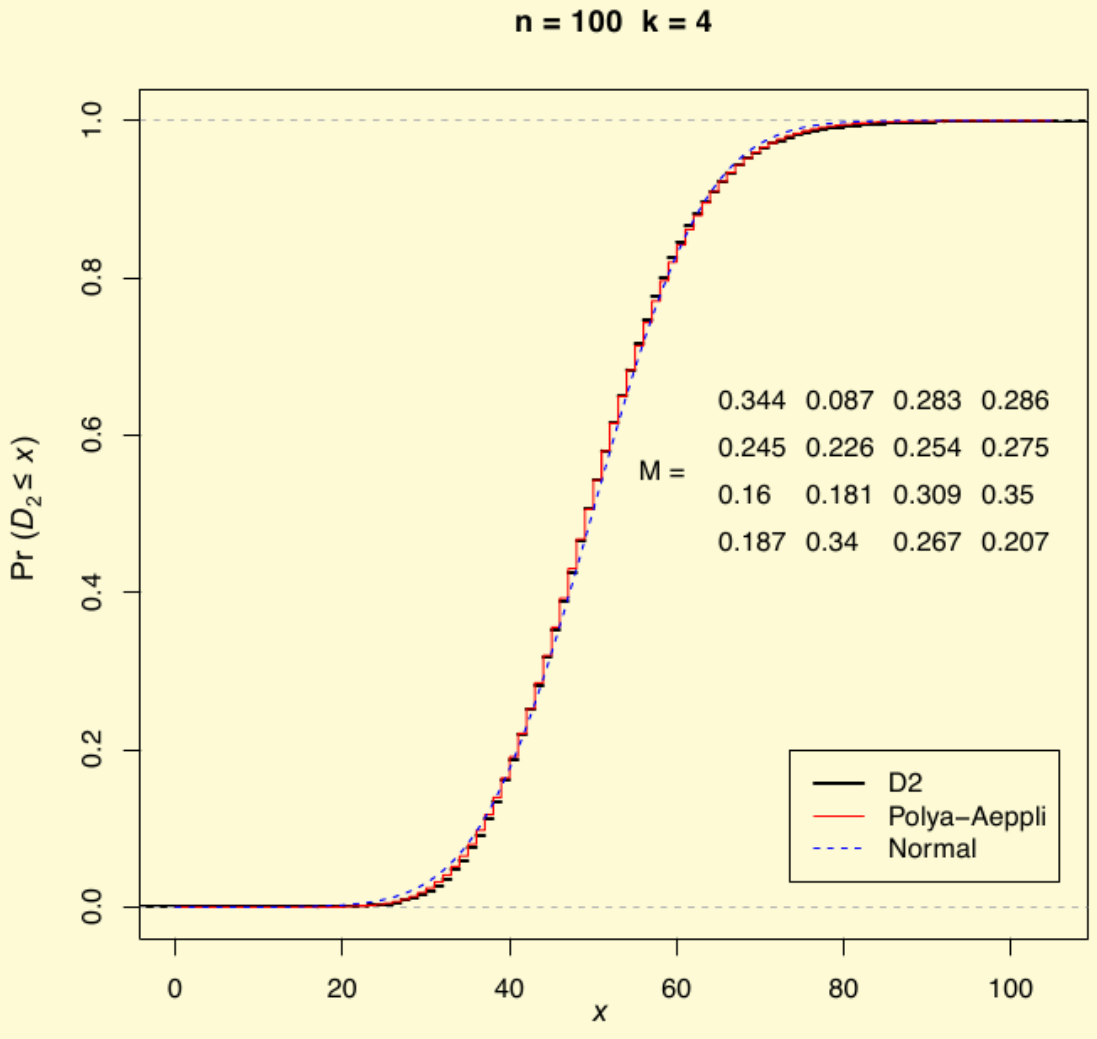
1. For a given order 1 Markov matrix, generate 10,000 random pairs of Markovian sequences with periodic boundary conditions (R scripts).

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

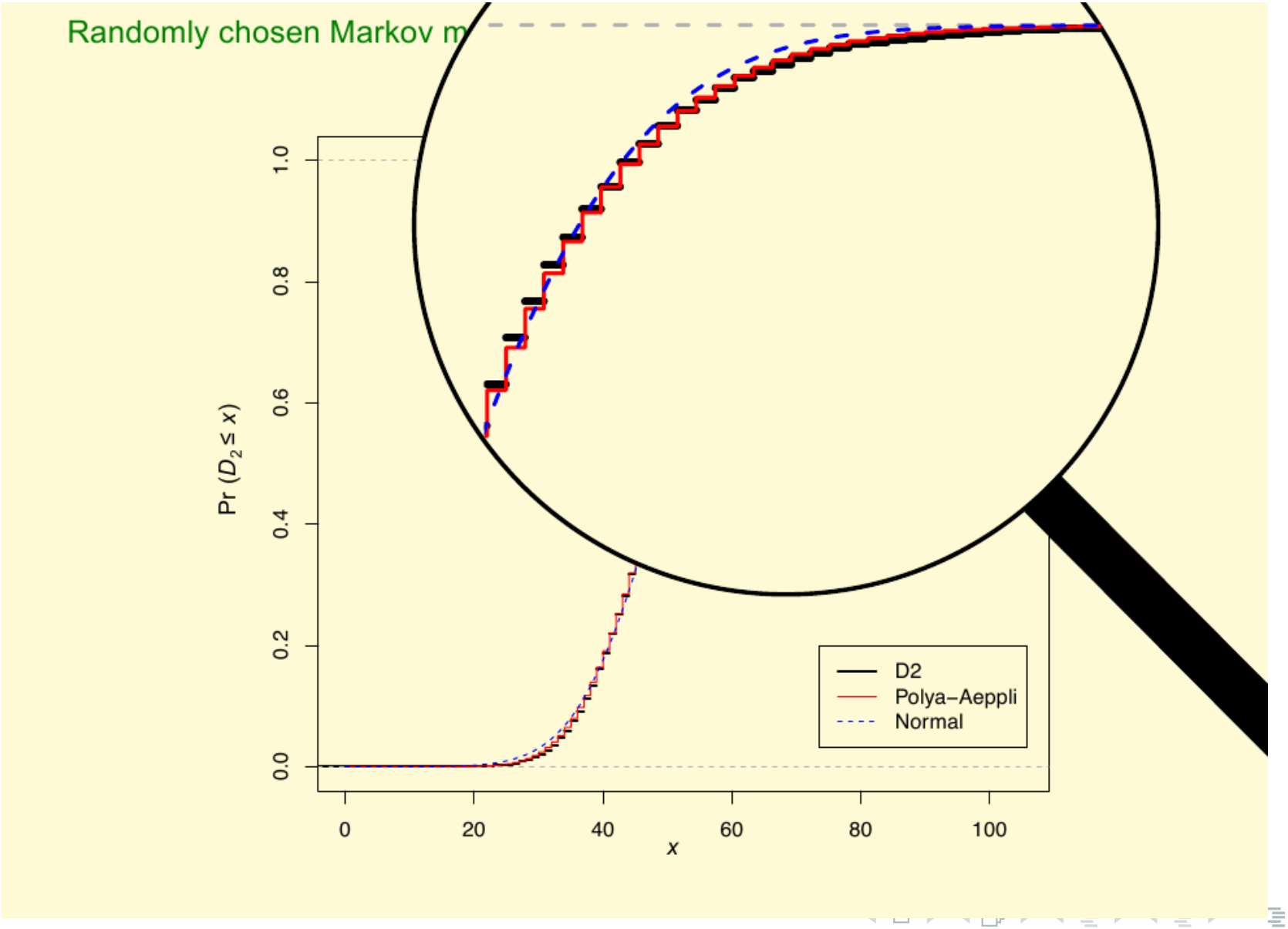
3. Compare empirical cumulative distribution function of $\boldsymbol{D_2}$ with that of Normal and Pólya-Aeppli (compound Poisson) distributions using theoretical $\mathbf{E}(\boldsymbol{D_2})$ and $\mathbf{Var}(\boldsymbol{D_2})$ (R scripts).

Results for a random Markov matrix

Randomly chosen Markov matrix *M*



Results for a random Markov matrix



DNA is messy

Real DNA is messy

messy with repeateats of different length and complexity,

and contains unknown r-gi-ns.

- The Ensembl database marks unknown regions and masks repetitive regions including *tandem repeats*.
- The tantan program masks simple repeats.

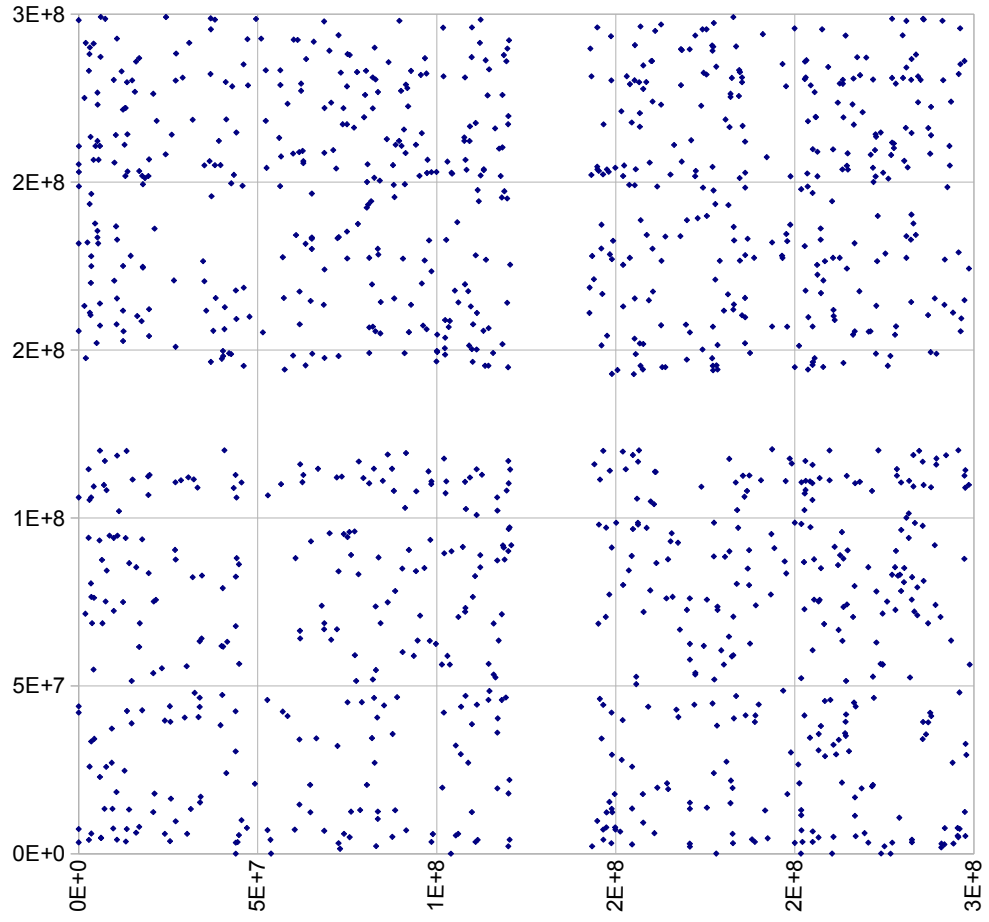
(Ensembl: Wellcome Trust Sanger Institute and European Bioinformatics Institute, 2012; tantan: Firth, 2011)

To compare D_2 from DNA with Markov models:

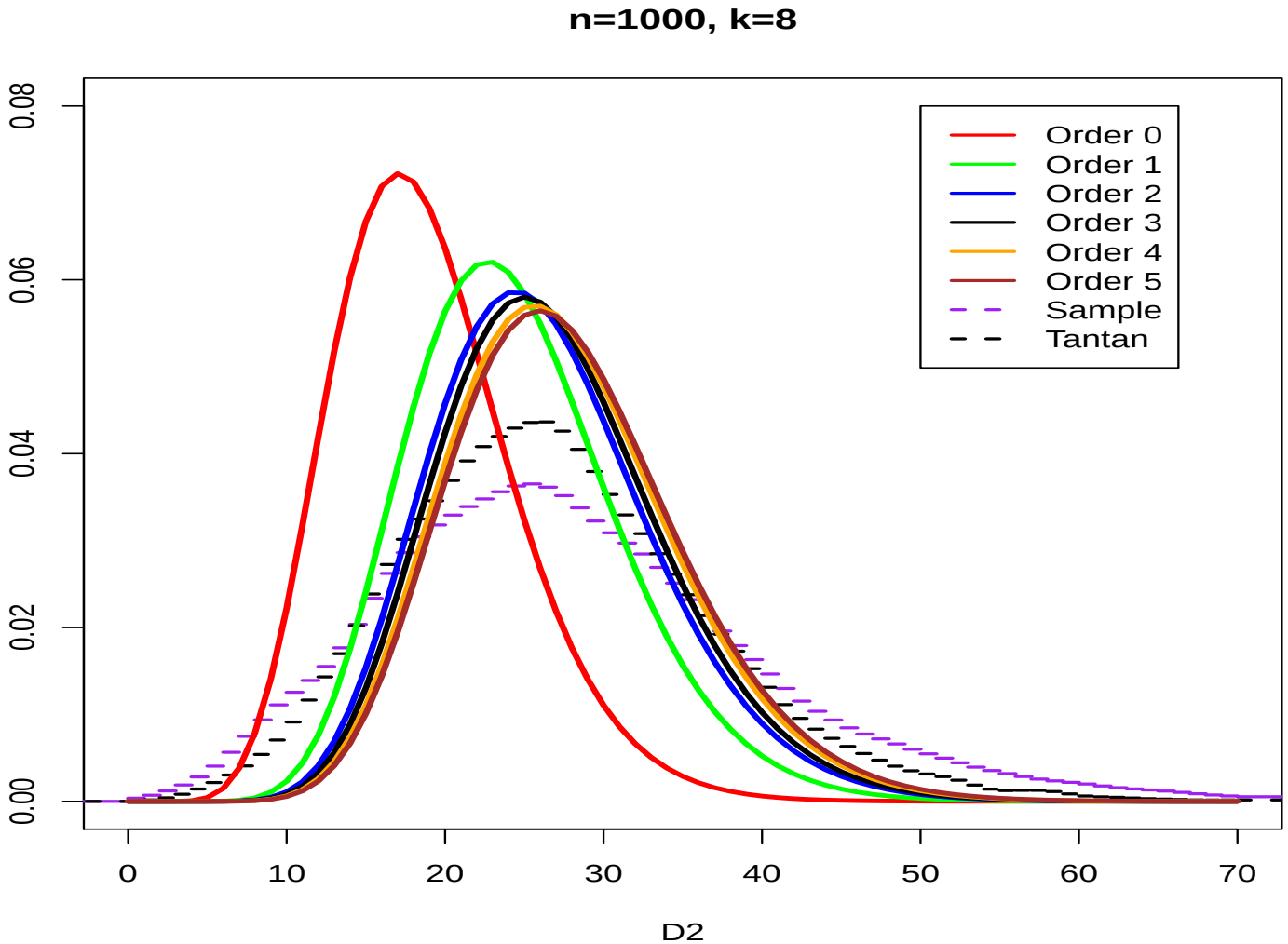
1. Obtain and mask a DNA sequence, yielding a series of unmasked regions;
2. For a fixed length n, produce a random sample of 10,000 pairs of sequences from the regions, using word length k to yield sequences with periodic boundary conditions;
3. Use SAFT program to calculate the D_2 value of each pair;
4. Given Markov order ω, compute the Markov matrix M from the DNA regions;
5. Given Markov matrix M, word length k and sequence length n, compute the theoretical mean and variance;
6. Compare the empirical distribution of D_2 values with a Gamma distribution using the theoretical mean and variance.

Human Chromosome 1: sample pairs

Note the gap around the centromere.



Human Chromosome 1: D_2 vs Markov models



What's next?

1. Compare chromosomes with their *stationary* k*-mer spectrum* to look for regions of interest.
   Have determined the theoretical mean and variance for this case. The formula for the variance is much simpler than the variance for D_2 between two sequences.
   Need to modify the SAFT program to compare a stationary k-mer spectrum to a database of sequences.
2. Scale up the SAFT program to work quickly with large databases.
   This includes testing parallel code on NCI clusters during 2013.
3. Release the SAFT program as open source software.
   Anticipated some time in 2013.