Poisson Process Approximation for Sequence Repeats, and Sequencing by Hybridization

Richard Arratia[1]

Daniela Martin[1]

Gesine Reinert[1]

Michael S. Waterman[1,2]

Department of Mathematics[1]

Department of Biological Sciences[2]

University of Southern California

Los Angeles, CA 90089-1113

January 18, 2001

Supported by grants from the National Institute of Health (GM 36230) and the National Science Foundation (DMS 90-05833).

Abstract

Sequencing by hybridization is a tool to determine a DNA sequence from the unordered list of all l-tuples contained in this sequence; typical numbers for l are $l = 8, 10, 12$. For theoretical purposes we assume that the multiset of all l-tuples is known. This multiset determines the DNA sequence uniquely if none of the so-called Ukkonen transformations are possible. These transformations require repeats of $(l-1)$-tuples in the sequence, with these repeats occurring in certain spatial patterns. We model DNA as an i.i.d. sequence.

We first prove Poisson process approximations for the process of indicators of all leftmost long repeats allowing self-overlap and for the process of indicators of all leftmost long repeats without self-overlap. Using the Chen-Stein method, we get bounds on the error of these approximations. As a corollary, we approximate the distribution of longest repeats. In the second step we analyze the spatial patterns of the repeats. Finally we combine these two steps to prove an approximation for the probability that a random sequence is uniquely recoverable from its list of l-tuples. For all our results we give some numerical examples including error bounds.

Keywords. Sequencing by hybridization, sequence repeats, DNA sequences, Chen-Stein method, Poisson process approximation, Ukkonen transformations

Contents

1 Introduction

One of the primary goals of the Human Genome Project is to increase the rate of DNA sequencing and to reduce its costs. While gel-based methods for determining the sequence of nucleotides (A, G, C, T) are being automated and improved, new approaches to DNA sequencing are being explored. Sequencing by hybridization (SBH) is a novel approach for determining DNA sequences that was proposed by several groups around the same time [Drmanac and Crkvenjakov (1987); Bains and Smith (1988); Lysov *et al.* (1988); Southern (1988); Macevicz (1989)].

Sequencing by hybridization is based on the following setup. A short single-stranded DNA of 8–25 letters is called a probe. The probe will bind or hybridize to a single-stranded target DNA if the substring complementary to the probe exists in the target. If the target is presented to all probes of length l (called l-tuples), then the l-tuple content of the target is known, and this data can be used to partially or fully determine the sequence of the target.

To accomplish the repeated probing of all 4^l probes of length l, all the probes are attached to the surface of a substrate where each probe is at a known position. This is called a sequencing chip. Then the labeled target is presented to the sequencing chip, and hybridizations are detected by an instrument sensitive to the label. The experimental challenges to making this approach successful include synthesizing and fixing DNA to the substrate in a reliable manner, devising efficient detection systems for DNA–DNA hybridization (*i.e.* for label detection), controlling the substantial differences between the binding energies of complementary DNAs of the same length but different base composition, and detecting perfectly complementary duplexes from those that are complementary except for one mismatched pair of bases. There have been rigorous efforts to overcome these challenges and significant progress has been made, although determination of longer sequences is not yet routine [Pevzner and Lipshutz (1994)]. Fodor and colleagues have developed light-directed polymer synthesis [Fodor *et al.* (1991)]; [Fodor *et al.* (1993)]; [Pease *et al.* (1994)] and recently synthesized a sequencing chip with all 4^8 8-tuples. A sequencing chip with all 4^{10} 10-tuples is a near term possibility.

Certainly the experimental aspects of sequencing by hybridization are of importance in developing the technology; in addition, the computational and mathematical sides of sequencing by hybridization are critical too.

To understand the basic problem, we consider a mathematical idealization. A sequence $\mathbf{a} = a_1a_2...a_m$ is to be sequenced, and the data is the multiset of all l-tuples present in the sequence, known as the l-spectrum of $\mathbf{a}$, $S_l(\mathbf{a})$. The multiset forgets the order in which the l-tuples occur, but it does keep track of multiple occurrences. This multiplicity information is not currently present in the physical data, but makes the mathematical analysis tractable. [It is natural to first pose the sequence recoverability problem as a traveling salesman or Hamiltonian path problem. The graph $G_{\mathcal{H}} = (V_{\mathcal{H}}, E_{\mathcal{H}})$ for the Hamiltonian path problem has vertex set $V_{\mathcal{H}} =$ the set underlying $S_l(\mathbf{a})$, and $(\mathbf{u}, \mathbf{v}) \in E_{\mathcal{H}}$ when $\mathbf{u} = u_1u_2 \dots u_l, \;\; \mathbf{v} = v_1v_2 \dots v_l \in S_l(\mathbf{a})$ and $u_2u_3 \dots u_l = v_1v_2 \dots v_{l-1}$. This Hamiltonian path problem, visiting all the vertices, is computationally difficult.] Pevzner (1989) employed de Bruijn sequences (see [van Lint and Wilson (1992)]) to treat this problem as an Eulerian path problem, finding a path which uses all the edges. The vertices of the graph for the Eulerian path problem are the $(l-1)$-tuples from $S_{l-1}(\mathbf{a})$, and the directed edges, with multiplicities, correspond to $S_l(\mathbf{a})$. Formally, the de Bruijn graph for $\mathbf{a}$ is $G_{\mathcal{E}} = (V_{\mathcal{E}}, E_{\mathcal{E}})$ where $V_{\mathcal{E}} =$ the set underlying $S_{l-1}(\mathbf{a})$ and $E_{\mathcal{E}} = S_l(\mathbf{a})$ is the edge multiset; an edge $c_1c_2 \cdots c_l$ goes from vertex $c_1c_2 \cdots c_{l-1}$ to vertex $c_2c_3 \cdots c_l$. The problem of determining the sequence $\mathbf{a}$ is translated into an Eulerian path problem, one for which there is an efficient solution. Furthermore, a word is uniquely recoverable from its l-spectrum if and only if there is only one Eulerian path for its de Bruijn graph.

A concrete example may help the reader get oriented. There are three very short examples at the start of section 3, but they are all atypical in that they involve self-overlapping repeats. For a longer but typical example, we take $m = 24, l = 4$, and the word $\mathbf{a}$ = GTGAC CATGG AAGAC TTGGA AGTT. The 4-spectrum is a multiset containing 21 4-tuples, of which only 18 are distinct. To emphasize that the multiset does not report the order in which its elements occur, we present it in alphabetical order; when the multiplicity of an element is greater than one, the multiplicity is given as a superscript. The 4-spectrum is $S_4 = \{$ AAGA, AAGT, ACCA, ACTT,AGAC,AGTT,ATGG,CATG,CCAT,CTTG, GAAG^2,GACC,GACT,GGAA^2, GTGA,TGAC,TGGA^2,TTGG $\}$. It is indeed hard to verify the above data, so we present the 4-spectrum again, in the same order, but with some extra information: each 4-tuple is subscripted by the position or positions where it begins, for example we write GTGA$_1$ and TGAC$_2$. Thus, the 4-spectrum, with additional information, is $S_4 = \{$ AAGA$_{11}$, AAGT$_{20}$, ACCA$_4$, ACTT$_{14}$,AGAC$_{12}$,AGTT$_{21}$,ATGG$_7$,CATG$_6$,CCAT$_5$, CTTG$_{15}$, GAAG$^2_{10,19}$,GACC$_3$,GACT$_{13}$,GGAA$^2_{9,18}$, GTGA$_1$,TGAC$_2$,TGGA$^2_{8,17}$,TTGG$_{16}\}$. This word $\mathbf{a}$ is not uniquely recoverable from its 4-spectrum because another word, namely $\mathbf{a}' =$ GTGAC TTGGA AGACC ATGGA AGTT, has the same 4-spectrum. The reader can verify this by brute force, but it is more easily checked by finding the de Bruijn graph of $\mathbf{a}$. For this example, the de Bruijn graph has 21 edges. There are 18 distinct edges, and three of these have multiplicity two. There are 17 distinct vertices, five of which are visited twice. The word $\mathbf{a}'$ has the same de Bruijn graph; the two words $\mathbf{a}$ and $\mathbf{a}'$ correspond to two different Eulerian paths in this graph.

Computational difficulties arise when the data only give the set underlying the l-spectrum, i.e. there is no information on multiple occurrences of l-tuples. A tougher problem is that the data may have errors. Nevertheless, it is instructive to first handle the mathematically idealized problem. We ask how big must l be to expect to uniquely determine a random sequence from its l-spectrum. For the random analysis, we assume that the m letters of the given word are independent and identically distributed; the distribution *may* be to assign probability 1/4 to each of A, C, G, T. This problem was the subject of a recent paper [Dyer, Frieze, Suen (1994)]. The relation between that paper and ours is described at the end of this section.

The first-order intuition for the answer is easy to derive; the critical consideration is whether l is large or small in comparison to $\log_{1/p}(m^2)$, where $p = \mathbf{P}$(two random letters match). The crude heuristic suggests first that for sequencing by hybridization data to give a unique answer, there should not be any l-tuple repeats. There are about $\binom{m}{2}p^l$ expected repeats of length l, and solving $1 = \binom{m}{2}p^l$ yields a critical boundary at $l = \log_{\frac{1}{p}}\binom{m}{2}$; so the longest repeat in a random sequence is approximately $l = \log_{\frac{1}{p}} \binom{m}{2} \approx \log_{\frac{1}{p}}(\frac{m^2}{2})$; see [Arratia, Waterman (1985)]. Making this intuition precise involves a Poisson process approximation to keep track of how many repeats there are and where they occur. The distributional limit theorem for the length of the longest repeat in a single sequence is proved in [Zubkov, Mikhailov (1974)], and also occurs as a special case of Theorem 7.2 in [Karlin, Ost (1987)]. Here we strengthen that result by giving error bounds.

A more careful analysis of the probability of unique recoverability starts with the Ukkonen-Pevzner criterion for unique recoverability, which we present as Theorem 6. In this criterion, the overwhelmingly most likely cause for a sequence to be not uniquely recoverable from its l-spectrum is having an interleaved pair of repeats of t-tuples, where $t = l - 1$. Loosely speaking, this cause is that the sequence has the form $\cdots a \cdots b \cdots a \cdots b \cdots$, where a, b denote t-tuples, and such a sequence is not recoverable because it has the same l-spectrum as the sequence obtained by swapping the two substrings that form the $\cdots$ in $a \cdots b$. [See section 3 for a precise description, which is valid even when the repeating t-tuples overlap.] Thus the probability of unique recoverability is approximately the probability of not having any interleaved pair of t-tuple repeats. In our previous example, the sequence is GTGAC CATGG AAGAC TTGGA AGTT, with $l = 4, t = 3$, and here $a =$ GAC, which begins at positions 3 and 13, while $b =$ TGG, which begins at positions 8 and 16. The $\cdots$ in the first $a \cdots b$ is CA and the $\cdots$ in the second $a \cdots b$ is T. Exchanging these produces GTGAC TTGGA AGACC ATGGA AGTT, a different sequence with the same 4-spectrum.

The next step is to use a Poisson approximation for the number of pairs of repeats. Repeats come in clumps, and indeed the number of repeats is not close to Poisson, so something like "maximal repeats" or "leftmost repeats" of length at least $t = l - 1$ must be considered. Returning to our guiding heuristic, the expected number of such repeats is about $\lambda \approx \binom{m}{2}(1-p)p^t$. A Poisson approximation takes the form: $\mathbf{P}(\text{ k repeats}) \approx e^{-\lambda}\lambda^k/k!$. An argument involving the Catalan numbers, $C_k = \frac{1}{k+1}\binom{2k}{k}$, shows that when there are k repeats, the probability of having no interleaved pair is $\approx k!2^k C_k/(2k)!$. Averaging over k yields that the probability of unique recoverability for a sequence of length m, from its l-spectrum, is approximately

$$f(\lambda) = \sum_{k \geq 0} e^{-\lambda} \frac{\lambda^k}{k!} \frac{k! 2^k C_k}{(2k)!} = e^{-\lambda} \sum \frac{(2\lambda)^k}{k!(k+1)!}, \tag{1}$$

using $\lambda = \binom{m}{2}(1-p)p^t$, with $t = l - 1$.

Here begins an overview of our paper. Section 2 give the details of a Poisson approximation for leftmost pairs of repeats, even allowing self-overlap. Poisson approximations for repeats within a single sequence or for matches between two sequences have occured in previous papers, for example [Zubkov, Mikhailov (1974)], [Arratia, Gordon, Waterman (1986)], [Karlin, Ost (1987)], [Arratia, Gordon, Waterman (1990a)], [Novak (1995)], and [Waterman (1995)]. At the level of these last three references, where a Poisson approximation is shown with an error bound of the form $O(m^{-\epsilon})$, two problems are very similar: the analysis of matching between two sequences, and the analysis of repeats within a single sequence. One novelty in the present paper is that we strive for very small error bounds, even for moderate values of m. At this level of careful bounding the two problems have substantial differences. For a more detailed discussion of the relationship between repeats within a single sequence and matches between two sequences, see [Reinert (1996)].

The reader mainly interested in sequencing by hybridization should simply accept, as the result of section 2, that the process $\mathbf{X}$ indicating where all leftmost repeats occur can be approximated by a much simpler process $\mathbf{Y}$ having independent Poisson coordinates, with an error of at most $b(m,t) \equiv \overline{b_1} + \overline{b_2}$. This error bound is a complicated but computable function of m, t, and the distribution used in our assumption that all letters are i.i.d. . The reader interested in sequence matching may find section 2 quite informative in its details. Our guiding principle was: with $t = 7, 8, 9, 10$ or 11, and m between 100 and 1000, what could lead to say a ten percent improvement in the overall upper bound $b(m,t)$? Thus for example, we kept track of the "declumping" factors of $(1-p)$ whenever they occur in a dominant term. In contrast, for the two sequence matching problem, the most careful treatment to date, [Waterman (1995)], gives away some of these factors. Even in an asymptotically non-dominant term, we didn't settle (see (68)) for the "easy bound" of the form $mt^3(\xi_*)^t$ on the net covariance contribution from two pairs of repeats where all four t-tuples form one island; here $\xi_* = \max_a(\mathbf{P}(A_i = a))$ is the probability of the most likely letter, and $m(\xi_*)^t$ is of the order m^{-1} in the i.i.d. uniform case. Instead we perform a careful combinatorial analysis to replace the t^3 factor by a constant $c = cd_\infty(\xi_*)$, with $c = 1.0445$ for the important special case of a uniform distribution over four letters. As a corollary of the careful Poisson process approximation, we obtain good bounds on approximations for the length of the longest repeat within a sequence, both allowing and forbidding self-overlap in the repeats.

A major issue for section 2 is how to handle self-overlapping repeats, which do play a role in sequencing by hybridization. The easy strategy, embodied in Corollary 1, is to bound the expected number of repeats involving self-overlap, and apply the second moment calculations of the Chen-Stein method only to repeats without self-overlap. The more difficult strategy is to apply Chen-Stein directly to all repeats. A priori, there is no way to predict which strategy will yield better bounds for realistic values of m and t; it is necessary to carry out both strategies. The last two colums of tables 1 and 2 show a comparison of the bounds; for example with $m = 50, t = 5$ the easy strategy is better, for $m = 200, t = 7$ it is about 3% worse, and for $m = 1600, t = 11$ the easy strategy gets a bound almost twice as large: .0035 versus .0018.

Section 3 shows how to adapt the Ukkonen-Pevzner characterization of unique recoverability, given in terms of where repeats occur, to the process needed for probability approximation, which only says where *leftmost* repeats occur. The three examples at the start of section 3 show that for each of the three classes of transformations considered by Ukkonen and Pevzner, there are nontrivial examples where the only repeats are self-overlapping.

Section 4 combines the Poisson process approximations of section 2 with the deterministic manipulations of section 3, and comes up with an overall error bound that we analyze as the sum of five contributions: one for Poisson approximation, one for each of the two classes of transformations that are unlikely, one for symmetrizing the Poisson intensities to remove the irregularity from the boundary and from self-overlap, and one for the tie breaking that is needed in the argument involving Catalan numbers.

We tabulate some values of the error bounds for realistic instances of l and m, for examples corresponding to DNA with either the four letters equally likely, or with p_A, p_C, p_G, p_T coming from the proportions of bases in the complete chloroplast genome of the liverwort Marchantia Polymorpha, with 42,896 A's, 17,309 C's, 17,556 G's, and 43,263 T's; see [Arratia, Gordon, Waterman (1990a)]. Tables 1, 2 and 3 discuss Poisson approximations. Tables 4 and 5 concern the distribution of the length of the longest repeat. Tables 6 and 7 analyze the five sources of error for the problem of sequencing by hybridization, and Tables 8 and 9 give the probability of unique recoverability, with error bounds and performance guarantees.

In the analysis of a DNA sequence for repeats, only rare repeats - those corresponding to small values of λ - are of interest. However, in SBH the goal is to gain information about the DNA sequence, even if it is incomplete information. An SBH experiment is worth running even if the probability of reading a unique sequence is 0.5. For this probability, the tables show values of λ as large as 2.747.

We now discuss the relation between our paper and [Dyer, Frieze, Suen (1994)]. They introduce the formula (1) and explain the connection between Catalan numbers and the probabilistically dominant Ukkonen-Pevzner transformation, transposition using an interleaved pair of repeats. Our theorem differs from theirs in that we give a bound on the error, which is important for applications, where m and l are only moderately large. Furthermore we had difficulty constructing a rigorous reading of [Dyer, Frieze, Suen (1994)].

Section 5 considers directions for future work. The Poisson process approximation for repeats that we use should be robust enough to help analyze more realistic questions. For one example, one may want to approximate the probability of being able to reconstruct a sequence from the set underlying the l-spectrum without knowing multiplicities, or in the presence of errors. For a second example, if the sequence isn't uniquely recoverable, what sort of information is given by the l-spectrum. What is the distribution of the lengths of the fragments that can be recovered, and what is the distribution of the number of sequences that share the same spectrum?

The formulas as used by the program DERIVE, to compute our tables, can be found at `http://www-hto.usc.edu/papers/abstracts/sbh.html`.

We recommend skipping past section 2 for the first reading of this paper.

2 Poisson approximations for repeats

Notation We write $f \asymp g$ to mean that the ratio f/g is bounded away from zero and infinity, and $f \sim g$ to mean $f/g \to 1$. In contrast, in heuristics we write $f \approx g$ to mean that f and g are approximately equal, with no specific requirement.

Throughout this paper we assume that the letters $A_1, A_2, \ldots$ are i.i.d. (independent and identically distributed) with

$$0 < \xi_a = \mathbf{P}(A_i = a) < 1 \tag{2}$$

for $a \in S$, a finite or countably infinite alphabet. The case $S = \{\text{A,C,G,T}\}$ of size $s = 4$ is our motivation. For $k = 2, 3, \ldots$ let

$$p \equiv p_2; \qquad p_k = \mathbf{P}(A_1 = A_2 = \cdots = A_k) = \sum_{a \in S} (\xi_a)^k. \tag{3}$$

The special case of a uniform distribution has $\xi_a = 1/s$ for all $a \in S$, so $p = 1/s$, and $p_k = s^{-(k-1)} = p^{k-1}$.

For the sake of expressing the growth rates of our error bounds as functions of m when $m, t \to \infty$ with $\lambda \asymp 1$, we define two parameters, γ and ϵ. As is proved below, γ and ϵ satisfy the inequalities

$$0 < \gamma \leq 1, \quad 0 < \epsilon \leq 1/3, \quad \gamma \leq 3\epsilon, \tag{4}$$

all with equality holding if and only if we have the uniform case.

Write $\xi_* \equiv \max_{a\in S} \xi_a$ for the maximal single letter probability. Since at least two different letters have positive probability, it follows that $(\xi_*)^2 < p$, so that γ defined by

$$(\xi_*)^2 = p^{1+\gamma} \tag{5}$$

satisfies $\gamma > 0$. For $r = 1, 2, 3, \dots,$

$$p_{r+1} \equiv \sum (\xi_a)^{r+1} \leq \sum (\xi_a)(\xi_*)^r = (\xi_*)^r, \tag{6}$$

with equality if and only if ξ is uniform. The case $r = 1$ shows that $\gamma \leq 1$, with $\gamma = 1$ only for the uniform case.

We define ϵ by

$$p_3 = p^{\frac{3}{2}(1+\epsilon)}. \tag{7}$$

Consider Hölder's inequality, as the statement that the l_q norm of a function is decreasing, strictly so if the function is nonzero in at least two points. Applied to ξ and $q = 2, 3, 4, \dots$ this implies that $p^{1/2} > (p_3)^{1/3} > (p_4)^{1/4} > \cdots$. This shows that $\epsilon > 0$. From the Cauchy-Schwarz inequality, applied to a random variable with $D = \xi_a$ on the event $\{A_1 = a\}$, we have $\mathbf{E}D = p$ and

$$p^2 = (\mathbf{E}D)^2 \leq \mathbf{E}D^2 = p_3, \tag{8}$$

hence $\epsilon \leq 1/3$. Furthermore, equality holds if and only if D is constant, i.e. the distribution of A_i is uniform. From $p_3 = \sum(\xi_a)^3 \leq \sum(\xi_*)(\xi_a)^2 = (\xi_*)\, p$, with equality only for the uniform case, it follows $\gamma \leq 3\epsilon$, with equality only for the uniform. In (50) we also show an additional inequality: $3\epsilon < 2\gamma$.

Given a sequence of letters $A_1A_2\cdots A_m$ and a test length t, we say that there is a repeat at (i, j) if $i < j$ and the t-tuples $A_{i+1}A_{i+2}\cdots A_{i+t}$ and $A_{j+1}A_{j+2}\cdots A_{j+t}$ following positions i and j are identical. [This choice, rather than $A_iA_{i+1}\cdots A_{i+t-1} = A_jA_{j+1}\cdots A_{j+t-1}$, turns out to be convenient. It is more suggestive to call our choice a "repeat following (i, j)" but we will usually use the simpler phrase "repeat at (i, j)"; it is after all a matter of taste, and not one of technical correctness, since (i, j) is a point in the plane and not *a place* in the sequence $A_1A_2\cdots$.] We will throughout assume that $2 \leq t$ to avoid trivialities, and $2t \leq m$ to avoid unnecessary complication in the expression for the expected number of repeats. We keep track of all repeats within $A_1A_2\cdots A_m$ by restricting to the index set I defined by

$$I \equiv I(m, t) \equiv \{\alpha = (i, j) : \ 0 \leq i < j \leq m - t\}. \tag{9}$$

The size of this index set is $|I| = \binom{m-t+1}{2} = (m-t+1)(m-t)/2$.

We define the indicator function that a repeat occurs at $\alpha = (i,j) \in I$ by

$$R_\alpha \equiv R_{i,j} \quad \equiv \quad \mathbf{1}(A_{i+1} \cdots A_{i+t} = A_{j+1} \cdots A_{j+t}) \tag{10}$$

Our notation is $\mathbf{1}(C)$ for the indicator function of an event C, i.e. the random variable with values $\mathbf{1}(C) = 1$, if C occurs, and $\mathbf{1}(C) = 0$ otherwise. We work with indicators because their sum,

$$N \equiv \sum_{\alpha \in I} R_\alpha \tag{11}$$

counts the number of repeats.

There is a general phenomenon of clumping that may occur in Poisson approximation, described for instance in [Aldous (1989)] and [Arratia, Tavaré (1993)]. Here, repeats come in clumps; for instance with $t = 3, m = 16$, $A_1 \cdots A_m$ =CTATA ATGGT ATAAT C, which has TATAAT $= A_2 \cdots A_7 = A_{10} \cdots A_{15}$, we say there are repeats (of 3-tuples) following $(1,9), (2,10), (3,11)$, and $(4,12)$. Counting all repeats, the result would be that the distribution of the number N is not approximately Poisson but rather compound Poisson. More importantly, the process $(R_\alpha)_{\alpha \in I}$ cannot be approximated by any process having independent coordinates; see e.g. section 4.2.1 in [Arratia, Goldstein, Gordon (1990b)]. For many purposes, including the analysis of unique recoverability, it is enough to count clumps of repeats. There are many ways to give a precise definition for clumps; we choose one of these, which puts clumps in one to one correspondence with "leftmost" repeats, and makes it easy to establish a Poisson process approximation. [Another workable strategy is to identify clumps of repeats with "maximal repeats of length $\geq t$"; this was used in [Dyer, Frieze, Suen (1994)], and a Poisson process in this context can again be established using the Chen-Stein method, as in section 4.2.1 of [Arratia, Goldstein, Gordon (1990b)]. Using leftmost repeats (of length exactly t) is simpler than using maximal repeats (of length $\geq t$) for the purposes of Poisson process approximation.]

Formally, a repeat at (i,j) is leftmost if there is not also a repeat at $(i-1, j-1)$. Thus we define the indicator function that a leftmost repeat occurs at $\alpha = (i,j) \in I$ by

$$\begin{aligned} X_\alpha \equiv X_{i,j} \quad &\equiv \quad \mathbf{1}(A_1 \cdots A_t = A_{j+1} \cdots A_{j+t}) \text{ if } i = 0 \\ &\equiv \quad \mathbf{1}(A_i \neq A_j, A_{i+1} \cdots A_{i+t} = A_{j+1} \cdots A_{j+t}) \text{ if } i \neq 0. \end{aligned} \tag{12}$$

The sum of these indicators,

$$W \equiv W(m,t) \equiv \sum_{\alpha \in I} X_\alpha \tag{13}$$

counts the number of leftmost repeats. Note that since $X_{i,j} = R_{i,j}\mathbf{1}(i = 0 \text{ or } R_{i-1,j-1} = 0)$, the process $(X_\alpha, \alpha \in I)$ carries no additional information compared to the $(R_\alpha, \alpha \in I)$. There are examples, such as that in the remark following (95), to show that collectively the X_α carry strictly less information than the R_α. Nevertheless, as we show in section 3, the indicators X_α carry enough information to determine unique recoverability.

A repeat at (i, j) would naturally be called "self-overlapping" if and only if the two t-tuples, $A_{i+1}A_{i+2}\cdots A_{i+t}$ and $A_{j+1}A_{j+2}\cdots A_{j+t}$, share some common A_k, i.e. $|i-j| < t$. However, since our concern is leftmost repeats, which involves $(t+1)$-tuples, we will also classify the situation $j = i + t$ as having self-overlap. Thus we define the index set I^* for "non self-overlapping repeats" (of t-tuples, taken leftmost) by

$$I^* \equiv I^*(m,t) \equiv \{\alpha = (i,j) : \; 0 \le i < i+t < j \le m-t\}. \tag{14}$$

Note that $I^* \subset I$ and

$$|I^*| = \binom{m-2t+1}{2}. \tag{15}$$

The number of non self-overlapping leftmost repeats is defined to be

$$W^* \equiv W^*(m,t) \equiv \sum_{\alpha \in I^*} X_\alpha. \tag{16}$$

The process of indicators of leftmost repeats is

$$\mathbf{X} \equiv (X_\alpha)_{\alpha \in I}, \tag{17}$$

and the process of indicators of non self-overlapping leftmost repeats is

$$\mathbf{X}^* \equiv (X_\alpha)_{\alpha \in I^*}. \tag{18}$$

Note, these processes take values in $\{0,1\}^I$ and $\{0,1\}^{I^*}$ respectively. Compared with W and W^*, the total numbers of leftmost repeats, the processes give additional information: where these repeats occur.

2.1 Expected number

For non self-overlapping leftmost repeats, i.e. for $i + t < j$, the $2(t+1)$ indices into the sequence $A_{(\cdot)}$ come in $t+1$ disjoint pairs, and for $i \ne 0$ we have $X_{i,j} = \mathbf{1}(A_i \ne A_j)\mathbf{1}(A_{i+1} = A_{j+1})\cdots\mathbf{1}(A_{i+t} = A_{j+t})$, with the factors being indicators of independent events. Being careful with the special case $i=0$, where the factor $\mathbf{1}(A_i \neq A_j)$ is not present, we have the probability $\mathbf{E}X_\alpha$ of a leftmost repeat at $\alpha=(i,j)\in I^*$ given by

$$\begin{aligned} \mathbf{E}X_\alpha &= p^t \qquad ; i=0 \\ &= (1-p)p^t \quad ; i>0. \end{aligned} \tag{19}$$

Since I^* has $m-2t$ elements with $i=0$, (namely $(0,j)$ with $j=t+1$ to $m-t$), and $\binom{m-2t}{2}$ elements with $i>0$, the expected number of non self-overlapping leftmost repeats is

$$\lambda^* \equiv \lambda^*(m,t) \equiv \mathbf{E}W^* = \binom{m-2t}{2}(1-p)p^t + (m-2t)p^t = \binom{m-2t+1}{2}(1-p)p^t + (m-2t)p^{t+1}. \tag{20}$$

The second equality above can be derived by algebraic manipulation, or seen directly from the point of view that all points $(i,j)\in I^*$ have intensity $\mathbf{E}X_{i,j}=p^t-p^{t+1}$ or more, and that the exceptional case with $i=0$ has an extra p^{t+1} of intensity.

For m,t both large, the interesting case for distributional approximations is that λ^* is bounded away from zero and infinity, and it is fairly easy to see from (20) that this occurs if and only if the difference between t and $2\log_{1/p} m$ is bounded, i.e.

$$\lambda^* \asymp 1 \text{ if and only if } t - 2\log_{1/p} m = O(1) \text{ as } m,t\to\infty. \tag{21}$$

It is also easy to see that

$$\lambda^* - \frac{m^2}{2}(1-p)p^t \to 0 \tag{22}$$

whenever $m,t\to\infty$ with λ^* bounded away from zero and infinity. These qualitative relations (21) and (22) will also be true for repeats allowing self-overlap, i.e. for $\lambda\equiv\mathbf{E}W$ replacing λ^*. The remainder of this subsection gives an exact formula for λ and proves these two qualitative relations. This exact formula for λ is

$$\lambda \equiv \lambda(m,t) \equiv \mathbf{E}W = \lambda^* + \sum_{d=1}^{t}(p_{q+1})^r(p_q)^{d-r} + \sum_{d=1}^{t}(m-t-d)(p_q-p_{q+1})(p_{q+1})^r(p_q)^{d-r-1}. \tag{23}$$

Allowing self-overlap, the simplest case is $\alpha=(0,1)$, with $X_\alpha \equiv \mathbf{1}(A_1A_2\cdots A_t = A_2A_3\cdots A_{t+1}) = \mathbf{1}(A_1=A_2=A_3=\cdots=A_{t+1})$, so that $\mathbf{E}X_\alpha = p_{t+1}$. The next simplest case is $\alpha=(0,2)$ and t even, with $X_\alpha\equiv\mathbf{1}(A_1A_2A_3\cdots A_t = A_3A_4A_5\cdots A_{t+2}) = \mathbf{1}(A_1=A_3=\cdots=A_{t+1})\mathbf{1}(A_2=A_4=\cdots=A_{t+2})$, so that $\mathbf{E}X_\alpha=(p_{(t+2)/2})^2$. For the same $\alpha=(0,2)$ but t odd, $X_\alpha=\mathbf{1}(A_1=A_3=\cdots=A_{t+2})\,\mathbf{1}(A_2=A_4=\cdots=A_{t+1})$, so that $\mathbf{E}X_\alpha = p_{(t+1)/2}p_{(t+3)/2}$.

The general self-overlap situation has $j = i + d$, for d in the range 1 to t. The overlapping matching $A_{i+1} \cdots A_{i+t} = A_{i+d+1} \cdots A_{i+d+t}$ forces periodicity with period d; specifically the word $A_{i+1} \cdots A_{i+t+d}$ of length $t + d$ is a d-tuple, repeated over and over $(t + d)/d = q + (r/d)$ times. In detail, divide d into $t + d$ to get quotient q and remainder r, so that

$$\alpha = (i, j), \quad j = i + d, \ 1 \le d \le t, \ t + d = dq + r, \ 0 \le r < d. \tag{24}$$

First consider the indicator $R_{i,j} \equiv \mathbf{1}(A_{i+1} \cdots A_{i+t} = A_{j+1} \cdots A_{j+t})$, so that there is no declumping factor $\mathbf{1}(A_i \ne A_j)$ to complicate things. The indicator R_α with α satisfying (24) involves $t+d$ letters, and $t+d = dq+r = (d-r)q+r(q+1)$, so it is plausible that the matches break up d disjoint groups of letters, corresponding to d independent events, with r groups of $q + 1$ letters and $d - r$ groups of q letters. In fact, the r factors of R_α having probability p_{q+1} each are $\mathbf{1}(A_{i+1} = A_{i+d+1} = A_{i+2d+1} = \cdots = A_{i+qd+1})$ through $\mathbf{1}(A_{i+r} = A_{i+d+r} = A_{i+2d+r} = \cdots = A_{i+qd+r})$, and the remaining $d - r$ factors, having probability p_q each, are $\mathbf{1}(A_{i+r+1} = A_{i+r+d+1} = \cdots = A_{i+r+(q-1)d+1})$ through $\mathbf{1}(A_{i+d} = A_{i+2d} = \cdots = A_{i+qd})$. This proves that for α satisfying (24),

$$\mathbf{E}(R_\alpha) = (p_{q+1})^r (p_q)^{d-r}. \tag{25}$$

[Check that the special case $i = 0, j = t$, which isn't really self-overlapping, but is included in the above discussion, reduces correctly, with $q = 2, r = 0$, to $\mathbf{E}X_\alpha = (p_{q+1})^r (p_q)^{d-r} = p_3^0 p_2^d = p^d = p^t$.]

Now consider X_α, for $i > 0$, so that compared with the analysis of R_α in the previous paragraph, there is also a declumping factor $\mathbf{1}(A_i \ne A_j)$. The effect is to change the last of the indicators involving q letters, $\mathbf{1}(A_{i+d} = A_{i+2d} = \cdots = A_{i+qd})$, to an indicator involving $q + 1$ letters, namely $\mathbf{1}(A_i \ne A_{i+d} = A_{i+2d} = \cdots = A_{i+qd})$. The expectation of this last indicator is $p_q - p_{q+1}$. [To check this claim note that the event $\{A_2 = A_3 = \cdots = A_k\}$ is the disjoint union of the events $\{A_1 = A_2 = \cdots = A_k\}$ and $\{A_1 \ne A_2 = A_3 = \cdots = A_k\}$, so $p_{k-1} = p_k + \mathbf{P}(A_1 \ne A_2 = A_3 = \cdots = A_k)$.] Thus for the general self-overlapping index $\alpha = (i, j) \in I \setminus I^*$, using (24) to define d, q, r as functions of i, j, we have

$$\begin{aligned} \mathbf{E}X_\alpha &= (p_{q+1})^r (p_q)^{d-r} & ; i = 0 \\ &= (p_q - p_{q+1})(p_{q+1})^r (p_q)^{d-r-1} & ; i > 0. \end{aligned} \tag{26}$$

As a check, we observe that in the uniform case the above simplifies to $\mathbf{E}X_\alpha = p^t$ if $i = 0$ and $\mathbf{E}X_\alpha = (1 - p)p^t$ if $i > 0$, the same as in the non self-overlapping case.

The expected number λ of leftmost repeats is the sum of $\mathbf{E}X_\alpha$ for α with and without self-overlap. The terms without self-overlap have a net contribution λ^*. For the α with self-overlap, for each of $d = 1, 2, \ldots, t$, there is exactly one term with $i = 0$ and there are $m - t - d$ terms with $i > 0$, namely $i = 1, 2, \ldots, m - t - d$. Our restriction $2t \le m$ is to assure $m - t - d \ge 0$ so that truncation is not needed in the expression for λ below. We have shown that with $\lambda^* = (m - 2t)p^t + \binom{m-2t}{2}(1-p)p^t$, as given by formula (20), the formula (23) is valid.

Since for applications the value of t might be 7,8,9,10, or 11, the above expression, with 2 terms for λ^* and then $2t$ additional terms, is tractable. For simplicity of understanding, it is worth having a simple upper bound on $\lambda - \lambda^*$, the expected number of leftmost self-overlapping repeats. We get such a bound in (27) below. To motivate the bound, we observe that for $\alpha = (i, i+d), \mathbf{E}X_\alpha$ is nonincreasing as d increases from 1 to t; this holds both for $i = 0$, where there is no declumping factor, and for $i > 0$, where there is. [We do not present the proof of this.] Having identified that the "worst case" is $\alpha = (0, 1)$ with $\mathbf{E}X_\alpha = p_{t+1}$, we content ourselves with the easily proved bound that $\mathbf{E}X_\alpha \le (\xi_*)^t$, regardless of the amount of self-overlap. To prove this, we use (25), together with the bounds $p_{q+1} \le (\xi_*)^q$ and $p_q \le (\xi_*)^{q-1}$. The resulting power of ξ_* simplifies as $qr + (q-1)(d-r) = qd + r - d = t + d - d = t$. Thus, for α satisfying (24),

$$\mathbf{E}X_\alpha \le \mathbf{E}R_\alpha = (p_{q+1})^r (p_q)^{d-r} \le ((\xi_*)^q)^r ((\xi_*)^{q-1})^{d-r} = (\xi_*)^t. \tag{27}$$

The net result is

$$\lambda - \lambda^* \le mt(\xi_*)^t. \tag{28}$$

Note that this is not of the form $\lambda - \lambda^* \asymp \cdots$; we do not know such an expression.

A simplified form of (28), with γ defined by (5), is that $\lambda - \lambda^* = O(m^{-\gamma} \log m)$, since

$$\lambda - \lambda^* \le mt(\xi_*)^t \asymp m^{-\gamma} \log m. \tag{29}$$

The asymptotic bound in the last line is valid uniformly in $t, m \to \infty$ with λ^* bounded away from zero and infinity. To check this we write a series of equivalent statements that two functions have the same asymptotic order of magnitude: $\lambda^* \asymp 1$, $m^2 p^t \asymp 1$, $p^{t/2} \asymp m^{-1}$, $p^{(1+\gamma)t/2} \asymp m^{-1-\gamma}$, $mtp^{(1+\gamma)t/2} \asymp m^{-\gamma} t$. Finally, recall that $\lambda^* \asymp 1$ implies that $t - 2\log_{1/p} m = O(1)$, which in turn implies that $t \asymp \log m$.

The bound (29) shows that $\lambda - \lambda^* \to 0$ when $m, t \to \infty$ with $\lambda^* \asymp 1$. A corollary is that for $m, t \to \infty$, $\lambda \asymp 1$ if and only if $t - 2\log_{1/p} m = O(1)$, and that if $\lambda \asymp 1$ then

$$\lambda - (m^2/2)(1-p)p^t \to 0. \tag{30}$$

2.2 Review of total variation approximations

Our analysis of unique recoverability is based on the process $\mathbf{X}$ of indicators of leftmost repeats. This process has a complicated dependence structure, but the dependencies are weak and have only a small influence on the probability of unique recoverability. To make this rigorous, we compare $\mathbf{X}$ to a "nearby" process $\mathbf{Y}$ having independent coordinates, and the same marginal intensities. The notion "nearby" is quantified by the total variation distance, as follows.

For any two random processes $\mathbf{X}, \mathbf{Y}$ both with values in the same space T, the total variation distance is defined by

$$d_{TV}(\mathbf{X}, \mathbf{Y}) = \sup |\mathbf{P}(\mathbf{X} \in B) - \mathbf{P}(\mathbf{Y} \in B)|,$$

where the supremum is taken over all (measurable) subsets $B \subset T$. One consequence, which we apply in section 4, is that for any indicator of an event, i.e. a measurable functional h from T to $\{0,1\}$, there is an error bound of the form $|\mathbf{E}h(\mathbf{X}) - \mathbf{E}h(\mathbf{Y})| \leq d_{TV}$. [Another consequence is that for a functional $g : T \to T'$, the random elements $\mathbf{X}' = g(\mathbf{X})$ and $\mathbf{Y}' = g(\mathbf{Y})$ with values in T' are no further apart: $d_{TV}(\mathbf{X}', \mathbf{Y}') \leq d_{TV}(\mathbf{X}, \mathbf{Y})$. This is useful in comparing the two conclusions of Theorem 1 below, where the functional g is "summing the coordinates," and the images $\mathbf{X}'$ and $\mathbf{Y}'$ are called W and K. The random variable bound (32), without the "magic" factor $(1 - e^{-\lambda})/\lambda$, would simply be a corollary of its process bound (31).]

There following process approximation theorem first appears, with an extra factor of 2 in the upper bound, in [Arratia, Goldstein, Gordon (1989)]. The bound (32) originates [Chen (1975)], using Stein's method. A friendly discussion of the Chen-Stein method and its application to sequence matching is [Arratia, Goldstein, Gordon (1990b)]. The book [Barbour, Holst, Janson (1992)] presents much more, including the improvement by a factor of 2; see a related book review [Arratia, Tavaré (1993)].

Theorem 1 *Suppose $\mathbf{X} \equiv (X_\alpha)_{\alpha \in I}$ is a process of indicator random variables with $\mathbf{E}X_\alpha = \mathbf{P}(X_\alpha = 1) = 1 - \mathbf{P}(X_\alpha = 0)$. Let $\mathbf{Y} \equiv (Y_\alpha)_{\alpha \in I}$ be a process with independent Poisson distributed coordinates Y_α, with $\mathbf{E}Y_\alpha = \mathbf{E}X_\alpha$. Suppose for each $\alpha \in I$ there is a $B_\alpha \subset I$ such that X_α is independent of the sigma-algebra generated by all $X_\beta, \beta \in I \setminus B_\alpha$. Let*

$$b_1 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} \mathbf{E}X_\alpha \mathbf{E}X_\beta, \qquad b_2 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha, \beta \neq \alpha} \mathbf{E}(X_\alpha X_\beta).$$

Then

$$d_{TV}(\mathbf{X},\mathbf{Y}) \le (b_1+b_2). \tag{31}$$

Let $W=\sum_{\alpha\in I} X_\alpha,\quad \lambda=\sum_{\alpha\in I}\mathbf{E}X_\alpha,\quad K=\sum_{\alpha\in I} Y_\alpha$. *Assume* $0<\lambda<\infty$. *[It follows that* K *is a Poisson random variable and* $\mathbf{E}W=\mathbf{E}K=\lambda$.*] Then*

$$d_{TV}(W,K) \le \frac{1-e^{-\lambda}}{\lambda}(b_1+b_2). \tag{32}$$

and in particular

$$|\mathbf{P}(W=0)-e^{-\lambda}| \le \frac{1-e^{-\lambda}}{\lambda}(b_1+b_2). \tag{33}$$

We will apply this theorem in two situations, in the next two sections. In both cases, the indicator variables X_α are (among) those defined by (12). The only difference is in the index sets playing the role of I for Theorem 1; I^* defined by (14) for the process of non self-overlapping leftmost repeats, and I defined by (9) for the process of all leftmost repeats. Recall that $I^* \subset I$. In both cases the neighborhoods B_α we choose are defined via the symmetric "overlap" relation: for $\mathbf{X}^*$ we will use $B_\alpha \equiv \{\beta\in I^* : \alpha,\beta \text{ overlap each other }\}$, and for $\mathbf{X}$ we will use $B_\alpha \equiv \{\beta\in I : \alpha,\beta \text{ overlap each other }\}$. Formally, the overlap relation is given by: for $\alpha=(i,j), \beta=(i',j')\in I$

$$\alpha\sim\beta \text{ if and only if } \min(|i-i'|,|i-j'|,|j-i'|,|j-j'|)\le t. \tag{34}$$

The motivation is that the indicator X_α involves the set of positions $\{i,i+1,\ldots,i+t\}\cup\{j,j+1,\ldots,j+t\}$, X_β corresponds to another set of positions; we defined $\alpha\sim\beta$ to be the condition that these sets overlap. Thus it is easy to check that if α and β do not overlap, written $\alpha\not\sim\beta$, then X_α and X_β are independent. There is a subtlety: it is also necessary (and easy in this case) to check that X_α is independent of the sigma-algebra generated by $(X_\beta, \beta\not\sim\alpha)$.

[For a concrete example of this subtlety in operation, consider defining $B_\alpha \equiv \{\beta\in I : \alpha\sim\beta\}$ as before, but changing the neighbor relation $\sim$ on I to $\alpha\not\sim\beta$ if and only if X_α and X_β are independent. Take the special case where the A_i are uniformly distributed. This makes the new B_α much smaller, which would yield better bounds b_1 and b_2. But X_α is not independent from $\sigma(X_\beta,\beta\notin B_\alpha)$, so Theorem 1 does not apply to this choice of the B_α. [Actually, the full version of the theorem, in [Arratia, Goldstein, Gordon (1989)], does apply; there is an error term b_3 to control the departure from independence.] To check this failure of independence, consider the alphabet $S=\{\text{A,T}\}$, and take $t=2, m=11, \alpha=(1,4), \gamma=$

$(1,6), \gamma' = (1,7), \delta = (4,8), \delta' = (4,9)$. It is easy to check that X_α is independent of each of the variables $X_\gamma, X_{\gamma'}, X_\delta$ and $X_{\delta'}$. But $X_\gamma X_{\gamma'} = 1$ implies $A_2 = A_7 \neq A_1$, and $X_\delta X_{\delta'} = 1$ implies $A_5 = A_9 \neq A_4$. Thus $X_\gamma X_{\gamma'} X_\delta X_{\delta'} = 1$ implies $A_1 \neq A_2, A_4 \neq A_5$, which implies $X_\alpha = 0$. The event $X_\alpha = 1$ has positive probability [1/8]. The event $X_\gamma X_{\gamma'} X_\delta X_{\delta'} = 1$ has positive probability [in fact, 2^{-10}]; we could have $A_1 \cdots A_m =$ TAATA AAAAA A or ATTAT TTTTT T. Hence X_α is not independent of $\sigma(X_\gamma, X_{\gamma'}, X_\delta, X_{\delta'})$. This phenomenon is also true for alphabets of larger size s; it can be seen that the conditional probability $\{\mathbf{P}(X_\alpha = 1 | X_\gamma X_{\gamma'} X_\delta X_{\delta'} = 1) = \frac{s-2}{s-1} \neq \mathbf{P}(X_\alpha = 1) = (1-\frac{1}{s})s^{-2}$, so that X_α is not independent of $\sigma(X_\gamma, X_{\gamma'}, X_\delta, X_{\delta'})$.]

2.3 The uniform case

When the i.i.d. letters $A_1, A_2, \ldots$ are uniformly distributed over a finite alphabet, rather than having a general distribution, many of the quantities we analyze, including expectations and covariances for the indicators X_α of leftmost repeats, are much simpler. In this section we collect the results of all such computations for this uniform special case. Some proofs involve complicated arguments about cycles in the graph of matching edges. In this section, we only quote the results as they simplify for the uniform case; the proofs are saved for the sections on the general, not necessarily uniform case.

For the uniform case, with an alphabet of size s, we have $p \equiv \mathbf{P}(A_1 = A_2) = 1/s$. More generally for $r = 2, 3, \ldots, p_r \equiv \mathbf{P}(A_1 = A_2 = \cdots = A_r) = s^{1-r} = p^{r-1}$.

In the uniform case the expected number λ of leftmost repeats and the expected number λ^* of leftmost non self-overlapping repeats are

$$\begin{aligned} \lambda &= \binom{m-t+1}{2}(1-p)p^t + (m-t)p^{t+1} \qquad (35) \\ \lambda^* &= \binom{m-2t+1}{2}(1-p)p^t + (m-2t)p^{t+1}. \end{aligned}$$

For λ, the expression above can easily be derived using $|I| = \binom{m-t+1}{2}$, noting that for $\alpha \in I$, regardless of self-overlap, $\mathbf{E}X_\alpha \geq (1-p)p^t$. Equality holds except for $m-t$ cases $\alpha = (0,j)$, which have $\mathbf{E}X_\alpha = p^t = p^{t+1} + (1-p)p^t$. It requires some work to check that (23), the expression for λ in the general case, simplifies to the same. The general expression for λ^* is given by (20), which is identical to the uniform special case above.

Next we look at the bounds for the Chen-Stein method, Theorem 1. The net result [for both cases, not allowing or allowing self-overlap,] will be an upper bound on the total variation distance between the process marking repeats and a process with independent coordinates, such that the bound is order of $\log m/m$ when $m, t \to \infty$ with $\lambda \asymp 1$.

Remark The upper bound may not be sharp; the factor of $\log m$ may just be an artifice of our method. The best lower bound we can find is order of $1/m$, and comes from looking at configurations of the form $X_\alpha = X_\beta = X_\gamma = 1$ where $\alpha = (i,j), \beta = (i,k), \gamma = (j,k)$.

Next we consider the case of non self-overlapping repeats; the bounds for the Chen-Stein method are denoted b_1^* and b_2^*, with the asterisk to denote that self- overlap is excluded. The exact value for b_1^* is given by (42), with asymptotics $b_1^* \sim (8t/m)(\lambda^*)^2$ from (43); the same expressions are used for both the uniform and nonuniform cases. For b_2^* in the uniform case we use an upper bound $\overline{b_2^*}$ which can be defined simply by $\overline{b_2^*} \equiv b_1^*$. The inequality

$$b_2^* \leq b_1^* \tag{36}$$

needs to be justified; it does not hold in the nonuniform case. Here is an outline of that justification. For the "parallel overlapping" case $(i',j') = (i+d, j+d)$ with $0 \neq |d| \leq t$ we have $X_\alpha X_\beta \equiv 0$ thanks to the declumping factors. Excluding the parallel overlapping case, uniformity leads to independence: for $\alpha \neq \beta$, $\mathbf{E}(X_\alpha X_\beta) = \mathbf{E}X_\alpha \mathbf{E}X_\beta$. The argument to see this involves a graph; one needs to check that the matches and mismatches required by $X_\alpha X_\beta = 1$ form no cycles. Except for the declumping factors, we give the details of this in the discussion leading up to equation (46). That equation reduces in the uniform case, via $p_{r+1} = p^r$, to $\mathbf{E}(R_\alpha R_\beta) = p^{e(1)} p_3^{e(2)} \cdots = p^{e(1)+2e(2)+3e(3)+\cdots} = p^{2t} = \mathbf{E}R_\alpha \mathbf{E}R_\beta$. To handle declumping factors, if one or two are present, we argue that the graph $S_{\alpha\beta}$, augmented by edges corresponding to mismatches, still has no cycles. Start with the argument to show that $S_{\alpha\beta}$ has no cycles, and augment S_α or S_β or both. Effectively, we need to use $t+1$ instead of t in one or two places, but the geometry remains the same. The net result is that for $\alpha, \beta \in I^*$ with $\alpha \neq \beta$, excluding the parallel overlapping case,

$$\mathbf{E}(X_\alpha X_\beta) = \mathbf{E}X_\alpha \mathbf{E}X_\beta. \tag{37}$$

In summary, for the uniform case $b_2^* < b_1^*$, and the only differences between b_1^* and b_2^* are that b_2^* excludes the terms with $\alpha = \beta$, and for b_2^*, the parallel overlapping terms are zero.

Now we look at the case of repeats, allowing self-overlap; the bounds for the Chen-Stein method are denoted b_1 and b_2. An exact expression for b_1 would be exceedingly complicated, but it is easy to give an upper bound $\overline{b_1}$ and to show that $b_1 \sim \overline{b_1}$ (for $m, t \to \infty$ with $\lambda \asymp 1$). To derive this, we need an upper bound on $\sum_{\beta:\alpha\sim\beta} \mathbf{E}X_\beta$. First note that the number of terms in this sum is at most $(4t+2)(m-t)$. [Given $\alpha = (i,j)$, there are at most $4t+2$

choices for an integer x with $|x-i| \le t$ or $|x-j| \le t$. There are $m-t$ choices for $y \neq x$ with $0 \le y \le m-t$. Now take $\beta = (i', j')$ with $i' = \min(x,y), j' = \max(x,y)$; this accounts for all possible β as well as some extraneous values.] The value of $\mathbf{E}X_\beta$ is either $(1-p)p^t$ or p^t, and (disregarding the requirement that $\alpha \sim \beta$) there are exactly $m-t$ choices β with $\mathbf{E}X_\beta = p^t = (1-p)p^t + p^{t+1}$. Thus for any α

$$\sum_{\beta:\alpha\sim\beta} \mathbf{E}X_\beta \le (4t+2)(m-t)(1-p)p^t + (m-t)p^{t+1},$$

and hence for the uniform case we can define an upper bound $\overline{b_1}$ by

$$\begin{aligned} b_1 &\le \sum_{\alpha\in I} \mathbf{E}X_\alpha \sum_{\beta\in I:\alpha\sim\beta} \mathbf{E}X_\beta \\ &\le \sum_{\alpha\in I} \mathbf{E}X_\alpha \left[(4t+2)(m-t)(1-p)p^t + (m-t)p^{t+1}\right] \\ &= \lambda\left[(4t+2)(m-t)(1-p)p^t + (m-t)p^{t+1}\right] \\ &\equiv \overline{b_1}. \end{aligned} \tag{38}$$

For asymptotics, when $m, t \to \infty$ with $\lambda \asymp 1$, we have $\lambda \sim (m^2/2)(1-p)p^t$ and $(m-t) \sim m$ so $\overline{b_1} \sim (8t/m)\lambda^2$. That

$$b_1 \sim \overline{b_1} \sim (8t/m)\lambda^2 \tag{39}$$

now follows from $b_1^* \sim (8t/m)(\lambda^*)^2$ from (43), together with $b_1^* \le b_1 \le \overline{b_1}$ and the fact that $\lambda^* \sim \lambda$, which is easily seen from (35). We get the same asymptotics in (64) for the general case, using a more complicated argument.

For b_2 it is essential to notice that even in the uniform case, there are situations where X_α and X_β are highly positively correlated. The extreme example is $\alpha = (0,1), \beta = (0,2)$, so that $X_\alpha X_\beta = \mathbf{1}(A_1 = A_2 = \cdots = A_{t+2})$ with $\mathbf{E}(X_\alpha X_\beta) = p^{t+1} \gg p^{2t} = \mathbf{E}X_\alpha \mathbf{E}X_\beta$. These situations are collected under "case one" in the general analysis of b_2, and the upper bound (78) applies here with $\xi_* = 1/s = p$. For the other situations included in b_2 but not already part of b_2^*, which form "case two", we show that the graph $S_{\alpha\beta}$ has no cycles, see (65). As in the argument leading up to (37), putting in the declumping factors doesn't cause cycles, and in case two $\mathbf{E}(X_\alpha X_\beta) = \mathbf{E}X_\alpha \mathbf{E}X_\beta$.

[Uniformity is needed to simplify the following calculation: for any tree with $r+k$ edges corresponding to $r+k+1$ letters, with r edges requiring matches and k edges requiring mismatches, regardless of where the mismatch edges appear in the tree, the probability of the corresponding event is $(1-p)^k p^r$. Here is an example of how the probability corresponding to a tree with r matching edges and k mismatching edges doesn't vary with the tree, but only for the uniform case. First consider $t = 2, \alpha = (1,5), \beta = (2,5)$, which contributes to b_2^*. Here $k = 2, r = 4$ and $\mathbf{E}(X_\alpha X_\beta) = \mathbf{P}(A_1 \neq A_5 \neq A_2 = A_6 = A_3 = A_7 = A_4) = (1-p)p_5$. In contrast consider $t = 2, \alpha = (1,2), \beta = (4,5)$, which contributes to case two. Again $k = 2, r = 4$, but now $\mathbf{E}(X_\alpha X_\beta) = \mathbf{P}(A_1 \neq A_2 = A_3 = A_4 \neq A_5 = A_6 = A_7) = (p_3 - p_4)p_3$. In general these are not equal, but in the uniform case they are.]

In summary, for the uniform case we have that b_2 is at most b_1 plus the contribution (78) from the terms in case one, so we define an upper bound $\overline{b_2}$ for b_2 by

$$b_2 \leq \overline{b_2} \equiv \overline{b_1} + (1 + (1-p)(m-t))p^t c d_\infty(p). \tag{40}$$

For alphabets of size 4,3, or 2, the relevant upper bounds for the uniform case are $cd_\infty(1/4) < 1.0445, cd_\infty(1/3) < 1.981, cd_\infty(1/2) < 22.09$.

2.4 Repeats, not allowing self-overlap

We apply Theorem 1 to the process $\mathbf{X}^*$ defined in (18), with $B_\alpha = \{\beta \in I^* : \ \beta \sim \alpha\}$, the relation "$\sim$" being defined by (34). To distinguish the Chen-Stein error bounds b_1 and b_2 in this case from those in the next section, we call them b_1^* and b_2^*. Hence we have the same random variables, but different index sets: $b_1^* = \sum_{\alpha,\beta \in I^*:\ \alpha\sim\beta} \mathbf{E}X_\alpha \mathbf{E}X_\beta, b_2^* = \sum_{\alpha,\beta\in I^*:\ \alpha\sim\beta,\alpha\neq\beta} \mathbf{E}(X_\alpha X_\beta)$, while $b_1 = \sum_{\alpha,\beta\in I:\ \alpha\sim\beta} \mathbf{E}X_\alpha \mathbf{E}X_\beta, b_2 = \sum_{\alpha,\beta\in I:\ \alpha\sim\beta,\alpha\neq\beta} \mathbf{E}(X_\alpha X_\beta)$.

We will give an exact expression for b_1^*, and an upper bound $\overline{b_2^*}$ for b_2^*. We begin with b_1^*. The intensity function $\mathbf{E}X_\alpha$, given by (19), is constant except for the boundary effect at $i = 0$. Hence the chief problem is to determine the size of the neighbor relation, i.e. the number of ordered pairs of neighbors, i.e. $|G|$ where

$$G = \{(\alpha,\beta) : \ \alpha, \beta \in I^*, \alpha \sim \beta\}.$$

It is then easy to make a correction for the boundary effects. To get a handle on $|G|$, consider the complementary relation on I^*, namely

$$H = \{(\alpha,\beta) : \ \alpha, \beta \in I^*, \alpha \not\sim \beta\}.$$

Writing $\alpha = (i,j), \beta = (i',j')$, the map $(\alpha,\beta) \mapsto \{i,i',j,j'\}$ is a $\binom{4}{2} = 6$ to one correspondence from H onto the set J of sets of four points all more than t apart:

$$J = \{C \subset \{0,1,\ldots,m-t\} : |C| = 4, a \neq b \in C \text{ implies } |a-b| > t\}.$$

To see that the correspondence is $\binom{4}{2}$ to one, note that a priori $i < j$ and $i' < j'$, so picking a set of two of the four elements of a C to serve as $\{i,j\}$ determines (α,β). It is elementary that $|J| = \binom{(m-t+1)-3t}{4}$. To see this, write $C = \{k_1, k_2, k_3, k_4\}$ with $k_1 < k_2 < \cdots$, and let $j_1 = k_1, j_2 = k_2 - t, j_3 = k_3 - 2t, j_4 = k_4 - 3t$. This gives a set of four distinct elements $\{j_1, \ldots, j_4\} \subset \{0, 1, \ldots, (m-t) - 3t\}$ with $j_1 < \cdots < j_4$, and it is easy to check that this is a one to one correspondence between J and the set of all four-subsets of $\{0, 1, \ldots, m-t-3t\}$. We have shown that $|H| = 6\binom{m-4t+1}{4}$, so in terms of $|I^*|$ given by (15), the number of ordered pairs of neighbors, as a function of m and t, is

$$c(m,t) \equiv |G| = (|I^*|)^2 - |H| = \binom{m-2t+1}{2}^2 - 6\binom{m-4t+1}{4} = 2m^3t + m^3 - 18m^2t^2 + \cdots. \tag{41}$$

Without additional work, the number of ordered pairs of neighbors in which both $i, i' \neq 0$ is $c(m-1, t)$, because eliminating 0 from $\{0, 1, 2, \ldots, m-t\}$ has the same effect as eliminating $m-t$, or equivalently, reducing m by one. Since there are $m-2t$ choices for $\alpha = (i,j) \in I^*$ with $i = 0$, and any two of these are neighbors of each other, there are exactly $(m-2t)^2$ ordered pairs of neighbors in which both have first coordinate zero. [The following argument is summation by parts, and is easily understood by comparison to the second part of (20).] Since the $\mathbf{E}X_\alpha \mathbf{E}X_\beta = (1-p)^2 p^{2t}$ when neither first coordinate is zero, and the increments are $(1-p)p^{2t+1}$ and p^{2t+2} successively when one and then the other first coordinate becomes zero, we have, with $c(m,t)$ given by (41),

$$b_1^* = c(m,t)(1-p)^2 p^{2t} + (c(m,t) - c(m-1,t))(1-p)p^{2t+1} + (m-2t)^2 p^{2t+2}. \tag{42}$$

For asymptotics, $c(m,t) \sim 2m^3 t$ for $m, t \to \infty$ with $t/m \to 0$, so for $m, t \to \infty$ with $\lambda^* \asymp 1$ we have

$$b_1^* \sim 2m^3 t(1-p)^2 p^{2t} = (8t/m)((1-p)p^t m^2/2)^2 \sim (8t/m)(\lambda^*)^2. \tag{43}$$

Although it is possible to give an exact expression for b_2, both the expression and derivation would be exceedingly complicated, so we just give an upper bound, which we will denote $\overline{b_2^*}$.

We begin with a sketch of the analysis. Recall the definition (34) of the overlap relation. We say that the "degree" of overlap is the number of inequalities $|i - i'| \leq t, |i - j'| \leq t, |j - i'| \leq t, |j - j'| \leq t$ which are satisfied. Here the possible degrees of overlap for a pair $\alpha \sim \beta$ are $d = 1, 2, 3$, thanks to the restriction of no self-overlap in α, β individually. As a guide to which pairs (α, β) require careful bounding of $\mathbf{E}(X_\alpha X_\beta)$, we observe without proof that there are on the order of $m^{4-d}t^d$ pairs (α, β) having overlap of degree $d, d = 1, 2$, or 3. The dominant contribution to b_2^* turns out to be from overlap of degree one.

We bound the number of pairs (α, β) having overlap of degree two or more, as follows. There are $\binom{4}{2} = 6$ ways to specify a set of two out of the four inequalities, but one of these, namely $|i-j'| \leq t, |j-i'| \leq t$, cannot occur, thanks to α and β not having self-overlap. In each of the remaining cases, we can designate one of α, β which is "tied down" in both its components. [For example, the case $|i-i'| \leq t, |i-j'| \leq t$ ties down both components of β (and not both components of α); the case $|i-j'| \leq t, |j-j'| \leq t$ ties down both components of α (and not both components of β); for the case $|i-i'| \leq t, |j-j'| \leq t$ both α, β have both components tied down, and our canonical choice is to designate α.] For the element of I^* which is not designated as tied down, there are at most $|I^*| = \binom{m-2t+1}{2}$ choices. Then for the other element of I^* there are at most $(2t+1)^2$ choices. Combining these, we have at most

$$5\binom{m-2t+1}{2}(2t+1)^2 \sim 10m^2t^2 \tag{44}$$

choices for $(\alpha, \beta) \in (I^*)^2$ with overlap of degree two or more.

We will now establish an upper bound on $\mathbf{E}(X_\alpha X_\beta)$, valid for all cases. We identify an index $\alpha = (i,j)$ with the set of t undirected edges

$$S_\alpha = \{\{i+1, j+1\}, \{i+2, j+2\}, \ldots, \{i+t, j+t\}\}, \tag{45}$$

so that an edge $\{u, v\}$ has $1 \leq u \neq v \leq m$ and corresponds to the indicator that $A_u = A_v$. For $\alpha \in I^*$, no two edges in S_α share a vertex. [A similar structure arises in analyzing matching between two random sequences, except there the graphs are bipartite, see [Arratia, Gordon, Waterman (1986)].] Observe that in the "parallel, overlapping case" $(i', j') = (i+d, j+d)$ with $0 \neq |d| \leq t$, we have that $X_\alpha X_\beta$ is identically zero, thanks to the declumping factors $\mathbf{1}(A_i \neq A_j)$ and $\mathbf{1}(A_{i'} \neq A_{j'})$.

Excluding the parallel, overlapping case, for $\alpha \neq \beta$ the two sets of edges S_α and S_β have no edges in common, so there are $2t$ edges in the union; let $S_{\alpha\beta}$ denote the resulting graph. Different components in the graph have disjoint vertex sets; hence the events corresponding to components are mutually independent. Let $e(k)$ be the number of components having k edges, so that $2t = \sum ke(k)$. For $\alpha, \beta \in I^*$, each vertex in $S_{\alpha\beta}$ has degree at most 2. We claim that the graph $S_{\alpha,\beta}$ has no cycles.

[Proof: Write $d = j - i, e = j' - i'$ and without loss of generality, since the parallel case has been excluded, assume $d < e$. Note that since $\alpha \in I^*$, we have $j > i + t$; hence for an edge $\{w, w'\} \in S_\alpha$, either $w \leq i + t$ and then $w' = w + d$, or else $w > i + t$ and hence $w' = w - d$. Suppose there were a cycle. Take a simple cycle, with leftmost vertex v and rightmost vertex v'. There are two paths from v to v', and each of these paths must alternately use edges from S_α and from S_β. The path starting with S_α must begin $v, v + d$

since we chose v leftmost; and it follows that $v+d > i+t$. The path starting with S_β must begin $v, v+e$. Since $v' \geq v+e > v+d$ the path that started in S_α must continue past $v+d$; it must begin with $v, v+d, v+d+e$ since $v+d-e < v$ would contradict v being leftmost. Thus $v' \geq v+d+e > v+e$ so the path starting with S_β must continue past $v+e$. It must begin $v, v+e, v+e-d$; the path starting $v, v+e, v+e+d$ is excluded because with $v+e$ in the role of w four sentences back, $v+e > v+d > i+t$. It follows, from $\{v+e-d, v+e\}$ being an edge in S_α, with $v+e-d$ in the role of w, that $v+e-d \leq i+t$ and hence $v+e-d < v+d$. Using the above steps again, the path starting with S_β must begin $v, v+e, v+e-d, v+e-d+e, v+e-d+e-d$ and it follows that $v+e-d+e-d < v+d$. Iterating we get that $v+k(e-d) < v+d$ for $k = 1, 2, \ldots$, a contradiction.]

Since each component is a tree, a component with r edges has $r+1$ vertices and thus corresponds to requiring $r+1$ of the random letters to match, which has probability p_{r+1}. Giving away the declumping factors in the first inequality, we have, for $\alpha \neq \beta \in I^*$,

$$\mathbf{E}(X_\alpha X_\beta) \leq \mathbf{E}(R_\alpha R_\beta) = p^{e(1)}(p_3)^{e(2)}(p_4)^{e(3)} \cdots \quad . \tag{46}$$

For pairs (α, β) with overlap of degree two or more, the $O(m^2t^2)$ bound (44) on the number of such pairs is so small that there is only a relatively small loss in the overall upper bound on b_2^* if we use the following coarse treatment of (46), without enumerating cases according to the values of $e(1), e(2), \ldots$. Using (6), that $p_{r+1} \leq (\xi_*)^r$, together with the simple upper bound in (46) yields

$$\begin{aligned} \mathbf{E}(X_\alpha X_\beta) &\leq \mathbf{E}(R_\alpha R_\beta) = \prod (p_{r+1})^{e(r)} \\ &\leq (\xi_*)^{(e(1)+2e(2)+3e(3)+\cdots)} = (\xi_*)^{2t}. \end{aligned} \tag{47}$$

Recall that the number of pairs (α, β) with overlap of degree two or more is order of m^2t^2, and that for $\lambda^* \asymp 1$ we have $t \asymp \log m$ and $m^2p^t \asymp 1$. Using the notation defined by (5), the net contribution to b_2^* from these pairs having overlap of degree two or more is at most order of $t^2m^2(\xi_*)^{2t} = t^2m^2p^{t(1+\gamma)} \asymp t^2p^{t\gamma} \asymp t^2m^{-2\gamma} \asymp (\log m)^2m^{-2\gamma}$.

For pairs (α, β) with overlap of degree one, it is not hard to see that the graph $S_{\alpha,\beta}$ has only components with one or two edges. Thus (46) simplifies to

$$\mathbf{E}(X_\alpha X_\beta) \leq \mathbf{E}(R_\alpha R_\beta) = p^{e(1)}(p_3)^{e(2)}, \tag{48}$$

with $e(1) + 2e(2) = 2t$. Recall, from the discussion following (7), that $p^2 \leq p_3$, with equality in case of the uniform distribution. One could use the bound

$$\mathbf{E}(X_\alpha X_\beta) \leq p^{e(1)}(p_3)^{e(2)} \leq (p_3)^{(e(1)+2e(2))/2} = (p_3)^t = p^{3t(1+\epsilon)/2}. \tag{49}$$

We will show below that the number of pairs (α, β) with overlap of degree one is order of $m^3 t$. Recall that for $\lambda^* \asymp 1$ we have $t \asymp \log m$ and $m^2 p^t \asymp 1$. Using (49), the net contribution to b_2^* from pairs with overlap of degree one would be at most order of $m^3 t p^{3t(1+\epsilon)/2} \asymp t p^{3t\epsilon/2} \asymp (\log m) m^{-3\epsilon}$. To show that this upper bound has larger order than our upper bound on the contribution from overlap of degree two or more, we have to show that $3\epsilon < 2\gamma$. As in (8), consider the random variable D with $D = \xi_a$ on the event $\{A_1 = a\}$. Condition on the event M that $A_1 = A_2$. We have

$$p_4/p = \mathbf{E}(D^2|M) \geq (\mathbf{E}(D|M))^2 = (p_3/p)^2, \tag{50}$$

so that $p_4 \leq (p_3)^2/p$. We also use the inequality $p_4 > (\xi_*)^4$, which follows from there being more than one letter with positiive probability. Unraveling some notation, we have $p^{2+2\gamma} = (\xi_*)^4 < p_4 \leq (p_3)^2/p = p^{2+3\epsilon}$, which proves that $3\epsilon < 2\gamma$. Thus the contribution from pairs with overlap of degree two or more is smaller than that from pairs with overlap of degree one, by at least some power of m.

Since pairs with overlap of degree one yield the main contribution to b_2^*, it is worth some additional effort to give a better upper bound than (49) on these $\mathbf{E}(X_\alpha X_\beta)$. There are two ways to improve the estimate. The first, which has no effect in the uniform case, is to classify the different types of overlap of degree one according to the displacement k involved in the overlap; for the non-uniform case we save a factor asymptotic to $tr/(1-r)$ where $r = p^2/p_3 < 1$ is constant, while $t \asymp \log m$. The second, which is useful for both uniform and non-uniform distributions, and which applies to all degrees of overlap, is to include the declumping factor $\mathbf{1}(A_i \neq A_j)$, which carries over to an improvement of $(1-p)^2$ in the upper bound on $\mathbf{E}(X_\alpha X_\beta)$ for most pairs (α, β). The complication comes in counting the exceptional cases where there aren't two independent declumping effects.

For the first improved upper bound we distinguish cases according to the displacement k between overlapping indices. To be specific, recall that overlap of degree one means that exactly one of the four inequalities $|i-i'| \leq t, |i-j'| \leq t, |j-i'| \leq t, |j-j'| \leq t$ is satisfied. We treat the subcase $|i-i'| \leq t$ so that $k = |i-i'|$; the other three subcases have the same structure. When $k=0$, we have $e(1) = 0, e(2) = t$ so the second inequality in (49) holds with equality. For $1 \leq k \leq t$, we have $e(1) = 2k, e(2) = t-k$ so that $\mathbf{E}(R_\alpha R_\beta) = p^{2k}(p_3)^{t-k} = (p_3)^t r^k$ where $r = p^2/p_3$.

Next we count ordered pairs (α, β) having overlap of degree one together with a specified k, using reasoning similar to that used to derive (41). The number of instances of $k=0$ and overlap of degree one is exactly $6\binom{m-t+1-2t}{3}$ To see this, consider first the subcase $|i-i'| \leq t$.

Picking (α, β) involves choosing values for i, j, j' all in the range 0 to $m-t$, mutually more than t apart; the binomial coefficient gives the number of ways to do this. We need $i' = i$ and thus i is the smallest of the three values chosen, but there remains a two way choice for assigning j, j' to the two larger values. The subcase $|j - j'| \leq t$ also contributes a factor of two. In the subcase $|i - j'| \leq t$ there is only one choice, namely $i' < i = j' < j$, and similary there is only one choice in the subcase $|j - i'| \leq t$. Thus the factor 6 comes from the four subcases as $2 + 2 + 1 + 1$. For each k from 1 to t, there is an additional two way choice, corresponding on the first subcase to $i' = i + k$ versus $i' = i - k$. Note that, in contrast to the case $k = 0$, here $12\binom{m-t+1-2t}{3}$ is not exactly the count of pairs (α, β) having overlap of degree one and a specified k because some of the specified configurations will have overlap of degree two, [for example, $k = 1, i = 3, j = 4 + t, j' = 8 + 2t, i' = i + k$] or the index specified by displacement may be out of range [for example, $k = 1, i = 0, j = 4 + t, j' = 8 + 2t, i' = i - k$].

We have shown that the net contribution to b_2^* from its terms $\mathbf{E}(X_\alpha X_\beta)$ having overlap of degree one and displacement k is $6\binom{m-3t+1}{3}(p_3)^t$ for $k = 0$, and at most $12\binom{m-3t+1}{3}(p_3)^t r^k$ for each of $k = 1, 2, \ldots, t$. Summing over k, we have the following upper bound on the contribution to b_2^* from pairs having overlap of degree one. Recall that $r = p^2/p_3$. In the first inequality, both sums are taken over all pairs having overlap of degree one; we will return to this inequality later to use the declumping factors for a further improvement.

$$\sum_{\alpha,\beta \in I^*, \alpha \sim \beta \text{ degree } 1} \mathbf{E}(X_\alpha X_\beta) \leq \sum \mathbf{E}(R_\alpha R_\beta) \leq (1 + 2r + 2r^2 + \cdots + 2r^t) 6 \binom{m - 3t + 1}{3} (p_3)^t \tag{51}$$

In the nonuniform case, we have $r \equiv p^2/p_3 < 1$ so the first factor is bounded by $1 + 2r/(1-r)$, a constant even as t increases.

The second way to improve the upper bound on $\mathbf{E}(X_\alpha X_\beta)$ is to take account of the declumping factors. Recall the discussion leading up to (46), where, excluding the parallel overlapping case, the graph $S_{\alpha\beta}$ has exactly $2t$ edges, each edge corresponding to one of the matches required by $R_\alpha R_\beta = 1$. Except when $i = 0$ or $i' = 0$, to have $X_\alpha X_\beta = 1$ requires two additional conditions, namely $A_i \neq A_j$ and $A_{i'} \neq A_{j'}$, which correspond to two distinct edges $\{i, j\}$ and $\{i', j'\}$ not in $S_{\alpha\beta}$. The effect of adding these two edges to $S_{\alpha\beta}$ is the same as the effect of increasing t to $t + 1$, so in particular the augmented graph has vertices of degrees one and two only, and no cycles. A component of $S_{\alpha\beta}$ having r edges corresponded to the requirement that $r + 1$ letters match, which has probability p_{r+1}. If in the augmented graph one of these new edges forms a component by itself, the new component corresponds to an event of the form $\{A_i \neq A_j\}$, having probability $(1 - p)$. On the other hand, if in the augmented graph, one of the new edges joins an old component with r edges to form a new component with $r+1$ edges, then this new component corresponds to an event of the form $A_1 = A_2 = \cdots = A_{r+1} \neq A_{r+2}$, having probability $p_{r+1} - p_{r+2}$; see the argument before (26). The net effect of including this one declumping factor is to replace p_{r+1} by $(p_{r+1} - p_{r+2})$, i.e. to multiply by a factor which varies with $r = 1, 2, \ldots, 2t$. Fortunately, this factor is no greater than the simple declumping factor:

$$\frac{p_{r+1} - p_{r+2}}{p_{r+1}} \leq \frac{p - p_3}{p} \leq 1 - p. \tag{52}$$

To see this inequality, one method is to treat r as a continuous variable, and differentiate. Here is a probabilistic proof. As in (8), consider the random variable D with $D = \xi_a$ on the event $\{A_1 = a\}$. Condition on the event M that $A_1 = A_2 = \cdots = A_k$, for $k = 1, 2, \ldots$. We have

$$p_{k+2}/p_k = \mathbf{E}(D^2|M) \geq (\mathbf{E}(D|M))^2 = (p_{k+1}/p_k)^2, \tag{53}$$

so that $p_{k+2}/p_{k+1} \geq p_{k+1}/p_k \cdots \geq p_3/p_2 \geq p_2/p_1 \equiv p$. This proves (52).

Consider cases where $i > 0$ and $i' > 0$, so that there are two extra edges for declumping. If (α, β) has overlap of degree two or more, it can be seen that at least one of the new edges joins an old component, so by (52) the upper bound on $\mathbf{E}(X_\alpha X_\beta)$ in (46) can be improved by multiplying in a factor of $(p - p_3)/p$. If instead (α, β) has overlap of degree one, then we consider cases according to the displacement k, as in the discussion leading up to (51). For $k = 0$ the two new edges together form one new component, corresponding to a new factor of $\mathbf{P}(A_1 \neq A_2, A_1 \neq A_3) = \sum \mathbf{P}(A_1 = a, A_2 \neq a, A_3 \neq a) = \sum \xi_a (1 - \xi_a)^2 = 1 - 2p + p_3$. [In the subcase $|i - i'| = 0$ the new component corresponds to the event $\{A_i \neq A_j, A_i \neq A_{j'}\}$.] For $k = 1$ to t one new edge forms a component by itself, and the other new edge joins with an old, single edge component, corresponding to new factors of $(1-p)$ and $(p-p_3)/p$ respectively. [In the subcase $i' = i + d$, the first factor corresponds to the event $\{A_i \neq A_j\}$, and the second factor corresponds to replacing the event $\{A_{i+d} = A_{j+d}\}$ by the event $\{A_{j'} \neq A_{i+d} = A_{j+d}\}$.]

In cases where $i = 0$ or else $i' = 0$, so that there is exactly one declumping factor, we simply use (52) to save a factor of $(1 - p)$, i.e. we use $\mathbf{E}(X_\alpha X_\beta) \leq (1 - p)\mathbf{E}(R_\alpha R_\beta)$, without trying to classify subcases.

Putting the above considerations together, we have an upper bound $\overline{b_2^*}$ for b_2^* for the nonuniform case. The first term, which is the dominant contribution, is an upper bound on the contribution from overlap of degree one and both $i, i' > 0$. Compared with (51) there are declumping factors, and the top of the binomial coefficient is reduced by 1 since the three designated indices, spaced more than t apart, are chosen from $\{1, \ldots, m - t\}$ instead of from

$\{0, 1, \ldots, m-t\}$. The second term is an upper bound on the contribution from overlap of degree two or more, and both $i, i' > 0$. From (44) the number of pairs (α, β) here is at most $5\binom{m-2t}{2}(2t+1)^2$, with $m-2t$ in place of $m-2t+1$ because 0 is excluded, as above. For the upper bound on $\mathbf{E}(X_\alpha X_\beta)$ we combine $(\xi_*)^2$ from (47) with a declumping factor $(p-p_3)/p$. The third and final term is the smallest for practical cases of m, t; it bounds the contribution from all pairs (α, β) with $i = 0$ or $i' = 0$ (or both), and overlap (of degree one or more). For an upper bound on $\mathbf{E}(X_\alpha X_\beta)$ we use $(\xi_*)^2$ as given by (47), without attaching any declumping factor. The number of such pairs is at most $2m^2(5t+3)$,, using the following argument. There are $(m-t+1)(m-t+2)/2 < m^2/2$ ways to choose w, x, y with $w = 0 \le x \le y \le m-t$. Given such w, x, y, there are at most $5t+3$ ways to pick z with $0 \le z \le m-t$ so that z is within distance t of at least one of w, x, y. Next, in choosing which two of w, x, y, z should serve as the coordinates of α, with the other two variables serving as the coordinates of β, of the $6 = \binom{4}{2}$ ways, at least two are excluded by the requirement that α and β each has no self-overlap. For i.i.d. letters A_i with a *nonuniform* distribution, so that $r \equiv p^2/p_3 < 1$, we define

$$\begin{aligned}\overline{b_2^*} \;\equiv\; & \left((1-2p+p_3) + 2(1-p)\frac{p-p_3}{p}(r + r^2 + \cdots + r^t)\right) 6\binom{m-3t}{3}(p_3)^t \qquad (54)\\ & +5\binom{m-2t}{2}(2t+1)^2\frac{p-p_3}{p}(\xi_*)^{2t} + 2m^2(5t+3)(\xi_*)^{2t}.\end{aligned}$$

For the uniform case, our definition of $\overline{b_2^*}$ is

$$\overline{b_2^*} \equiv b_1^*, \qquad (55)$$

with b_1^* given by (42). We have shown that in both the nonuniform and uniform cases

$$b_2^* \le \overline{b_2^*}.$$

Now assume that $m, t \to \infty$ with $\lambda^* \asymp 1$. ¿From the discussion at (50) we see that the second and third terms are negligible compared with the first. Recall that $\lambda^* \asymp 1$ implies $\lambda^* \sim m^2 p^t (1-p)/2$ and that ϵ is defined by $p_3 = p^{3(1+\epsilon)/2}$, so that $0 < \epsilon < 1/3$ in the nonuniform case. It follows that

$$\overline{b_2^*} \sim \left((1-2p+p_3) + 2(1-p)\frac{p-p_3}{p}\frac{r}{1-r}\right) m^3 (p_3)^t \asymp m^{-3\epsilon}. \qquad (56)$$

Notice that if the underlying distribution of letters were uniform, in the above asymptotics we would have $\epsilon = 1/3$, but the geometric series, instead of being summable, would be t, so the conclusion of (56) would be that the complicated expression in (54) is $\asymp tm^{-1}$, which is no worse than the order of magnitude of b_1^* given by (43).

For i.i.d. letters A_i with a *uniform* distribution, recall from (55) that we use b_1^* as our upper bound $\overline{b_2^*}$; compared to the long expression in (54) this bound is smaller (but of the same asymptotic order), in addition to being more easily derived.

To summarize, for the process $\mathbf{X}^* \equiv (X_\alpha)_{\alpha \in I^*}$ of indicators of leftmost, non self-overlapping repeats of t-tuples in an i.i.d. sequence $A_1A_2\cdots A_m$, we have calculated the exact value λ^* of the expected number of such repeats, and for the ingredients b_1^*, b_2^* needed for Theorem 1, we have an exact expression for the first, and an upper bound $\overline{b_2^*}$ for the second. Thus we have proved

Theorem 2 *Assume that $A_1, A_2, \ldots$ are independent and identically distributed as specified by (2). Let $t \geq 2, m \geq 2t$. Take λ^* given by (20), b_1^* given by (42), and $\overline{b_2^*}$ given by (54) in the nonuniform case and $\overline{b_2^*} \equiv b_1^*$ in the uniform case. For the process $\mathbf{X}^*$ of indicators of leftmost non self-overlapping repeats defined by (18), compared to the Poisson process $\mathbf{Y}^* \equiv (Y_\alpha)_{\alpha \in I^*}$ having independent coordinates and $\mathbf{E}(Y_\alpha) = \mathbf{E}(X_\alpha)$, the total variation distance satisfies*

$$d_{TV}(\mathbf{X}^*, \mathbf{Y}^*) \leq b^*(m,t) \equiv b_1^* + \overline{b_2^*} \quad = 2b_1^* \sim 16(\lambda^*)^2 \tfrac{t}{m} \quad \textit{uniform case,} \tag{57}$$

$$\asymp m^3 (p_3)^t \asymp m^{-3\epsilon} \quad \textit{nonuniform case.} \tag{58}$$

For the random variable $W^* \equiv \sum_{\alpha \in I^*} X_\alpha$, compared to a Poisson random variable K^* with $\mathbf{E}(K^*) = \lambda^*$,

$$d_{TV}(W^*, K^*) \leq r^*(m,t) \equiv \frac{1 - e^{-\lambda^*}}{\lambda^*}(b_1^* + \overline{b_2^*}) < b^*(m,t) \equiv (b_1^* + \overline{b_2^*}). \tag{59}$$

For the asymptotics, $\epsilon > 0$ is defined in (7), and in the nonuniform case, $3\epsilon < 1$. The asymptotics are valid as $m, t \to \infty$ in any way with λ^* bounded away from zero and infinity, equivalently, with $t = \frac{\log m^2}{\log(1/p)} + O(1)$.

æ

2.5 Repeats, allowing self-overlap

We now carry out a process approximation for $\mathbf{X} \equiv (X_\alpha)_{\alpha \in I}$, the process of indicators of leftmost repeats of t-tuples in the i.i.d. sequence $A_1A_2\cdots A_m$. The comparison process is $\mathbf{Y} \equiv (Y_\alpha)_{\alpha \in I}$ where the Y_α are independent Poisson random variables with $\mathbf{E}Y_\alpha = \mathbf{E}X_\alpha$. We will apply the Chen-Stein method, given in Theorem 1, directly to the process $\mathbf{X}$, so our job is to compute (good upper bounds on) b_1 and b_2. Most of the summands involved in b_1 and b_2 have both indices α, β having no self-overlap, and these have been handled in the previous section. The only cases that remain are those in which α or β or both have self-overlap.

Before we proceed with our direct analysis, we employ an easy strategy for process approximation which avoids second moments for self-overlapping repeats. Tables 1,2, and 3 show that this easy strategy has worse error bounds for some realistic values of m and t; see also the discussion after (68). Basically, the process $\mathbf{X}$ for repeats allowing self- overlap can be handled as a corollary to the treatment of the process $\mathbf{X}^*$ excluding self-overlaps, using first moments to bound the additional error. For $\lambda - \lambda^*$, the expected number of repeats having self-overlap, the exact value is given by (23), and (28) states that $\lambda - \lambda^* = O(m^{-\epsilon} \log m)$. It only requires the triangle inequality to get a conclusion having the same form as Theorem 3:

Corollary 1 *Under the same hypotheses as Theorem 3,*

$$d_{TV}(\mathbf{X}, \mathbf{Y}) \leq 2(\lambda - \lambda^*) + d_{TV}(\mathbf{X}^*, \mathbf{Y}^*). \tag{60}$$

Proof. The basic idea is very simple: $d_{TV}(\mathbf{X}, \mathbf{X}^*) \leq \lambda - \lambda^*$ and $d_{TV}(\mathbf{Y}, \mathbf{Y}^*) \leq \lambda - \lambda^*$. There is a slight technicality in that the processes being compared, say $\mathbf{X}$ and $\mathbf{X}^*$, do not take values in the same space. To get around this, we need to redefine $\mathbf{X}^*$ by taking $\mathbf{X}^* \equiv (X^*_\alpha)_{\alpha \in I}$ where X^*_α is defined to be X_α if $\alpha \in I^*$ and to be zero if $\alpha \in I \setminus I^*$. With this definition, the two processes $\mathbf{X}, \mathbf{X}^*$ have values in the same space, so that $d_{TV}(\mathbf{X}, \mathbf{X}^*)$ is meaningful. Furthermore, they are both constructed on the same probability space, so that $d_{TV}(\mathbf{X}, \mathbf{X}^*) \leq \mathbf{P}(\mathbf{X} \neq \mathbf{X}^*) = \mathbf{P}(\cup_{\alpha \in I \setminus I^*} \{X_\alpha > 0\}) \leq \sum_{\alpha \in I \setminus I^*} \mathbf{E} X_\alpha = \lambda - \lambda^*$. The same argument works to show $d_{TV}(\mathbf{Y}, \mathbf{Y}^*) \leq \lambda - \lambda^*$. ■

We return to the approximation of $\mathbf{X}$ by the direct application of the Chen-Stein method. Recall that b_1 is a sum over pairs $(\alpha, \beta) \in I^2$ with $\alpha \sim \beta$; see (34) for the discussion of the overlap relation. Each of α, β is classified according to whether or not it has self-overlap, leading to a partition of the terms of b_1 into four cases. The first case, no self-overlap in either index, results in b_1^*, treated in the previous section. The second and third cases, with self-overlap in exactly one of the indices, have the same sum. For the fourth case, both indices have self-overlap. Thus

$$b_1 = b_1^* + 2 \sum_{\alpha \in I \setminus I^*, \beta \in I^*, \alpha \sim \beta} \mathbf{E} X_\alpha \mathbf{E} X_\beta + \sum_{\alpha, \beta \in I \setminus I^*, \alpha \sim \beta} \mathbf{E} X_\alpha \mathbf{E} X_\beta. \tag{61}$$

For an upper bound on the middle sum in (61), note first that given α with self-overlap, the number of $\beta \in I^*$ satisfying $\alpha \sim \beta$ is at most $(3t+1)(m-2t+1)$. [To see this, α has self-overlap, i.e $|i-j| \le t$, so there are at most $3t+1$ integers x within t of i or j. There are at most $m-2t+1$ integers y with $0 \le y \le m-t$, $|x-y| > t$. We form $\beta = (\min(x,y), \max(x,y))$; this accounts for all $\beta \in I^*$ and also for some extraneous β.] For $\mathbf{E}X_\beta$ we use the upper bound p^t. [Although most of the terms have $\mathbf{E}X_\beta = (1-p)p^t$, classifying the cases would add complexity]. Thus

$$2 \sum_{\alpha \in I \setminus I^*, \beta \in I^*, \alpha \sim \beta} \mathbf{E}X_\alpha \mathbf{E}X_\beta \le 2 \sum_{\alpha \in I \setminus I^*} \mathbf{E}X_\alpha (3t+1)(m-2t+1)p^t = 2(\lambda - \lambda^*)(3t+1)(m-2t+1)p^t. \tag{62}$$

If $m, t \to \infty$ with $\lambda \asymp 1$ then from (29), $\lambda - \lambda^* = O(m^{-\gamma} \log m)$ and the upper bound in (62) is $O(m^{-1-\gamma}(\log m)^2)$.

For an upper bound on the last sum in (61), note first that given α with self-overlap, the number of $\beta \in I \setminus I^*$ satisfying $\alpha \sim \beta$ is at most $(3t+1)(2t)$. [The difference with the previous paragraph is that β has self-overlap, so there are at most $2t$ choices for y with $0 < |x-y| \le t$.] For an upper bound on $\mathbf{E}X_\beta$ we use (27). Thus

$$\sum_{\alpha, \beta \in I \setminus I^*, \alpha \sim \beta} \mathbf{E}X_\alpha \mathbf{E}X_\beta \le \sum_{\alpha \in I \setminus I^*} \mathbf{E}X_\alpha (3t+1)(2t)(\xi_*)^t = (\lambda - \lambda^*)(3t+1)(2t)p^{(1+\gamma)t/2}. \tag{63}$$

If $m, t \to \infty$ with $\lambda \asymp 1$ then from (29), $\lambda - \lambda^* = O(m^{-\gamma} \log m)$ and the upper bound in (63) is $O(m^{-1-2\gamma}(\log m)^3)$.

Putting these together we have the following upper bound $\overline{b_1}$ for b_1:

$$b_1 \le \overline{b_1} \equiv b_1^* + 2(\lambda - \lambda^*)(3t+1)\left((m-2t+1)p^t + tp^{(1+\epsilon)t/2}\right) \sim \frac{8t}{m}\lambda^2. \tag{64}$$

Recall, the value of b_1^* is given by (42), the value and bound for $\lambda - \lambda^*$ are given by (20) and (28), and the asymptotics are valid as $m, t \to \infty$ with $\lambda \asymp 1$. Note that b_1^*, from the case with no self overlap, makes the dominant contribution to $\overline{b_1}$, and that $\overline{b_1} - b_1^* = O(m^{-1-\gamma}(\log m)^2)$.

For b_2 we use a different grouping of the terms not incorporated in b_2^*. Case one contains all (α, β) such that the union of the four intervals $[i, i+t], [j, j+t], [i', i'+t], [j', j'+t]$ is one connected interval *and* the graph $S_{\alpha\beta}$ of matching edges, defined following (45), has cycles. Case two contains all the remaining terms. In case one, even for uniformly distributed letters, X_α and X_β are not independent; in some cases they are highly positively correlated. The simplest example is $\alpha = (0,1), \beta = (0,2)$, for which $X_\alpha X_\beta = R_\alpha R_\beta = \mathbf{1}(A_1 = A_2 = \cdots = A_{t+2})$ with $\mathbf{E}(X_\alpha X_\beta) = p_{t+2} \gg (p^t)^2 = \mathbf{E}X_\alpha \mathbf{E}X_\beta$. In case two, we can show that $S_{\alpha\beta}$ has no cycles, and it is easy to show that the net contribution from case two is negligible.

First we analyze case two. The number of instances is at most $2m^2t(3t+1)$. To see this, the first factor of two corresponds to designating which of α, β is required to have self-overlap. If α has self-overlap then there are at most mt choices for α, at most $3t+1$ values for the coordinate of β that overlaps α, and at most m values for the other coordinate of β.

Next we argue that in case two, $S_{\alpha\beta}$ has no cycles. Assume that the union of the four intervals $[i,i+t],[j,j+t],[i',i'+t],[j',j'+t]$ is not one connected interval. Since $\alpha \sim \beta$, this forces at least one of α,β to have no self-overlap. We have already excluded terms where both α,β have no self-overlap since these are incorporated into b_2^*. Assume that α has self-overlap, so $[i,i+t] \cup [j,j+t]$ forms one connected interval. Thus either $[i',i'+t]$ or else $[j',j'+t]$ must join this interval; the other must form a separate interval. Assume that $[i,i+t] \cup [j,j+t] \cup [i',i'+t]$ form one connected interval, so that $[j,j+t]$ does not intersect this interval. Note that both the graphs S_α and S_β have no cycles. Furthermore, as $\beta \in I^*$, in S_β each vertex has degree exactly 1; no two edges share a vertex. Any edge in S_β connects a vertex from the connected set $\{i,\ldots,i+t\} \cup \{j,\ldots,j+t\} \cup \{j',\ldots,j'+t\}$ with a vertex in the isolated set $\{j',\ldots,j'+t\}$, and from this vertex no further edge is emanating. Thus, no cycle in $S_{\alpha\beta}$ can contain an edge from S_β, and must thus consist solely of edges in S_α. But S_α itself contains no cycles. The same argument applies if we interchange i' and j', or if we interchange α and β. This proves that $S_{\alpha\beta}$ contains no cycles.

Since in case two, the graph $S_{\alpha\beta}$ has no cycles, we can apply the bound (47), stating $\mathbf{E}(X_\alpha X_\beta) \le \mathbf{E}(R_\alpha R_\beta) \le (\xi_*)^{2t}$. Recall that ξ_* is the probability of the most likely letter; see (5). Thus the net contribution to b_2 from case two satisfies the upper bound

$$\sum_{\text{case two}} \mathbf{E}(X_\alpha X_\beta) \le 2m^2 t(3t+1)(\xi_*)^{2t}, \tag{65}$$

which is of order $m^{-2\gamma}(\log m)^2$ when $m,t \to \infty$ with $\lambda \asymp 1$.

We will develop our upper bound on case one in several stages. There is an easily derived upper bound on case one, which converges to zero as $m,t \to \infty$ with $\lambda \asymp 1$, as follows. There are at most order of mt^3 terms in case one. The worst case value for $\mathbf{E}(X_\alpha X_\beta)$ is $p_{t+2} \le (\xi_*)^{t+1} = O(m^{-1-\gamma})$. This yields an upper bound $O(m^{-\gamma}t^3)$ on the net contribution from case one.

Working harder we can derive (68) where the factor t^3 is replaced by a constant $c_0(\xi_*)$. For the application to DNA, $\xi_* \ge 1/4$ and some values of the constant are $c_0(1/4) = 28$ (exactly), $c_0(.3) = 46.9154$, $c_0(.35) = 76.7355$ and $c_0(1/2) = 324$ (exactly). For current SBH chips, $l \ge 8$ and so $t \ge 7$, and the $O(t^3)$ bound would be much worse.

For case one, the union of the four intervals $[i,i+t],[j,j+t],[i',i'+t],[j',j'+t]$ is one connected interval, say of length $t+k$, corresponding to $t+k$ letters involved in the matches indicated by $R_\alpha R_\beta$. Exclude the parallel, overlapping case, so that the graph $S_{\alpha\beta}$ has exactly

$2t$ edges, and the sum of the vertex degrees is exactly $4t$. Let c be the number of connected components of this graph; enumerate these and let $n(1), n(2), \ldots, n(c)$ be the number of vertices in these components. The upper bound corresponding to (47) is now

$$\mathbf{E}(X_\alpha X_\beta) \leq \mathbf{E}(R_\alpha R_\beta) = p_{n(1)} \cdots p_{n(c)} \leq (\xi_*)^{n(1)-1+\cdots+n(c)-1} = (\xi_*)^{t+k-c}. \tag{66}$$

The next step is to prove that $c \leq 2k/3$; the method involves summing up the degrees of the vertices of $S_{\alpha\beta}$. As in the argument leading to (47), write $d = j - i, e = j' - i'$ and without loss of generality, since the parallel case has been excluded, assume $d < e$. A vertex v has degree at most four, since its only possible neighbors are $v \pm d, v \pm e$. We claim that for a component with n vertices, $4n$ exceeds the sum of the degrees, of the vertices in that component, by at least six. [To see this, for $n = 2$, check that the component has exactly one edge; for $n = 3$ check that the component has at most three edges. For $n \geq 4$ the leftmost vertex v has degree at most two (with neighbors $v + d$ and $v + e$), the second to leftmost vertex v has degree at most three (with neighbors $v - d, v + d, v + e$), and similarly at the rightmost two vertices.] Summing the degrees of all the $t + k$ vertices to get $4t$, and grouping by components, we have

$$4(t + k) \geq 4t + 6c,$$

and it follows that $c \leq 2k/3$, and since c is an integer, $c \leq \lfloor 2k/3 \rfloor$. Thus the upper bound (66) simplifies to

$$\mathbf{E}(X_\alpha X_\beta) \leq \mathbf{E}(R_\alpha R_\beta) \leq (\xi_*)^{t+k-c} \leq (\xi_*)^{t+\lceil k/3 \rceil}. \tag{67}$$

Next we bound the number of terms in case one for each possible k. Since $\alpha \neq \beta$, the smallest possible value for k is $k = 2$. The number of pairs (α, β) in case one, involving $t + k$ letters in matches, is at most m times $2(k-1)^2 + k(k-1) = 3k^2 - 5k + 2$; we exclude the parallel overlapping case, as in the discussion following (45). This bound does not use the requirement that there be a cycle, and is derived as follows. There are at most m choices for $b = \min(i, i')$. Either α or else β comes first in lexicographic order; the two cases are equinumerous and we assume α comes first, so that $i = b$. If $\alpha = (b, b + k)$ then there are at most $k(k-1)/2$ choices for β with $b < i' < j' \leq b + k$. If $\alpha \neq (b, b + k)$ then $(k-1)^2$ choices remain: we must have $j' = b + k$, there are $k - 1$ choices $j = b + 1, \ldots, b + k - 1$, and there are $k-1$ choices for $i' = b, b+1, \ldots, b+k-1$; excluding the one parallel overlapping case, namely $i' - b = b + k - j$. Thus there are at most $m(2[k(k-1)/2 + (k-1)^2] = m(k(k-1) + 2(k-1)^2)$ pairs (α, β) such that the union of the four intervals $[i, i + t], [j, j + t], [i', i' + t], [j', j' + t]$ is one connected interval.

Summing over all pairs (α, β) in case one, the contribution to b_2 is at most

$$\sum_{\text{case one}} \mathbf{E}(X_\alpha X_\beta) \leq \sum_{k \geq 2} m(3k^2 - 5k + 2)(\xi_*)^{t+\lceil k/3 \rceil} = m(\xi_*)^t c_0(\xi_*), \tag{68}$$

where $c_0(x) = \sum_{k\geq 2}(3k^2-5k+2)x^{\lceil k/3 \rceil} = 18x(2x^2+6x+1)/(1-x)^3$. (The sum was simplified using DERIVE.) For $m, t \to \infty$ with $\lambda \asymp 1$, the upper bound in (68) is order of $m^{-\gamma}$, with $\gamma > 0$ given by (5).

Even this bound is unsatisfactory for realistic instances of m, t. This can be seen in the uniform case. Here, $b_1^* + b_2^* \sim (16t/m)\lambda^2, \lambda - \lambda^* \sim mtp^t \sim 2(1-p)^{-1}(t/m)\lambda$, so the upper bound from Corollary 1 is asymptotically $(t/m)(16\lambda^2 + (4/(1-p))\lambda)$. In contrast, the contribution from case one is bounded above by (68) which simplifies to $mp^t c_0(p) \sim 2(c_0(p)/(1-p))(1/m)\lambda$. For small λ, in both our bounds the λ^2/m term is dominated by the λ/m term. Comparing coefficients of $(1/m)\lambda$, we have $4t/(1-p)$ from Corollary 1 versus $2c_0(p)/(1-p)$ from (68), which is $2t$ versus $c_0(p)$, and $c_0(1/4) = 28$. To handle small values like $t = 7, 9, 11$ we need to work harder!

There is a lot of slack in the upper bound on the number of components, $c \leq 2k/3$, used to get (67). For example, for $k = 3$ the bound says $c \leq 2$, but for $t \geq 3$, all $2{\cdot}7$ instances of (α, β) have $c = 1$. For $k = 4$, the bound says $c \leq 2$, but of the $2 \cdot 15$ instances (for sufficiently large t that the "has cycles" requirement of case one is satisfied) most have $c = 1$; the exceptions (given in the special case $i = 0$ or $i' = 0$) are (α, β) or $(\beta, \alpha) = ((0,2),(0,4)), ((0,4),(1,3))$, or $((0,4),(2,4))$. In these last three examples, it is easy to see that the displacements $j - i$ and $j' - i'$ involved in α and β have greatest common divisor $d = 2$ and hence $S_{\alpha\beta}$ has at least d components. This line of attack, trying to describe the number of components in terms of the two displacements and their greatest common divisor, seems like it might yield exact results, but there is an additional complication from varying t. Given (α, β), the number of components of $S_{\alpha\beta}$ is really $c \equiv c(\alpha, \beta; t)$. It is easy to see that this is nonincreasing as a function of t, using the case one requirement that the union of four intervals is one interval. [Explicitly, with $b = \min(i, i')$, the vertex set of $S_{\alpha\beta}$ is $\{b+1, b+2, \ldots, b+k+t\}$. When t is replaced by $t+1$, the graph $S_{\alpha\beta}$ acquires two new edges, namely $\{i+t+1, j+t+1\}$ and $\{i'+t+1, j'+t+1\}$, and one new vertex, namely $b+k+t+1$, but this new vertex is connected to an old vertex, and hence to an old component, by one or both of the new edges.] We also note that case one is really case one(t) and the indicator $\mathbf{1}((\alpha, \beta) \in \text{case one})$ is nondecreasing in t. The complexity which prevents us from proving the conjecture following (75) is that, for example when $d = 2$ so that there are at least two components, namely one for even integers and another for odd integers, as t increases to the threshold of (α, β) being in case one, a cycle may have formed in one component but not the other. To summarize what we have proved, for all $t \geq 2$, for all (α, β), with $k \equiv k(\alpha\beta) = \max(j, j') - \min(i, i')$,

$$(k - c(\alpha, \beta; t))\mathbf{1}((\alpha, \beta) \in \text{ case one } (t)) \leq (k - c(\alpha, \beta; t+1))\mathbf{1}((\alpha, \beta) \in \text{ case one } (t+1)). \quad (69)$$

Thus, we enlisted a computer to generate all instances of case one (with $b \equiv \min(i, i') = 0$), for $k \leq 30$, and to find exactly the number c of components. For each instance, the value $k - c$ ends up as the exponent of (ξ_*) in (67), and taking out the common factor $(\xi_*)^t$ gives

$$(\xi_*)^{-t}\mathbf{E}(R_\alpha R_\beta) \leq (\xi_*)^{k-c}. \quad (70)$$

For $t = 7$, the overlap requirement of case one forces $k < 22$, and summing x^{k-c} over all instances of (α, β) with $\min(i, i') = 0$ yields a polynomial c_7 with

$$c_7(x) = 2(2x + 10x^2 + 18x^3 + 21x^4 + 42x^5 + 89x^6). \quad (71)$$

In the upper bound (68), $c_7(x)$ can be used in place of $c_0(x)$, and for the uniform case on four letters, we have $c_7(1/4) =$ (exactly) $6353/2048 \approx 3.102$; instead of $c_0(1/4) = 28$.

Once the computer has been enlisted, it is very easy to also handle the declumping factors. Given (α, β), say with $i' > i$, we know that in addition to the matches required for $R_\alpha R_\beta = 1$ to have $X_\beta = 1$ requires that $A_{i'} \neq A_{j'}$. If i' and j' are in the same component of $S_{\alpha\beta}$, then $R_\alpha R_\beta = 1$ requires $A_{i'} = A_{j'}$, and hence

$$\mathbf{E}(X_\alpha X_\beta) = 0. \quad (72)$$

Eliminating these instances leads to replacing $c_7(x)$ by $cd_7(x)$ where

$$cd_7(x) = 2(x + 3x^2 + 3x^3 + 7x^4 + 3x^5 + 11x^6). \quad (73)$$

For the uniform case on four letters, we have $cd_7(1/4) =$ (exactly) $2119/2048 \approx 1.0347$.

Furthermore when $b = \min(i, i')$ satisfies $b > 0$ there is a declumping factor of $(1 - p)$ or less, from the requirement that A_b does not match another position (if the position corresponds to a vertex of $S_{\alpha\beta}$ in a component of size r, then this declumping factor is exactly $p_r - p_{r+1}$; see the discussion before (26) and the inequality (52)). This reduces our upper bound on the net contribution from case one, *only in the case* $t = 7$, to $(1 + (m - t)(1 - p))cd_7(\xi_*)$; the term 1 without the factor $1 - p$ corresponds to the case $b = 0$.

Corresponding to (73), which gives the counts for the case $t = 7$, for $t = 2, 3, \ldots, 11$ the polynomials $cd_t(x)$ given below can be used in upper bounds of the form

$$\sum_{\text{case one}} \mathbf{E}(X_\alpha X_\beta) \leq (1 + (1 - p)(m - t))(\xi_*)^t cd_t(\xi_*). \quad (74)$$

In $cd_t(x)$, the coefficient of x^r is the number of instances of (α,β) with $\min(i,i')=0$, such that $\mathbf{E}(X_\alpha X_\beta)>0$ and the upper bound (70) applies with $k-c=r$. By brute force enumeration, the polynomials are

$$\begin{aligned}
cd_2(x) &= 2x\\
cd_3(x) &= 2(x+3x^2)\\
cd_4(x) &= 2(x+3x^2+3x^3)\\
cd_5(x) &= 2(x+3x^2+3x^3+7x^4)\\
cd_6(x) &= 2(x+3x^2+3x^3+7x^4+3x^5)\\
cd_7(x) &= 2(x+3x^2+3x^3+7x^4+3x^5+11x^6)\\
cd_8(x) &= 2(x+3x^2+3x^3+7x^4+3x^5+11x^6+5x^7)\\
cd_9(x) &= 2(x+3x^2+3x^3+7x^4+3x^5+11x^6+5x^7+13x^8)\\
cd_{10}(x) &= 2(x+3x^2+3x^3+7x^4+3x^5+11x^6+5x^7+13x^8+7x^9)\\
cd_{11}(x) &= 2(x+3x^2+3x^3+7x^4+3x^5+11x^6+5x^7+13x^8+7x^9+15x^{10}) \qquad (75)
\end{aligned}$$

While it appears almost certain from the data that there is a single power series such that for all $t\ge 2$, the polynomial cd_t is just the first $t-1$ terms, we have no proof of this conjecture.

In order to have good upper bounds available for all t, and without being able to prove the above conjecture, we have the following. For given (α,β), let $t_0\equiv t_0(\alpha,\beta)$ be the smallest t such that $(\alpha,\beta)\in$ case one(t). Using (69), for any t, $(\xi_*)^{-t}$ times the contribution $\mathbf{E}(X_\alpha X_\beta)$ to case one indexed by (α,β) can be bounded above by the $(k-c(t_0))$th power of ξ_*. Note that in the argument leading to (72), if i',j' are in the same component using $t=t_0$ then the same holds for all $t>t_0$. The following polynomial $c_{k\le 30}(x)$ has, as the coefficient of x^r, the number of (α,β) such that $\min(i,i')=0$ and $k-c(\alpha,\beta;t_0)=r$ and $\mathbf{E}(X_\alpha X_\beta)>0$ with $t=t_0$ in the declumping argument (72) and $k(\alpha,\beta)\le 30$.

$$\begin{aligned}
c_{k\le 30}(x) &= 2(x+3x^2+3x^3+7x^4+3x^5+11x^6+5x^7+13x^8+7x^9+15x^{10} \qquad (76)\\
&\quad +5x^{11}+27x^{12}+7x^{13}+15x^{14}+13x^{15}+28x^{16}+6x^{17}+32x^{18}+8x^{19}+30x^{20}\\
&\quad +16x^{21}+26x^{22}+8x^{23}+48x^{24}+14x^{25}+24x^{26}+16x^{27}+36x^{28}+8x^{29}).
\end{aligned}$$

For $k>30$, we simply bound the contributions as in (68). This gives

$$c_{k>30}(x)\equiv\sum_{k>30}(3k^2-5k+2)x^{\lceil k/3\rceil}=2x^{10}(1276x^3-504x^2-3513x+2822)/(1-x)^3.$$

(The sum was simplified using DERIVE.) Combining these, we have a power series $cd_\infty(x)$ given by

$$cd_\infty(x) = c_{k\le 30}(x) + c_{k>30}(x) \tag{77}$$

such that there are upper bounds analogous to (74), but valid for all $t \ge 2$:

$$\sum_{\text{case one}} \mathbf{E}(X_\alpha X_\beta) \le (1 + (1-p)(m-t))(\xi_*)^t cd_\infty(\xi_*). \tag{78}$$

This bound is not much worse than the bound using (74), for example with $t = 7$, $cd_\infty(1/4) \approx 1.0445$, versus $cd_7(1/4) \approx 1.0347$.

Combining our upper bound $\overline{b_2^*}$ on b_2^* with the estimates for case one and case two, we obtain the following expression $\overline{b_2}$ satisfying $b_2 \le \overline{b_2}$.

$$\bar{b}_2 \equiv \overline{b_2^*} + (1 + (1-p)(m-t))(\xi_*)^t cd_\infty(\xi_*) + 2m^2 t(3t+1)(\xi_*)^{2t}, \tag{79}$$

where $cd_\infty(\xi_*)$ is given by (78) and $\overline{b_2^*}$ is given by (54) or (55). For asymptotics in the uniform case as $m, t \to \infty$ with $\lambda \asymp 1$, the contribution from (78) is $\asymp 1/m$, and hence is dominated by the $\sim t/m$ terms, b_1^* and b_2^*. For asymptotics in the nonuniform case when $m, t \to \infty$ with $\lambda \asymp 1$, the dominant contribution is the term from (78), which is $\asymp m(\xi_*)^t \asymp m^{-\gamma}$. To see that this is the dominant term, recall from (4) that $\gamma \le 3\epsilon$, with equality only in the uniform case, so using (56) we conclude that $\bar{b}_2 = O(m^{-\gamma})$ in the nonuniform case.

Summarizing the above, we have derived the following Poisson process approximation for $\mathbf{X}$.

Theorem 3 *Assume that $A_1, A_2, \ldots$ are independent and identically distributed as specified by (2). Let $t \ge 2, m \ge 2t$. Take λ given by (23), $\overline{b_1}$ given by (64), and $\overline{b_2}$ given by (79). For the process $\mathbf{X}$ of indicators of leftmost repeats, allowing self-overlap, defined by (17), compared to the Poisson process $\mathbf{Y} \equiv (Y_\alpha)_{\alpha \in I}$ having independent coordinates and $\mathbf{E}(Y_\alpha) = \mathbf{E}(X_\alpha)$, the total variation distance satisfies*

$$d_{TV}(\mathbf{X}, \mathbf{Y}) \le b(m,t) \equiv \overline{b_1} + \overline{b_2} \qquad \sim 16(\lambda^*)^2 \tfrac{t}{m} \qquad \textit{uniform case,} \tag{80}$$

$$\asymp m(\xi_*)^t \asymp m^{-\gamma} \qquad \textit{nonuniform case.} \tag{81}$$

For the random variable $W \equiv \sum_{\alpha \in I} X_\alpha$, compared to a Poisson random variable K with $\mathbf{E}(K) = \lambda$,

$$d_{TV}(W, K) \le r(m,t) \equiv \frac{1 - e^{-\lambda}}{\lambda}(\overline{b_1} + \overline{b_2}) < b(m,t) \equiv (\overline{b_1} + \overline{b_2}). \tag{82}$$

The asymptotics are valid as $m, t \to \infty$ in any way with λ^* bounded away from zero and infinity, equivalently, with $t = \frac{\log m^2}{\log(1/p)} + O(1)$. For the nonuniform case, $0 < \gamma < 1$ is defined in (5).

In the following three tables, we illustrate the above approximation theorems with the alphabet $S = \{A, C, G, T\}$. The first six columns give the values of m and t, the expected number of repeats with and without allowing self-overlap, and (our upper bounds on) the Chen-Stein bounds on the total variation distance. In the last column, we give the bound that could be obtained using the "easy" strategy, Corollary 1, for the case allowing self-overlap. Tables 1 and 2 are for the uniform distribution, with $p = \xi_* = 1/4$. To illustrate the error associated with the asymptotic relation $\lambda \sim m^2 p^t (1-p)/2$, in which increasing t by one is exactly compensated by doubling m, Table 1 includes a progression of five pairs, $(m,t) = (50,5), (200,7), \ldots, (6400,13)$. For the next group of four entries, m is chosen so that λ is close to .6064; and for the last group in Table 1, m is chosen so that λ is close to 2.7465. The values are truncated after their fourth digit after 0. When bounds such as $b(m,t)$ are larger than 1, we report that large value because it reflects the behavior of second moments, but of course, as a bound on total variation distances for probabilities, they can always be truncated at 1.

Table 1. Poisson approximation in the uniform case

m	t	λ^*	λ	$b^*(m,t)$	$b(m,t)$	"easy strategy"
50	5	.6103	.7690	.5332	1.1676	.8506
200	7	.7989	.8599	.3632	.4701	.4852
800	9	.8766	.8969	.1438	.1579	.1844
1600	11	.2228	.2259	.0056	.0062	.0118
6400	13	.2270	.2280	.0017	.0018	.0035
169	7	.5558	.6068	.2057	.2808	.3078
659	9	.5892	.6059	.0786	.0883	.1119
2615	11	.6015	.6066	.0253	.0266	.0355
10430	13	.6049	.6064	.0075	.0077	.0159
353	7	2.6432	2.7532	2.3088	2.6621	2.5288
1394	9	2.7117	2.7473	.7952	.8387	.8663
5553	11	2.7359	2.7468	.2474	.2532	.2692
22182	13	2.7432	2.7464	.0732	.0742	.0796
$5.675 \cdot 10^6$	21	2.74597	2.74599	$4.570 \cdot 10^{-4}$	$4.598 \cdot 10^{-4}$	$4.984 \cdot 10^{-4}$
$5.81 \cdot 10^9$	31	2.74488	2.74488	$6.535 \cdot 10^{-7}$	$6.563 \cdot 10^{-7}$	$6.535 \cdot 10^{-7}$

The next table gives a few values that indicate how the bounds derived for the general case work when applied to the uniform case. The pairs (m, t) below all come from Table 1, so that values can be compared. In particular, it serves as a check that the values of λ below, using expression (23), agree with the values of λ above, using expression (35).

Table 2. Expressions for the general case, evaluated at the uniform case

m	t	λ^*	λ	$b^*(m,t)$	$b(m,t)$	"easy strategy"
169	7	.5558	.6068	.2483	.3346	.3504
659	9	.5892	.6059	.0830	.0925	.1163
5553	11	2.7359	2.7468	.2492	.2529	.2710
22182	13	2.7432	2.7464	.0733	.0738	.0798

The next table shows that the error bounds for our Poisson approximations, with similar m and λ, can be much worse in a non-uniform case. We used the distribution $p_A = 0.3544, p_C = 0.1430, p_G = 0.1451, p_T = 0.3575$; this distribution was found in the liverwort Marchantia polymorpha. For this distribution, $p = 0.294915, \xi_* = 0.3575$ and $cd_\infty = 2.483$. To illustrate the threshhold of informative Poisson approximation, the first group of entries has a fixed $m = 50$ and $t = 5, 6, 7$; then $m = 100$ and $t = 7, 8, 9$. The second group has, for each of $t = 7, 9, 11, 13$, a value m such that λ is close to 0.6064. In the last group, for values $t = 13, 15, 17$ and $t = 6.02 \cdot 10^{32}$, we used a value of m such that λ is close to 2.7465.

Table 3. Poisson approximation for $p_A = .3544, p_C = .1430, p_G = .1451, p_T = .3575, p = .2949$

m	t	λ^*	λ	$b^*(m,t)$	$b(m,t)$	"easy strategy"
50	5	1.316	1.741	18.15	34.88	19.007
50	6	.3511	.5082	2.668	5.801	2.982
50	7	.0931	.1504	.3875	.9668	.5020
100	7	.5167	.6436	2.288	4.386	2.542
100	8	.1454	.1912	.3319	.7067	.4235
100	9	.0409	.0573	.0480	.1179	.0809
97	7	.4816	.6043	2.1220	4.1022	2.3674
321	9	.5495	.6069	.7680	1.3345	.8828
1081	11	.5813	.6068	.2454	.4042	.2964
3663	13	.5956	.6066	.0744	.1190	.0963
7804	13	2.7234	2.7468	.5016	.6792	.5485
26470	15	2.7366	2.7466	.1613	.2092	.1811
89783	17	2.7424	2.7466	.0526	.0659	.0610
$6.02 \cdot 10^{23}$	88	2.7528	2.7528	$6.766 \cdot 10^{-18}$	$5.177 \cdot 10^{-16}$	$6.698 \cdot 10^{-16}$

2.6 Distribution of longest repeats

Consider the length of the longest repeat in a sequence. There are actually two problems, according to whether or not self-ovelap is allowed. For example, in the sequence TATATATA, the length is 6 if self-overlap is allowed, and 4 if self-overlap is excluded. ¿From the Poisson process approximation of where all leftmost repeats of matching t-tuples occur, an immediate corollary is a description of the length of the longest repeat, allowing self-overlap. For the situation excluding self-overlap, a little extra work is needed!

The length L_m of the longest repeat in $A_1 A_2 \cdots A_m$, allowing self-overlap, is naturally defined by

$$\begin{aligned} L_m \equiv \max \{ t : t = 0 \text{ or } & \text{ for some } i, j, 0 \leq i < j \leq m - t \text{ and} \\ & A_{i+1} A_{i+2} \cdots A_{i+t} = A_{j+1} A_{j+2} \cdots A_{j+t} \} . \end{aligned} \tag{83}$$

Similarly, the length L_m^* of the longest repeat not allowing self-overlap is defined as above but with the added restriction that $i + t \leq j$ so that the matching substrings $A_{i+1} A_{i+2} \cdots A_{i+t}$ and $A_{j+1} A_{j+2} \cdots A_{j+t}$ share no letters. Thus the natural definition is

$$\begin{aligned} L_m^* \equiv \max \{ t : t = 0 \text{ or } & \text{ for some } i, j, 0 \leq i < i + t \leq j \leq m - t \text{ and} \\ & A_{i+1} A_{i+2} \cdots A_{i+t} = A_{j+1} A_{j+2} \cdots A_{j+t} \} . \end{aligned} \tag{84}$$

Now, in terms of the approximately Poisson random variables $W \equiv W(m,t)$ and $W^* \equiv W^*(m,t)$ from (13) and (16) which count all leftmost repeats of t-tuples, with and without allowing self-overlap, for any $t \geq 2, m \geq 2t$ we have

$$\{L_m < t\} = \{W = 0\} \text{ but } \{L_m^* < t\} \neq \{W^* = 0\} . \tag{85}$$

The equality for the first part is obvious. If there were equality in the second part, then for example, with the word TATATATA, with $m = 8, t = 4$ we have $W^* = 0$ while with $m = 8, t = 3$ we have $W^* = 1$, leading to $L_8^* = 3$ instead of the correct $L_8^* = 4$. The failure of equality, for the case of no self-overlap, is due to our choice in (14) to define I^* with the condition $i + t < j$, instead of the $i + t \leq j$ that is natural in (84). [That choice, although clumsy for the purpose of approximating the distribution of L_m^*, simplifies the computation of b_1^* because it creates symmetry among the conditions involved in self-overlap and the conditions involved in the definition (34) of the overlap relation $\alpha \sim \beta$, which have to allow one extra position for the mismatch needed for declumping.]

Like λ, the error bounds for the Chen-Stein method are functions of m and t, for example $\overline{b_1} \equiv \overline{b_1}(m,t)$. This dependence is suppressed from the notation, except in writing out the error bound for the approximation of $\mathbf{P}(L_m = t)$, which involves both the cases (m, t) and $(m, t+1)$.

Theorem 4 *Under the conditions of Theorem 3, for $t \geq 2$ and $m \geq 2t$,*

$$|\mathbf{P}(L_m < t) - e^{-\lambda}| \leq r(m,t) \equiv \frac{1-e^{-\lambda}}{\lambda}(\overline{b_1} + \overline{b_2}). \tag{86}$$

If also $m \geq 2(t+1)$ then

$$|\mathbf{P}(L_m = t) - \left(e^{-\lambda(m,t+1)} - e^{-\lambda(m,t)}\right)| \leq r(m,t+1) + r(m,t). \tag{87}$$

Proof. The first result follows immediately from (85) combined with the Chen-Stein method in Theorem 1. The second result follows from the first using $\mathbf{P}(L_m = t) = \mathbf{P}(L_m < t+1) - \mathbf{P}(L_m < t)$ together with the triangle inequality. ∎

The following corollary requires only the approximation for λ given by (30), together with (80) or (81), which show that the error bounds tend to zero. It says that there is a family of limit distributions involving integerization of extreme value distributions. Define centering constants $c(m)$ by

$$c(m) \equiv \log_{1/p}(m^2(1-p)/2). \tag{88}$$

Corollary 2 *Under the conditions of the previous theorem, for any $T < \infty$, as $m \to \infty$, uniformly in $x \in [-T, T]$ such that $x + c(m) \in Z$,*

$$\mathbf{P}(L_m < x + c(m)) \to \exp(-p^x), \mathbf{P}(L_m = x + c(m)) \to \exp(-p^{x+1}) - \exp(-p^x). \tag{89}$$

For non self-overlapping repeats, there is a correction term needed to account for a consecutive repeats of a t-tuple; these were not included in W^*. Let

$$V = \sum_{0 \leq i \leq m-t} R_{i,i+t}$$

be the number of places where there is a consecutive repeat, so that $\mathbf{P}(V > 0) \leq \mathbf{E}V < mp^t$. We have

$$\{L_m^* < t\} = \{W^* = 0\} \cap \{V = 0\}, \text{ hence } \mathbf{P}(W^* = 0) - mp^t < \mathbf{P}(L_m^* < t) \leq \mathbf{P}(W^* = 0). \tag{90}$$

Apart from this correction term, the following results for non self-overlapping repeats are a direct translation of the previous results for repeats allowing self-overlap.

Theorem 5 *Under the conditions of Theorem 2, for $t \geq 2$ and $m \geq 2t$, with $r^*(m,t) \equiv \frac{1-e^{-\lambda^*}}{\lambda^*}(b_1^* + \overline{b_2^*})$*

$$|\mathbf{P}(L_m^* < t) - e^{-\lambda^*(m,t)}| \leq mp^t + r^*(m,t). \tag{91}$$

If also $m \geq 2(t+1)$ then

$$|\mathbf{P}(L_m^* = t) - \left(e^{-\lambda^*(m,t+1)} - e^{-\lambda^*(m,t)}\right)| \leq r^*(m,t+1) + mp^t + r^*(m,t). \tag{92}$$

Proof. These results follow immediately from (90) combined with the Chen-Stein method in Theorem 1. ∎

The following corollary is proved using the approximation (22) for λ^*, together with (57) or (58), which show that the error bounds tend to zero. It says that L_m and L_m^* have the same limits; i.e. the effects of self overlap show up in approximations and error bounds, but not in limits.

Corollary 3 *Under the conditions of the previous theorem, for any $T < \infty$, as $m \to \infty$, uniformly in $x \in [-T,T]$ such that $x + c(m) \in Z$,*

$$\mathbf{P}(L_m^* < x + c(m)) \to \exp(-p^x), \mathbf{P}(L_m^* = x + c(m)) \to \exp(-p^{x+1}) - \exp(-p^x). \tag{93}$$

In attempting numerical calculations of the point probabilities using (87) or (92), it becomes apparent that in some cases these bounds are crude and can be beat by simple alternatives. For example, the best lower bound on $\mathbf{P}(L_m = t)$ for small t is zero, and the best upper bound comes from $\mathbf{P}(L_m = t) \leq \mathbf{P}(L_m < t+i) \leq \exp(-\lambda(m,t+i)) + r(m,t+i)$ for some choice of $i > 0$. In contrast, for large t, it appears that the upper bound from (87) is always more effective than the alternate bound $\mathbf{P}(L_m = t) \leq \mathbf{P}(L_m \geq t) \leq \lambda(m,t)$. This last upper bound is elementary, not requiring the hard work needed for Poisson approximation: the event $\{L_m \geq t\}$ is equal to the event $\{W(m,t) \geq 1\}$, with $\mathbf{P}(W \geq 1) \leq \mathbf{E}W$; the argument is really just that the probability of a union is at most the sum of the probabilities. If our goal is to bound $\mathbf{P}(L_m \geq t)$, the bound analogous to (86) is $\mathbf{P}(L_m \geq t) \leq (1 - \exp(-\lambda(m,t)) + r(m,t)$, and a priori it is possible that this bound is better than the elementary upper bound, $\lambda(m,t)$. Intuitively, the smallest Chen-Stein bounds occur for the uniform case, so we compared the two bounds numerically for this case. As Table 4 shows, the elementary bound sometimes won. One case can be analyzed: if $m, t \to \infty$ with $\lambda \asymp 1$, then since $(1 - \exp(\lambda(m,t))) - \lambda = \lambda^2/2 - \lambda^3/6 + \cdots$ is bounded away from zero, while $r(m,t) \to 0$, it follows that for sufficiently large m and t, the Chen-Stein bound beats the elementary bound.

In Table 5, we show our approximations, with error bounds, for the distribution of L_m, for $m = 200, 400, 600$ and for $m = 1000, 10^6$ and 10^9. It can be seen that for moderate m, like $m = 200$, we have only some limited success in pinning down the distribution of L_m, while for large m our approximations become very precise. Two of the qualitative implications of Corollary 2 are easily seen in Table 5, namely that after centering the distributions of the L_m are tight, without any rescaling, and that the limit distribution varies with the value of the fractional part of $\log_{1/p} m^2$. Both tables 4 and 5 are for the uniform case on an alphabet of four letters.

Table 4. Upper bounds on $\mathbf{P}(L_m \geq t)$: Chen-Stein versus first order

m	t	$1 - e^{-\lambda} + r(m,t)$	λ	m	t	$1 - e^{-\lambda} + r(m,t)$	λ
200	7	.8921	.8599	200000	17	.5830	.8729
200	8	.2226	.2127	200000	18	.1961	.2182
200	9	.0539	.0526	200000	19	.05310	.05456
200	10	.0132	.0132	200000	20	.01355	.01364
200	11	.0032	.0032	200000	21	.003404	.003409
200	12	$8.06 \cdot 10^{-4}$	$7.97 \cdot 10^{-4}$	200000	22	$8.521 \cdot 10^{-4}$	$8.525 \cdot 10^{-4}$
200	13	$1.99 \cdot 10^{-4}$	$1.97 \cdot 10^{-4}$	200000	23	$2.1313 \cdot 10^{-4}$	$2.1311 \cdot 10^{-4}$
200	14	$4.93 \cdot 10^{-5}$	$4.87 \cdot 10^{-5}$	200000	24	$5.3277 \cdot 10^{-5}$	$5.32783 \cdot 10^{-5}$
200	15	$1.21 \cdot 10^{-5}$	$1.20 \cdot 10^{-5}$				

Table 5 shows the point estimates for $\mathbf{P}(L_m = t)$, using the upper and lower bounds explained above. The point probability estimates are listed together with the intervals for which we guarantee that the point probability lies in. If the interval is symmetric around the point estimate, we use the $\pm$ convention. For example, the entry $.0309 \in [0, .7385]$ for $m = 200, t = 5$ means that .0309 is the approximation $\exp(-\lambda(200,5)) - \exp(\lambda(200,6))$ for $\mathbf{P}(L_{200} = 5)$; the best lower bound is 0 and the best upper bound is .7385. The entry $.3851 \pm .3463$ for $m = 200, t = 7$ means that .3851 is the approximation for $P(L_{200} = 7)$, and we can guarantee this probability to lie in the interval $[.3851 - .3463; .3851 + .3463]$.

Table 5. Approximation and guarantees for $\mathbf{P}(L_m = t)$

t	$\mathbf{P}(L_{200} = t)$	t	$\mathbf{P}(L_{400} = t)$	t	$\mathbf{P}(L_{800} = t)$
5	$.0309 \in [0; 0.7385]$	6	$.0287 \in [0; 0.5973]$	7	$.0274 \in [0; 0.5121]$
6	$.3922 \in [0; 0.7385]$	7	$.3848 \in [0; 0.5973]$	8	$.3804 \in [0; 0.5121]$
7	$.3851 \pm .3463$	8	$.3892 \pm .2017$	9	$.3917 \pm .1144$
8	$.1403 \pm .0337$	9	$.1440 \pm .0195$	10	$.1461 \pm .0109$
9	$.0383 \pm .0029$	10	$.0396 \pm .0016$	11	$.0404 \pm 9.38 \cdot 10^{-4}$
10	$.0097 \pm 3.36 \cdot 10^{-4}$	11	$.0101 \pm 1.83 \cdot 10^{-4}$	12	$.0103 \pm 9.80 \cdot 10^{-5}$
11	$.0024 \pm 5.50 \cdot 10^{-5}$	12	$.0025 \pm 2.89 \cdot 10^{-5}$	13	$.0025 \pm 1.50 \cdot 10^{-5}$
12	$5.99 \cdot 10^{-4} \pm 1.17 \cdot 10^{-5}$	13	$6.31 \cdot 10^{-4} \pm 6.08 \cdot 10^{-6}$	14	$6.48 \cdot 10^{-4} \pm 3.10 \cdot 10^{-6}$

t	$\mathbf{P}(L_{1000} = t)$	t	$\mathbf{P}(L_{10^6} = t)$	t	$\mathbf{P}(L_{10^9} = t)$
				25	$6.88 \cdot 10^{-37} \in [0; 9.19 \cdot 10^{-6}]$
		16	$3.31 \cdot 10^{-10} \in [0; 0.0058]$	26	$.9.10 \cdot 10^{-10} \in [0; 9.19 \cdot 10^{-6}]$
7	$.0035 \in [0; 0.4105]$	17	$.0042 \in [0; 0.0058]$	27	$.0054 \pm 1.15 \cdot 10^{-5}$
8	$.2412 \in [0; 0.4105]$	18	$.2513 \pm .0019$	28	$.2667 \pm 2.81 \cdot 10^{-6}$
9	$.4509 \pm .1841$	19	$.4554 \pm 3.52 \cdot 10^{-4}$	29	$.4500 \pm 4.94 \cdot 10^{-7}$
10	$.2122 \pm .0199$	20	$.2072 \pm 3.56 \cdot 10^{-5}$	30	$.1995 \pm 4.81 \cdot 10^{-8}$
11	$.0622 \pm .0016$	21	$.0606 \pm 2.78 \cdot 10^{-6}$	31	$.0579 \pm 3.62 \cdot 10^{-9}$
12	$.0161 \pm 1.59 \cdot 10^{-4}$	22	$.0157 \pm 2.28 \cdot 10^{-7}$	32	$.0150 \pm 2.79 \cdot 10^{-10}$
13	$.0040 \pm 2.13 \cdot 10^{-5}$	23	$.0039 \pm 2.52 \cdot 10^{-8}$	33	$.0037 \pm 1.79 \cdot 10^{-11}$
14	$.0010 \pm 4.06 \cdot 10^{-6}$	24	$9.98 \cdot 10^{-4} \pm 4.21 \cdot 10^{-9}$	34	$9.52 \cdot 10^{-4} \pm 4.26 \cdot 10^{-12}$

3 Deterministic aspects of unique recoverability

Recall that the l–spectrum of a word $\mathbf{A} = A_1 A_2 \cdots A_m$ is the multiset whose elements are the l-tuples $A_{i+1} A_{i+2} \cdots A_{i+l} \in S^l$, for $i = 0$ to $m - l$. Note that repetitions are allowed, so that this multiset always has cardinality $m - l + 1$, and that the order of these l-tuples is not specified by the spectrum. Two different words can have the same l-spectrum; examples are given below. We say that a word is (uniquely) l-recoverable if no other word has the same l-spectrum. Two basic problems are to give a simple criterion for unique l-recoverability, and to estimate the probability that a randomly chosen word of length m is l-recoverable.

The characterization for l-recoverability which we use was given by Ukkonen and Pevzner. Ukkonen [Ukkonen (1992)] described three simple classes of transformations of words of length m which preserve the l-spectrum, and conjectured that any two words having the same l-spectrum would be connected by a series of such transformations. Pevzner [Pevzner (1995)] proved that conjecture. [A different characterization of l- recoverability was given in Pevzner

[Pevzner (1989)] .] The Ukkonen-Pevzner characterization of l-recoverability is in terms of repeats in the (ordered) list of overlapping t-tuples, with $t = l - 1$, that make up the word $\mathbf{A} = A_1 \cdots A_m$. We denote this list as $B_0, B_1, B_2, \ldots, B_{m-t}$, where $B_i \equiv A_{i+1} \cdots A_{i+t} \in S^t$. [A natural alternate choice of notation, with $B_i \equiv A_i A_{i+1} \cdots A_{i+t-1}$ is not as convenient.] Note that the list $B_0, B_1, B_2, \ldots, B_{m-t}$ and the t-spectrum have exactly the same elements; the difference is that the list is ordered and the spectrum is not.

We give three examples, which will correspond to the three classes of transformations.

Example 1. With $m = 5, l = 4, t = 3$, the word TATAT has 4-spectrum whose two elements are ATAT and TATA (since order is irrelevant, we present this multiset in alphabetical order) and its list of overlapping 3-tuples is TAT,ATA,TAT. Observe that this list has a repeat, the same first and last element. A different word, ATATA, has the same 4-spectrum, but its list of overlapping 3-tuples is ATA,TAT,ATA. [Not only are the two lists different, but so are the underlying multisets, i.e. the words TATAT and ATATA are distinguishable by their 3-spectra but not by their 4-spectra!]

Example 2. With $m = 10, l = 4, t = 3$, the word ACACA TACAG has 4-spectrum {ACAC, ACAG, ACAT, ATAC, CACA, CATA, TACA}, and its list of overlapping 3-tuples is ACA,CAC,ACA,CAT,ATA,TAC,ACA,CAG. Observe that this list has an element, ACA, which occurs three times. A different word, ACATA CACAG, has the same 4-spectrum, but its list of overlapping 3-tuples is ACA,CAT,ATA,TAC,ACA,CAC,ACA,CAG. [The two lists are different only in that they are in different order, i.e. our two words have the same 3-spectrum as well as the same 4-spectrum.]

Example 3. With $m = 11, l = 5, t = 4$, the word ACACA CTCAC A has 5-spectrum {ACACA, ACACT, ACTCA, CACAC, CACTC, CTCAC, TCACA}, and its list of overlapping 4-tuples is ACAC,CACA,ACAC,CACT,ACTC,CTCA,TCAC,CACA. Observe that this list has two elements which occur twice; a pair of ACAC and a pair of CACA, and furthermore these two repeats are interleaved. A different word, ACACT CACAC A, has the same 5-spectrum. [The two words also have the same 4-spectrum.]

As above,we use the notation $B_0, B_1, \ldots, B_{m-t}$ for the list of overlapping t-tuples of an m-word, with $t = l - 1$. We consider the commas as optional, and grouping would be optional too. Thus if a, b are t-tuples and α, β are lists of t-tuples, then as lists $a\alpha b\beta = a, \alpha, b, \beta = (a(\alpha)b)\beta$. [One way to formalize this would be to speak of strings (of t-tuples) and concatenations.]

Theorem 6 *(Pevzner & Ukkonen) Two words* $\mathbf{A}$ *and* $\mathbf{A}'$ *of length* m *have the same* l*-spectrum if and only if they can be transformed into each other by a series of the following operations.*

1. *Rotation.* $\mathbf{A}'$ *is a* rotation *of* $\mathbf{A}$ *if* $B_0 = B_{m-t}$ *and for some* $0 < i < m-t$*, as lists,* $B'_0, B'_1, \ldots, B'_{m-t} = B_i, B_{i+1}, \ldots, B_{m-t-1}, B_{m-t}, B_1, B_2, \ldots, B_{i-i}, B_i$*. Another way to express this is: the list of overlapping* t*-tuples for one word has the form* $a\ \alpha\ b\ \beta\ a$*, and the other word has list* $b\ \beta\ a\ \alpha\ b$*, where* $a, b \in S^t$ *and* $\alpha, \beta \in (S^t)^*$*, i.e.* α *and* β *are lists, possibly empty, of* t*-tuples.*

2. *Transposition with a three way repeat.* $\mathbf{A}'$ *is a* transposition *of* $\mathbf{A}$ *(using a three way repeat at* (i, j, k)*) if, for some* $0 \le i < j < k \le m-t$ *and* $a \in S^t$*, we have* $a = B_i = B_j = B_k$*, and*

$$\begin{aligned} & B'_0, B'_1, \ldots, B'_{m-t} \\ = \ & B_0, \ldots, B_{i-1}, a, B_{j+1}, B_{j+2}, \ldots, B_{k-1}, a, B_{i+1}, \ldots, B_{j-1}, a, B_{k+1}, \ldots, B_{m-t}. \end{aligned}$$

   The alternate expression of this is: the list of overlapping t*-tuples for one word has the form* $\alpha\ a\ \beta\ a\ \gamma\ a\ \delta$*, and the other word has list* $\alpha\ a\ \gamma\ a\ \beta\ a\ \delta$*, where* $a \in S^t$ *and* $\alpha, \beta, \gamma, \delta \in (S^t)^*$*.*

3. *Transpositions with two interleaved pairs of repeats.* $\mathbf{A}'$ *is a* transposition *of* $\mathbf{A}$ *(using two interleaved pairs of repeats) if, for some* $0 \le i < i' < j < j' \le m-t$ *and* $a, b \in S^t$*, we have* $a = B_i = B_j$ *and* $b = B_{i'} = B_{j'}$*, and*

$$\begin{aligned} & B'_0, \ldots, B'_{m-t} = \\ & B_0, \ldots, B_{i-1}, a, B_{j+1}, \ldots, B_{j'-1}, b, B_{i'+1}, \ldots, B_{j-1}, a, B_{i+1}, \ldots, B_{i'-1}, b, B_{j'+1}, \ldots, B_{m-t}. \end{aligned}$$

   The alternate expression of this is: the list of overlapping t*-tuples for one word has the form* $\alpha\ a\ \beta\ b\ \gamma\ a\ \delta\ b\ \epsilon$*, and the other word has list* $\alpha\ a\ \delta\ b\ \gamma\ a\ \beta\ b\ \epsilon$*, where* $a, b \in S^t$ *and* $\alpha, \beta, \gamma, \delta, \epsilon \in (S^t)^*$*.*

Remark. The transformations described above include "trivial" transformations, i.e. cases with $\mathbf{A}' = \mathbf{A}$. Example 1 fits case 1 of Theorem 6 with $a =$TAT, $b =$ATA, $\alpha = \beta = \emptyset$, the empty list. Example 2 fits case 2 with $\alpha = \emptyset, a =$ACA, $\beta =$ CAC, $\gamma =$ CAT,ATA,TAC, and $\delta =$ CAG. Example 3 fits case 3 with $\alpha = \emptyset, a =$ACAC, $\beta = \emptyset, b =$CACA, $\gamma = \emptyset$, $\delta =$CACT,ACTC,CTCA,TCAC, $\epsilon = \emptyset$.

The basis of the proof of the theorem above is the "de Bruijn graph" of the word $\mathbf{A}$, whose vertex set is $\{B_0, B_1, \ldots, B_{m-t}\}$ (as a set, so that repeats and order are irrelevant,) and with $m-t$ directed edges (B_{i-1}, B_i) for $i = 1$ to $m-t$, so that edges are in one to one correspondance to the spectrum. Note that multiple edges can occur. The (ordered) list $B_0, \ldots, B_{m-t}$ is a Eulerian path, i.e. it traverses each edge the number of times prescribed by its multiplicity, and each Eulerian path corresponds to a different word of length m having the same spectrum, so l-recoverability is the same as having a de Bruijn graph with a *unique* Eulerian path.

Our statement of Theorem 6 is precisely what Ukkonen conjectured and Pevzner proved, but there is a slight imprecision in their description of the transformations. For instance, (with q in the role of our l,) Ukkonen wrote "(rotation) If y can be written as $y = z_1\, y_1\, z_2\, y_2\, z_1$ for some $(q-1)$-grams z_1 and z_2 and for some strings y_1 and y_2...." From this statement we initially concluded that self-overlapping repeats need not be considered in analyzing recoverability. However, as our three examples above show, there need not be any strings between the repeated $(q-1)$-grams (our t-tuples), because these repeats may have self-overlap.

It is fairly easy to see from Theorem 6 that unique recoverability can be determined just from the indicators $R_{i,j}, (i,j) \in I$ of repeats of t-tuples (see definition (10)); the part that requires thought is distinguishing trivial from nontrivial transformations. When it comes to the probabilistic analysis of recoverability, counting repeats is much more complicated than counting "leftmost" repeats. Thus our next step is to restate the result of Pevzner and Ukkonen in terms of *leftmost* repeats; this requires considerable work.

Recall that we say a repeat occurs at (i,j) with $0 \le i < j \le m-t$ if for some $a \in S^t, B_i = a = B_j$. It is *not* leftmost if the preceding letters match, so that $B_{i-1} = B_{j-1}$ and there is another repeat at $(i-1, j-1)$. For example, with $m = 17, t = 3$, the word AACGT AGACG TATCG TG has five repeats, at $(2,8), (3,9), (3,14), (4,10)$, and $(9,14)$; there are three leftmost repeats, at $(2,8), (3,14)$ and $(9,14)$.

Let $\mathbf{A}$ be a randomly chosen word from S^m. Recall from (12) in section 2 that for $0 \le i < j \le m, X_{i,j}$ is the indicator that a leftmost repeat occurs at (i,j), namely $X_{i,j} = \mathbf{1}(B_i = B_j)\mathbf{1}(i = 0 \text{ or else } B_{i-1} \ne B_{j-1})$, and $\mathbf{X}$ is the process specifying where all leftmost repeats occur.

Lemma 1 *Whether or not a word $\mathbf{A} \in S^m$ is l-recoverable is measurable through the process $\mathbf{X}$ of indicators of leftmost repeats of t-tuples, with $t = l-1$. Moreover, in terms of the transformations in Theorem 6,* *1. a nontrivial rotation is possible if and only if*

1a) $X_{0,m-t} = 1$*, and there is a* $0 < i < m - t$ *such that* $X_{0,i} = 0$*.*

2. A nontrivial transposition using three way repeats is possible if and only if there are $0 \leq i < j < k \leq m - t$ *such that either*

2a) $X_{j,k} = 1$ *and (*$X_{i,j} = 1$ *or* $X_{i,k} = 1$ *or both), or*

2b) $X_{j,k} = 0$ *and* $X_{i,j} = 1$ *and* $X_{i,k} = 1$ *and, with* d *the greatest common divisor of* $j - i$ *and* $k - j$*, it is NOT the case that*

$$X_{i,k-d} = 1, 0 = X_{k-d,k} = X_{k-d-1,k-1} = \cdots = X_{i+2,i+d+2} = X_{i+1,i+d+1}. \tag{94}$$

3. A nontrivial transposition using two interleaved pairs of repeats is possible if

3a) $X_{i,j} = 1$ *and* $X_{i',j'} = 1$ *with* $0 \leq i < i' < j < j' \leq m - t$*.*

Conversely, if there is a nontrivial transposition using two interleaved pairs of repeats, then 3a) or there is a nontrivial transposition using a three way repeat.

The overall measurability may be described as:

$$\mathbf{1}(\mathbf{A} \textit{ is } l\textit{-recoverable}) = h(\mathbf{X}) \equiv \mathbf{1}(\textit{none of } 1a, 2a, 2b, 3a). \tag{95}$$

Remark. This lemma shows that having a nontrivial rotation is measurable with respect to $\mathbf{X}$, and that having a nontrivial transposition using a three way repeat is measurable with respect to $\mathbf{X}$, but it does *not* prove the analogous property for nontrivial transpositions using two way repeats. In fact, this last property is not $\mathbf{X}$-measurable. As an example, for $m = 15, t = 2$ consider the words GGCAT TGGCA TAGGT and GGAAT CGGCT TAGGT. Both have $X_{0,6} = X_{0,12} = X_{6,12} = 1$ and all other $X_{i,j}$ are zero. The first word has a nontrivial transposition with interleaved repeats, $a = AT$ at $(i, j) = (3, 9)$, and $b = GG$ at $(i', j') = (6, 12)$. The second word doesn't have two interleaved pairs of repeats. Note the process $(R_\alpha)_{\alpha \in I}$ has ones at $(0, 6), (0, 12), (1, 7), (2, 8), (3, 9), (6, 12)$ for the first word, and only at $(0, 6), (0, 12), (6, 12)$ for the second word; this shows that the process (R_α) carries strictly more information than the process $\mathbf{X} = (X_\alpha)$.

Proof: For item 1, the condition $X_{0,m-t} = 1$ implies that $a \equiv B_0 = B_{m-t}$, and the condition $X_{0,i} = 0$ implies that $b \equiv B_i$ satisfies $a \neq b$, so that the rotation is nontrivial. Conversely, if there is a rotation, then $B_0 = B_{m-t}$, and nontriviality rules out the case $B_0 = B_1 = B_2 = \cdots = B_{m-t}$.

For item 2, the only easy implication is that given part 2a), it follows that $B_i = B_j = B_k$ and $A_{j-1} \neq A_{k-1}$, hence the transposition using the three way repeat at (i, j, k) is nontrivial.

To see where the dichotomy a) versus b) in part 2 arises, we must discuss *leftmost* three way repeats. We say that (i, j, k) with $0 \leq i < j < k \leq m - t$ is a repeat if $B_i = B_j = B_k$, and that such a repeat is leftmost if $i = 0$ or ($i \geq 1$ and $(i-1, j-1, k-1)$ is not a repeat). Observe that, if (i, j, k) and $(i-1, j-1, k-1)$ are both repeats, then the transposition at one triplet is nontrivial if and only if the transposition at the other triplet is nontrivial. Thus, there is a nontrivial tranposition using a three way repeat if and only if there is a nontrivial transposition using a leftmost three way repeat. Next, observe that for a leftmost repeat (i, j, k), it is not possible that all three of the indicators $X_{i,j}, X_{i,k}$, and $X_{j,k}$ are zero. Finally, for a repeat (i, j, k) it is never possible that exactly two of the three indicators are zero. To see this, there are three cases; one of which is $X_{i,j} = X_{i,k} = 0, X_{j,k} = 1$. In that case $B_i = B_j$ and $X_{i,j} = 0$ implies $i \geq 1$ and $A_i = A_j$; similarly $A_i = A_k$, and hence $A_j = A_k$, which contradicts ($B_j = B_k$ and $j \geq 1$ and $X_{j,k} = 1$). The other two cases are similar. The net result of this paragraph is that we need only consider leftmost repeats, and for these, at most one of the three indicators $X_{i,j}, X_{i,k}$, and $X_{j,k}$ is zero.

For repeats (i, j, k) we need a necessary and sufficient condition for nontriviality of the transposition that can be expressed in terms of indicators X. This paragraph will derive such a condition, in terms periodicity, namely that with $d = \gcd(j - i, k - j)$, the list $w \equiv B_i, B_{i+1}, \ldots, B_k$, as a word of length $k - i + 1$ over the alphabet S^t, has the form $w = \rho^{(k-i)/d}a$, where $a \in S^t$ and $\rho \in (S^t)^d$ with a being the first element of the d-tuple ρ. To establish our necessary and sufficient condition, observe first that, in the alternate notation of Theorem 6, we want a criterion for when $\alpha a \ \beta a \ \gamma \ a \ \delta = \alpha \ a \ \gamma \ a \ \beta \ a \ \delta$, which is obviously equivalent to $a \ \beta \ a \ \gamma = a \ \gamma \ a \ \beta$. Here, the list $a\beta$ has length $j - i > 0$, and the list $a\gamma$ has length $k - j > 0$. Thus, we may relabel our objective: ($a\beta$ becomes σ, $a\gamma$ becomes $\tau, j - i$ becomes m, $k - j$ becomes n) we need to prove, for strings σ, τ of lengths $m, n \geq 1$ with $d = \gcd(m, n)$, that the concatenations $\sigma\tau = \tau\sigma$ if and only if there is a string ρ of length d such that $\sigma = \rho^{m/d}$ and $\tau = \rho^{n/d}$. One implication is obvious; for the other we start with the assumption $\sigma\tau = \tau\sigma$. In case $m = n$ then easily $\sigma = \tau$, so we are done, with $d = m, \rho = \sigma$. Otherwise $m \neq n$, and without loss of generality we may assume $m < n$. It follows that τ has the form $\tau = \sigma\tau'$, where τ' has length $n' \equiv n - m > 0$. Stripping off the initial σ from both sides of $\sigma\tau = \tau\sigma$, i.e. from $\sigma(\sigma\tau') = (\sigma\tau')\sigma$ we get $\sigma\tau' = \tau'\sigma$. Since $\gcd(n', m) = \gcd(m, n) = d$ and $n' + m < m + n$, we are done, appealing to induction on $m+n$; in detail there is a string ρ of length d with $\sigma = \rho^{m/d}, \tau' = \rho^{n'/d}$, hence $\tau = \sigma\tau' = \rho^{n/d}$.

[The foundation of this induction are the cases with $m = n$.]

Next we argue that in condition b), given the first line, the remaining conditions on the indicators X are precisely that the transposition be nontrivial. Thus, we assume $X_{j,k} = 0$ and $X_{i,j} = X_{i,k} = 1$, and show that the transposition is trivial if and only if it IS the case that (94), i.e. $X_{i,k-d} = 1, 0 = X_{k-d,k} = X_{k-d-1,k-1} = \cdots = X_{i+1,i+d+1}$. ¿From the preceding paragraph, a transposition is trivial if and only if it satisfies the periodic condition $B_i, B_{i+1}, \ldots, B_k = \rho^{(k-i)/d}a$, where $a = B_i = B_j = B_k$ and $d = \gcd(j-i, k-j)$. Using this condition, it is obvious that if the transposition is trivial, then the condition (94) is satisfied. Next we assume that (94) is satisfied, and show why the periodic condition is satisfied. Put $\rho \equiv B_{k-d}, B_{k-d+1}, \ldots, B_{k-1}$. The periodic condition is that $\rho = B_{k-cd}, B_{k-cd+1}, \ldots, B_{k-(c-1)d-1}$, for $c = 2, 3, \ldots, (k-i)/d$. This is equivalent to: for $s = 0, 1, \ldots, d, B_{i+s} = B_{i+d+s} = B_{i+2d+s} = \cdots = B_{k-d+s}$. Now from $X_{i,k-d} = 1$ it follows that $B_{k-d} = a = B_k$, and from $X_{k-d,k} = 0$, we obtain that $B_{k-d-1} = B_{k-1}$. Therefore $X_{k-d-1,k-1} = 0$ implies $B_{k-d-2} = B_{k-2}$. Iterating this establishes the periodic structure.

We have completed the proof of item 2; here is a summary. The previous paragraph shows that, if 2b) holds, then there is a nontrivial transposition at (i,j,k). We already observed that 2a) trivially implies a nontrivial transposition at (i,j,k). For the converse, that a nontrivial transposition implies the disjunction 2a) or 2b), we start with a nontrivial transposition, and shift left to get a leftmost repeat (i,j,k) with a nontrivial transposition. By "leftmost", at most one of the three indicators $X_{i,j}, X_{i,k}, X_{j,k}$ is zero. If none are zero, 2a) follows. If exactly one is zero, then if $X_{j,k} \neq 0$, then 2a) again holds, while if $X_{j,k} = 0$ then the first line of 2b) holds, and using the previous paragraph, nontriviality of the transposition implies (94), the remaining part of 2b).

Finally, we consider item 3. If 3a), then from $X_{i',j'} = 1$ and $i' > 0$ it follows that $B_{i'-1} \neq B_{j'-1}$. This shows that the transposition is nontrivial; in the notation of item 3 of Theorem 6, $\beta \neq \delta$, because they end in different elements of S^t.

Conversely, suppose there is a nontrivial transposition using two interleaved pairs of repeats, say with $i < i' < j < j'$ giving the locations of the $\cdots a \cdots b \cdots a \cdots b \cdots$. In the alternate notation of item 3 of Theorem 6, write $B_0, B_1, \ldots, B_{m-t} = \alpha\ a\ \beta\ b\ \gamma\ a\ \delta\ b\ \epsilon$, so that nontriviality is equivalent to the condition $\beta\ b\ \gamma\ a\ \delta \neq \delta\ b\ \gamma\ a\ \beta$. [Note, nontriviality implies $\beta \neq \delta$, but the converse is false, e.g. with $\delta = \beta b \gamma a \beta$, we have $\beta \neq \delta$, but the transposition *is* trivial.]

We have assumed that $B_i = B_j$ and $B_{i'} = B_{j'}$, so whether or not 3a) holds is a question of the repeats at (i,j) and (i',j') are leftmost. If both are, we are done. If neither is, then both can be shifted one place left, i.e. there is an interleaved pair of repeats at $(i-1, j-1)$ and $(i'-1, j'-1)$, and it is easily checked that the transposition using this new pair is also nontrivial. After iterating this left shift, we either get to new values $0 \le i < i' < j < j'$ for which both repeats are leftmost, or else case 1) $X_{i,j} = 0, X_{i',j'} = 1$, or else case 2) $X_{i,j} = 1, X_{i',j'} = 0$.

Case 1 is easier. Shift back (i, j) by one until either 3a) is satisfied (with the new values), or else $j = i'$. In this latter case, we have a three way repeat at $(i, j = i', j')$ with $X_{i',j'} = 1$, and the transposition using this three way repeat is nontrivial (as given by 2a) with (i, i', j') playing the role of (i, j, k).

Now suppose case 2. Shift back (i', j') until either 3a) is satisfied (with the new values), or either $i' = i$ or else $j' = j$. Note that both $i' = i$ and $j' = j$ cannot be simultaneously satisfied, because then $\beta = \gamma$, and this is excluded by nontriviality.

First assume $i' = i$. Then we have a three way repeat at $(i = i', j, j')$. It is not obvious to us that the transposition using this three way repeat must be nontrivial. If the original transformation, using the pair of interleaved repeats, transforms $\mathbf{A}$ to $\mathbf{A}'$, and the new transformation, using the three way repeat, transforms $\mathbf{A}$ to $\mathbf{A}''$, then, remarkably, it can be shown that $\mathbf{A}'' = \mathbf{A}'$, hence the new transformation using a three way repeat is nontrivial. To check this, we use the alternate notation from Theorem 6. The list of overlapping t-tuples for $\mathbf{A}$ has the form

$$w \equiv B_0, B_1, \ldots, B_{m-t} = \alpha\ a\ \beta\ b\ \gamma\ a\ \delta\ b\ \epsilon,$$

and the list for $\mathbf{A}'$ is

$$w' \equiv \alpha\ a\ \delta\ b\ \gamma\ a\ \beta\ b\ \epsilon.$$

That the repeat (i', j') can be shifted back until $i = i'$ while maintaining $j < j'$ implies that $\delta = \delta' a \beta$ (where the list δ' is possibly empty). Thus, with parentheses added to display the transposition,

$$w' = \alpha a(\delta' a \beta) b \gamma a(\beta) b \epsilon.$$

Writing w again in this form, with parentheses added to show where a transposition with a three way repeat will apply,

$$w = \alpha\ a(\beta b \gamma) a(\delta') a\ \beta\ b\ \epsilon,$$

so that the list w'' for $\mathbf{A}''$ is

$$w'' = \alpha\ a(\delta') a(\beta b \gamma) a\ \beta\ b\ \epsilon.$$

The grouping indicated by parentheses is not part of the list, and one can now easily see that $w = w''$.

The case where (i', j') is shifted back until $j = j'$ while $i < i'$ can be handled by a similar argument, that the resulting transposition using three way repeat is a nontrivial transformation, because in fact it has the same effect as the transposition using two interleaved pairs repeats.

□

4 Probability approximations for unique recoverability

Recall, we always assume $2 \le t = l - 1 \le m/2$.

We begin by repeating a little more carefully the overview, from the introduction, of the probabilistic analysis of unique recoverability from the l-spectrum. The key is to examine where there are repeats of t-tuples within the sequence, with $t = l - 1$. Repeats come in clumps, clumps correspond to leftmost repeats of t-tuples, and according to Lemma 1, knowledge of where the leftmost repeats occur is also sufficient to decide whether or not a given word of length m is l-recoverable. In terms of where the leftmost repeats occur, there are three ways that unique recoverability can be spoiled, corresponding to the three items in Lemma 1. The first two of these, corresponding to rotations and three way repeats, are extremely unlikely. The dominant contribution, given as item 3a), is to have an interleaved pair of leftmost repeats. The total number of leftmost repeats is random, with a distribution close to Poisson(λ), where λ is the expected number of leftmost repeats, given by formula (23). More importantly, the process $\mathbf{X}$ giving the locations of the repeats is close to a Poisson process $\mathbf{Y}$, with close to constant intensity, so that conditional on having k leftmost repeats, the $2k$ coordinates specifying locations for the matching t-tuples are distributed approximately as $2k$ independent integers chosen uniformly from $\{0, 1, 2, \ldots, m - t\}$. Finally, provided there are no duplicates in a list of $2k$ independent uniforms, their relative order is a permutation of $2k$ objects, with all $(2k)!$ possibilities equally likely. Of these $(2k)!$ permutations, exactly $2^k k! C_k$ correspond to having no interleaved pair; here $C_k \equiv \frac{1}{k+1}\binom{2k}{k}$ is the k-th Catalan number. Thus the relative fraction, $2^k k! C_k/(2k)! = 2^k/(k+1)!$, approximates the probability of unique recoverability, conditional on having k pairs of leftmost repeats. Averaging with respect to the Poisson distribution, $\mathbf{P}(k \text{ repeats }) \approx e^{-\lambda}\lambda^k/k!$, yields that the probability of unique recoverability is approximately $f(\lambda) = \sum_{k \ge 0} e^{-\lambda}\lambda^k/k! \;\; 2^k/(k+1)!$.

In the above analysis, the error for each approximation can be controlled. Our outline is

$$\begin{aligned} \mathbf{P}(l - \text{recoverable }) &= \mathbf{E}h(\mathbf{X}) \\ &\approx \mathbf{E}g_1(\mathbf{X}) \quad \text{error bound } R_1 \qquad \text{eliminate rotations} \end{aligned}$$

$$
\begin{array}{rll}
\approx \mathbf{E}g_1(\mathbf{Y}) & \text{error bound } R_2 & \text{Poisson process approximation} \\
\approx \mathbf{E}g_1(\mathbf{Z}) & \text{error bound } R_3 & \text{symmetrize intensities} \\
\approx \mathbf{E}g(\mathbf{Z}) & \text{error bound } R_4 & \text{eliminate three way repeats} \\
= \mathbf{P}(D) & & \text{(unfolding; coupling to i.i.d. positions)} \\
\approx f(\lambda) & \text{error bound } R_5 & \text{tie breaking, Catalan numbers.}
\end{array}
$$

The first approximation step is to eliminate consideration of rotations; in terms of Lemma 1, g_1 is the indicator that none of (2a,2b, or 3a) occurs. The error bound R_1 is simply the probability that the random word of length m is nonconstant but begins and ends with the same t-tuples, $R_1 = p^t - p_m$.

The second approximation step is to replace the process $\mathbf{X}$ of indicators of leftmost repeats in the random word by a Poisson process $\mathbf{Y}$ having exactly the same intensity. Here $\mathbf{X}$ has a very complicated dependence structure, while $\mathbf{Y}$ has independent coordinates. The error bound, from the Chen-Stein method in section 3, is $R_2 = \bar{b}_1 + \overline{b_2}$, where $\bar{b}_1$ is given by (64) and $\overline{b_2}$ is given by (79). In case the bound (60) from Corollary 1 is smaller, we use this better bound as the value of R_2.

The third approximation step is to slightly change the intensity of the Poisson process, to get a new Poisson process $\mathbf{Z}$ which corresponds to independent coordinates chosen uniformly from $\{0, 1, \ldots, m-t\}$. The overall intensity λ is unchanged. The original process $\mathbf{Y}$ has three sources of nonuniformity: the effects of self-overlap (which are absent when the independent letters $A_1, A_2, \ldots$ are chosen *uniformly*,) the lack of the declumping factor $1-p$ at points (i, j) with $i = 0$, and the restriction that no repeat can occur at (i, j) with $i = j$. The error bound is R_3 given by the sum over all i, j of the absolute value of the change in intensity at the point (i, j).

The next approximation is to eliminate consideration of three way repeats; in terms of Lemma 1, g is the indicator that 3a) does not occur. The error bound R_4 is an upper bound on the probability of a three way repeat, which is both smaller and easier working with $\mathbf{Z}$ instead of $\mathbf{X}$. Notice, for eliminating the effects of rotations in step 1, the opposite was true; there, $\mathbf{X}$ was more tractable than $\mathbf{Z}$.

The equality $\mathbf{E}g(\mathbf{Z}) = \mathbf{P}(D)$ just switches notation from expectations of an indicator functional to the probability of an event. This lets us avoid notational complications in constructing $\mathbf{Z}$ in terms of other processes. We realize the process $\mathbf{Z}$ by picking the total number K of points according to a Poisson distribution with parameter λ, and then, on the event $\{K = k\}$, taking $2k$ independent uniform random picks from $\{0, 1, \ldots, m-t\}$ to be the coordinates of those points. This well known construction of (constant intensity) Poisson processes provides a handle on all $(2k)!$ permutations being equally likely.

The last error, bounded by R_5, arises because our $2k$ independent uniform choices were from a discrete set rather than a continuum, so that ties occur with small but nonzero probability. When ties occur, linear ordering doesn't determine a unique permutation.

[Alternate strategies are available. A small change would be to bound the effects of rotations and three way repeats togther, as the first step. A more substantial change would be to eliminate all of section 2.5, at the price of looser error bounds, by the following simpler outline. Start with the value of $\lambda - \lambda^*$ from (23), or a simple upper bound on it, such as (28), namely $\lambda - \lambda^* \leq mt(\xi_*)^t$. For the processes $\mathbf{X}, \mathbf{X}^*$, the total variation distance is bounded by $\lambda - \lambda^*$, so the preliminary step is that $\mathbf{E}h(\mathbf{X}) \approx \mathbf{E}h(\mathbf{X}^*)$ with error at most $\lambda - \lambda^*$. Then proceed to follow the outline above. In the second step, the error bound now comes from the Chen-Stein method applied to $\mathbf{X}^*$, which is simpler. It also shrinks the error, since $b_1^* < \overline{b_1}$ and $\overline{b_2^*} < \overline{b_2}$. With our alternate strategy, the error bound R_3 for symmetrizing the intensities is larger by about $2(\lambda - \lambda^*)$. Thus the net effect is to change the overall error bound by about $-(\overline{b_1} + \overline{b_2} - (b_1^* + \overline{b_2^*})) + 3(\lambda - \lambda^*)$. Since $3(\lambda - \lambda^*) > 2(\lambda - \lambda^*)$, the alternate strategy compares unfavorably with our original strategy using the option of Corollary 1 for bounding R_2.]

0. A rigorous setup. We now make the above argument precise, proceeding according to the steps in the outline. The first step is to "lift" the functional h defined at (95), so that the functionals h, g_1, and g all are defined on a space large enough to serve as the set of possible values of the process $\mathbf{Z}$. We replace the choice of index set (9) used in previous sections, $I = \{(i,j) : 0 \leq i < j \leq m - t\}$, by the slightly larger set

$$I = \{(i,j) : 0 \leq i \leq j \leq m - t\}. \tag{96}$$

We extend the process $\mathbf{X}$ of indicators of leftmost repeats to live on (the new) I, by setting all $X_{i,i} = 0$.

Next, since the coordinates of $\mathbf{X}$ are $\{0,1\}$–valued, while the coordinates of $\mathbf{Y}$ and $\mathbf{Z}$ are $Z_+ \equiv \{0, 1, 2, \ldots\}$–valued, we need to extend the definition of the functional h. That is, we need to define $h : Z_+^I \to \{0,1\}$, compatible with (95) , and tractable when applied to the processes $\mathbf{Y}$ and $\mathbf{Z}$. For $\mathbf{x} \in Z_+^I$, [so that for $0 \leq i \leq j \leq m - t$, $x_{i,j} \in \{0, 1, 2, \ldots\}$], we define 4 conditions, motivated by Lemma 1:

1a) $x_{0,m-t} \geq 1$, and there is $0 < i < m - t$ such that $x_{0,i} = 0$.

2a) There are $0 \leq i < j < k \leq m - t$ such that $x_{j,k} \geq 1$ and ($x_{i,j} \geq 1$ or $x_{i,k} \geq 1$ or both).

2b) There are $0 \leq i < j < k \leq m-t$ such that $x_{j,k} = 0$ and $x_{i,j} \geq 1$ and $x_{i,k} \geq 1$ and, with d the greatest common divisor of $j-i$ and $k-j$, it is NOT the case that $[x_{i,k-d} \geq 1, 0 = x_{k-d,k} = x_{k-d-1,k-1} = \cdots = x_{i+2,i+d+2} = x_{i+1,i+d+1}]$.

3a) There are $0 \leq i < i' < j < j' \leq m-t$ with $x_{i,j} \geq 1$ and $x_{i',j'} \geq 1$.

Our new definition of the functional h is that $h(\mathbf{x}) = 0$ if 1a),2a),2b), or 3a) is true, and $h(\mathbf{x}) = 1$ otherwise:

$$h(\mathbf{x}) = \mathbf{1}(\text{ none of: } \text{1a),2a),2b),3a) }).$$

With these (new) definitons of I and X and h, Lemma 1 still applies, so that $h(\mathbf{X})$ is the indicator that our randomly chosen word $A_1A_2\cdots A_m$ is l-recoverable. This justifies the opening equality in (96).

1. Eliminate rotations. Eliminating rotations from consideration means that we define the functional $g_1 : Z_+^I \to \{0,1\}$ by copying the definition of h, but omitting the condition 1a):

$$g_1(\mathbf{x}) = \mathbf{1}(\text{ none of: } \text{2a),2b),3a) }).$$

When applied to $\mathbf{X}$, we have $|h(\mathbf{X}) - g_1(\mathbf{X})| \leq \mathbf{1}(\mathbf{X} \text{ satisfies 1a) })$. Note that, $\mathbf{X}$ satisfies 1a) if and only if $A_1A_2\cdots A_t = A_{m-t+1}\cdots A_t$ and not $(A_1 = A_2 = \cdots = A_m)$, so that

$$|\mathbf{E}h(\mathbf{X}) - \mathbf{E}g_1(\mathbf{X})| \leq \mathbf{P}(\mathbf{X} \text{ satisfies 1a) }) = p^t - p_m \equiv R_1. \tag{97}$$

For asymptotics as $m, t \to \infty$ with $\lambda \asymp 1$, $R_1 \sim p^t \asymp m^{-2}$.

2. Poisson process approximation. Since the functional g_1 takes values in $[0,1]$, the difference between expectations of g_1 applied to two processes is bounded by the total variation distance between the processes. Thus the bound

$$|\mathbf{E}g_1(\mathbf{X}) - \mathbf{E}g_1(\mathbf{Y})| \leq d_{TV}(\mathbf{X}, \mathbf{Y}) \leq R_2, \tag{98}$$

where R_2 is the minimum of the two bounds, either (60) from Corollary 1, or $\overline{b_1} + \overline{b_2}$ from (80) and (81) in Theorem 3. We have from Theorem 3 that $\overline{b_1} + \overline{b_2} \asymp \log m \; / \; m$ in the uniform case, and $\overline{b_1} + \overline{b_2} \asymp m^{-\gamma}$ in the nonuniform case.

3. Symmetrize intensities. The next step is to change from the Poisson process $\mathbf{Y}$ on I to another Poisson process $\mathbf{Z}$ on I, also having independent coordinates. Recall, the intensities of $\mathbf{Y}$ are the values $\mathbf{E}Y_{i,j} \equiv \mathbf{E}X_{i,j} \equiv p_{i,j}$ as given by formulas (19) and (26); note for $i = j$ we have $X_{i,j} \equiv 0$ so $p_{i,j} = 0$. The overall intensity is $\lambda = \sum_{0 \leq i \leq j \leq m-t} p_{i,j}$, with value computable using (23). Our goal is to use a constant intensity point process on the square $[0, m-t]^2$, which has $(m-t+1)^2$ points, and then for $i < j$ to map both points (i, j) and (j, i) to the point $(i, j) \in I$. This motivates the definition

$$\begin{aligned} \lambda_{i,j} \equiv \mathbf{E} Z_{i,j} \quad & \equiv \tfrac{\lambda}{(m-t+1)^2} \qquad \text{for } (i,j) \in I, i = j \\ \equiv \mu \equiv & \tfrac{2\lambda}{(m-t+1)^2} \qquad \text{for } (i,j) \in I, i < j. \end{aligned} \tag{99}$$

For two Poisson random variables, such as $Y_{i,j}$ and $Z_{i,j}$, the total variation distance is at most the absolute value of the difference in their expectations. For two processes, each having independent coordinates, the total variation distance is at most the sum of the total variation distances between corresponding coordinates. As before, since g_1 is an indicator functional, total variation distance gives an upper bound on the distance between expectations. Thus

$$|\mathbf{E} g_1(\mathbf{Y}) - \mathbf{E} g_1(\mathbf{Z})| \leq d_{TV}(\mathbf{Y}, \mathbf{Z}) \leq \sum_{(i,j) \in I} |p_{i,j} - \lambda_{i,j}| \equiv R_3. \tag{100}$$

An exact expression for R_3 is easy in the uniform case, and is comparable to (23) in the nonuniform case. For asymptotics as $m, t \to \infty$ with $\lambda \asymp 1$, we have $R_3 = O(1/m)$ in the uniform case, and $R_3 = O(\log m \ / \ m)$ in the nonuniform case.

4. Eliminate three way repeats. The next step is eliminating three way repeats from consideration. Thus we define the functional $g : Z_+^I \to \{0, 1\}$ by

$$g(\mathbf{x}) = \mathbf{1}(\text{ not 3a) }).$$

For the difference with g_1, we have $\{\mathbf{x} : \ g_1(\mathbf{x}) \neq g(\mathbf{x})\} \subset \{\mathbf{x} : \ \text{2a) or 2b) }\} \subset \cup_{0 \leq i \leq m-t} C_i$ where C_i is the event that for the "corner" at $(i, i), \mathbf{x}$ is greater than zero at two or more different locations. The "corner" at (i, i) is defined here as $\{(k, j) \in I : k = i \text{ or else } j = i\}$. Notice that the point (i, i) is excluded from the corner, which contains $m - t$ points, and hence there are $\binom{m-t}{2}$ ways to pick two different locations in the corner at i, regardless of the choice of $0 \leq i \leq m - t$. Our process $\mathbf{Z}$ has independent coordinates, and intensity $\mu \equiv \frac{2\lambda}{(m-t+1)^2}$ at the points which make up corners, so for a particular pair of points, the probability that $\mathbf{Z}$ is nonzero at those points is at most μ^2. Thus $\mathbf{P}(\mathbf{Z} \in C_i) \leq \binom{m-t}{2}\mu^2$. Summing over the $m - t + 1$ choices for i we get

$$|\mathbf{E} g_1(\mathbf{Z}) - \mathbf{E} g(\mathbf{Z})| \leq (m-t+1)\binom{m-t}{2}\mu^2 < \frac{2\lambda}{m-t} \equiv R_4. \tag{101}$$

5. Catalan numbers There are two considerations here. First a counting argument is needed to establish a connection between Catalan numbers and interleaved pairs of repeats;

there is no probability error introduced in this step. Second, a tie breaking argument, which accounts for the error term R_5, is necessary.

The deterministic connection between Catalan numbers and interleaved pairs was correctly given by [Dyer, Frieze, Suen (1994)], but for the sake of completeness, we present the argument here. Suppose there are $2k$ distinct objects, and k distinct labels, with each label assigned to exactly two objects. [In $a = B_i = B_j$ with $a \in S^t, i, j \in \{0, 1, \ldots, m-t\}$, the objects are i and j and the label is a; this correspondence is rough since the objects are not necessarily distinct, and neither are the labels.] There are $(2k)!$ different ways to linearly order the objects. Let G_k be the subset consisting of those linear orders which do not have any pair of interleaved labels $\cdots a \cdots b \cdots a \cdots b \cdots$. The claim is that $|G_k| = 2^k k! C_k$, where the kth Catalan number, $C_k = \frac{1}{k+1}\binom{2k}{k}$, is the cardinality of the set F_k of "well formed formulas" using k left parentheses '(' and k right parentheses ')'. Recall, a string of k '(' and k ')' corresponds to a well formed formula if and only if, reading from left to right, the count of '(' is never exceeded by the count of ')'. There is a $2^k k!$ to one correspondence between G_k and F_k, namely given a linear order in G_k, written out as a string, for each of the k labels a replace the first occurrence of a by '(' and the second by ')'. It is easy to check that this is a map to F_k; this uses only the two- to-one labeling of the $2k$ objects, and does not use the "no interleaved pairs" property of elements of G_k. To see that each element of F_k has exactly $k!2^k$ preimages in G_k, the key observation is that in a well formed formula there is a unique matching of the '(' and ')' such that there are no interleaved matching pairs; this is part of the standard combinatorial treatment. The factor $k!$ comes from assigning the k labels to the k matching pairs, and the factor 2^k comes from choosing, for each of the k labels, for the two objects sharing that label, which comes first in the linear order.

Finally, we need a probabilistic approximation to connect having k pairs of repeats with the situation of having all $(2k)!$ linear orders equally likely. This cannot (as far as we know) be done directly with the original setup of repeating t-tuples in a sequence of i.i.d. letters $A_1, A_2, \ldots, A_m$, even if the letters are uniformly distributed over a finite alphabet S. Even after passing to the Poisson process $\mathbf{Z}$, the following additional argument is needed. Let $K, L_1, L_2, \ldots$ be independent, with K being Poisson distributed with parameter λ and L_i is uniform in the set $\{0, 1, 2, \ldots, m-t\}$. On the event $\{K = k\}$ form k points $Q_1, Q_2, \ldots, Q_k$ from the $2k$ coordinates $L_1, \ldots, L_{2k}$ via $Q_i = (L_{2i-1}, L_{2i})$. [Note, neither the L_i nor the Q_i need to be distinct.] This yields a Poisson point process $\mathcal{P}$ on the square $\{0, 1, \ldots, m-t\}$ with intensity $\lambda/(m-t+1)^2$ at each point of the square; we think of $\mathcal{P}$ as a random set of points. The square is "folded" up into the index set I defined by (96), via the map

$\phi((x,y)) = (x,y)$ if $x \le y$ and $\phi((x,y)) = (y,x)$ if $y < x$, and folding the constant intensity Poisson process on the square yields the Poisson process $\mathbf{Z}$ whose intensity is given by (99). In terms of this construction, $g(\mathbf{Z})$ is the indicator of the event D that $\mathcal{P}$ does not have a pair of points Q, Q' such that, with $(i,j) = \phi(Q), (i',j') = \phi(Q'), i < i' < j < j'$. [A priori, two points of $\mathcal{P}$ may coincide in location, but the further condition $i \ne i'$ rules this out.] In shorthand, the complementary event D^c is defined by

$$D^c = \{\exists Q, Q' : i < i' < j < j'\}. \tag{102}$$

Define another event B by changing only the strictness of the inequalities:

$$B^c = \{\exists Q, Q' : i \le i' \le j \le j'\}. \tag{103}$$

The obvious dilemma here is whether the Catalan argument applies with D or with B; the resolution lies in between. To break ties between the coordinates $L_1, L_2, L_3, \ldots$ we introduce independent continuously distributed "marks" $U_1, U_2, \ldots$; to be concrete let U_i take values in $[0,1]$. We denote the "augmented" coordinates by $\hat{L}_i \equiv (L_i, U_i)$. We use $\prec$ to denote the natural lexicographic total order on the space of possible values of augmented coordinates, i.e. for $x, y \in \{0, 1, \ldots, m-t\}, u, v \in [0,1]$ define $(x,u) \prec (y,v)$ if and only if $((x < y)$ or else $(x = y, u < v))$. With augmented points $\hat{Q}$ and with folding defined in terms of $\prec$, we define an event C by changing the definition (102) to include the marks:

$$C^c = \{\exists \hat{Q}, \hat{Q}' : \hat{i} \prec \hat{i}' \prec \hat{j} \prec \hat{j}'\}. \tag{104}$$

Since for two augmented coordinates (x,u) and (y,v), we have deterministically that $x < y$ implies $(x,u) \prec (y,v)$ which further implies $x \le y$, it follows that

$$\{\exists Q, Q', i < i' < j < j'\} \subset \{\exists \hat{Q}, \hat{Q}', \hat{i} \prec \hat{i}' \prec \hat{j} \prec \hat{j}'\} \subset \{\exists Q, Q', i \le i' \le j \le j'\},$$

i.e. $D^c \subset C^c \subset B^c$, so that we see $B \subset C \subset D$. For C, the distinction between $\prec$ and $\preceq$, corresponding to the distinction between D and B, is immaterial, since ties for the augmented coordinates have probability zero. On the event $\{K = k\}$ we have with probability one that there are $2k$ distinct augmented coordinates, and since $(L_1, U_1), \ldots, (L_{2k}, U_{2k})$ are i.i.d., each of the $(2k)!$ possible linear ordering occurs with probability $1/(2k)!$. Now by the Catalan argument, $\mathbf{P}(C|K = k) = 2^k k! C_k/(2k)! = 2^k/(k+1)!$. Averaging over the distribution of K yields an exact relation:

$$\mathbf{P}(C) = \sum \mathbf{P}(K = k)\mathbf{P}(C|K = k) = \sum_{k \ge 0} e^{-\lambda} \frac{\lambda^k}{k!} \frac{2^k}{(k+1)!} \equiv f(\lambda) \tag{105}$$

To connect this with $\mathbf{P}(D)$, consider the event E that the $2K$ coordinates contain a duplicate:

$$E = \{|\{L_1, \ldots, L_{2K}\}| < 2K\}.$$

Since $B \subset D \subset \ B \cup E$, we have $|\mathbf{P}(B) - \mathbf{P}(D)| \leq \mathbf{P}(E)$. For an upper bound on $\mathbf{P}(E)$ we use the simplest upper bound for the birthday problem: when there are j people independently and uniformly picking birthdays from a year with n days, the probability of a coincidence is at most the expected number of (unordered) pairs of people sharing a birthday, i.e. $\binom{j}{2}/n$. Here $n = m - t + 1$ and $j = 2K$ with $\mathbf{E}\binom{2K}{2} = 2\lambda^2 + \lambda$. Thus $\mathbf{P}(E) \leq \frac{2\lambda^2+\lambda}{m-t+1}$, which we take for R_5. This justifies the last step in our outline, with

$$|\mathbf{P}(D) - f(\lambda)| = |\mathbf{P}(D) - \mathbf{P}(C)| \leq \mathbf{P}(E) \leq \frac{2\lambda^2 + \lambda}{m - t + 1} \equiv R_5. \tag{106}$$

We summarize this as

Theorem 7 *Assume that $A_1, A_2, \ldots$ are independent and identically distributed. Let $l \geq 3, t = l - 1, m \geq 2t$. Let $p \equiv p_2 \equiv \mathbf{P}(A_1 = A_2)$ and more generally, for $r = 2, 3, \ldots,\ p_k \equiv \mathbf{P}(A_1 = A_2 = \cdots = A_k)$. With λ given by (23) and f defined by (105),*

$$|\mathbf{P}(A_1 A_2 \cdots A_m \textit{ is } l\textit{-recoverable }) - f(\lambda)| \leq R \equiv R_1 + \cdots + R_5. \tag{107}$$

For $m, t \to \infty$ with $\lambda \asymp 1$, which is equivalent to $m, l \to \infty$ with $l = \frac{\log m^2}{\log(1/p)} + O(1)$, we have $R = O(\log m\ /\ m)$ in the uniform case, and $R = O(m^{-\gamma})$ in the nonuniform case; the dominant contribution being the term R_2 from the Poisson process approximation.

Remark Consider that $f(\lambda(m,t))$ is just a function of m and t, designed to approximate the probability of unique recoverability, with the crucial property that as $m, t \to \infty$ with $\lambda \asymp 1$, the approximation error tends to zero. There are lots of plausible alternate expressions. For example, it is easy to see that the function $f(\lambda)$ is Lipschitz continuous with Lipschitz constant 1. Therefore we could replace λ in (107) by an approximation of λ that has an easier form, e.g $\lambda' = (1 - p)p^t m^2/2$, and a bound similar to (107), provided we include an additional error bound R_6 such that $|\lambda' - \lambda| \leq R_6$; see (28). For another example, we might try to account for the probability of rotations or three way repeats. Since there is no theory to say how these might be dependent on the presence or absence of an interleaved pair of repeats, the simplest heuristic is to assume independence, and use

$$F(m,t) \equiv f(\lambda)(1 - p^t + p^m) \exp\left(-\binom{m-t}{3}(1-p)(1-p^2)p^{2t}\right).$$

Here we have assumed a uniform distribution; the second factor corresponds to rotations, which is item 1a) in Lemma 1; the last factor is a Poisson approximation for the probability of having no three way repeats, matching only condition 2a) of Lemma 1. Now for $m, t \to \infty$ with $\lambda \asymp 1$, the difference between $F(m,t)$ and $f(\lambda)$ goes to zero, but it is plausible that for small m and t, $F(m,t)$ is the better approximation to the true, unknowable probability of unique recoverability. At the level of (107) however, we would have a worse error bound. For the example $t = 7, m = 180$ cited in [Pevzner *et al.* (1991)] to have approximately a 95% probability of unique recoverability, we have $f(\lambda) = .9368$ and $F(m,t) = .9347$. Thus the change from $f(\lambda)$ to $F(m,t)$ is both too small and also in the wrong direction, for explaining the difference between their 95% and our $\approx 93\frac{1}{2}\%$.

For the next two tables, recall that $R \equiv R_1 + R_2 + R_3 + R_4 + R_5$ is our upper bound on the error in the approximation of the probability of unique recoverability by the function $f(\lambda)$. Tables 6 and 7 show the contributions of the different error bounds to R in the uniform case and in the "strongly nonuniform" case $p_A = .3544, p_C = .1430, p_G = .1451, p_T = .3575, p = .2949$. The by far largest term is the error bound R_2, coming from the Chen-Stein method. We pick some round values for m, t, and some values m, t so that λ is close to .60646. The values are truncated, not rounded.

Table 6: The error bounds in the uniform case

m	$t = l - 1$	λ	R_1	R_2	R_3	R_4	R_5
100	7	.2015	$6.103 \cdot 10^{-5}$	.0565	.0042	.0043	.0030
200	7	.8599	$6.103 \cdot 10^{-5}$	.4701	.00886	.00891	.0120
800	9	.8969	$3.814 \cdot 10^{-6}$	.1579	.002264	.002267	.0031
169	7	.6458	$6.068 \cdot 10^{-5}$	.2808	.00744	.00749	.00824
659	9	.6059	$3.814 \cdot 10^{-6}$	.0883	.001861	.001864	.00205
2615	11	.6066	$2.38 \cdot 10^{-7}$	.0266	.00046	.00046	.00051
10430	13	.6064	$1.49 \cdot 10^{-8}$	.00779	.000116	.000116	.000128

Table 7: The error bounds in the nonuniform case

m	$t = l - 1$	λ	R_1	R_2	R_3	R_4	R_5
100	7	.6436	$1.940 \cdot 10^{-4}$	2.542	.0081	.0138	.0156
100	9	.0573	$1.687 \cdot 10^{-5}$	.0809	.0127	.0012	$6.946 \cdot 10^{-4}$
321	9	.6069	$1.68 \cdot 10^{-5}$	.8828	.0494	.0038	.00429
1081	11	.6068	$1.467 \cdot 10^{-6}$	.2964	.0272	.0011	.0012
3663	13	.6066	$1.276 \cdot 10^{-7}$	.0963	.0136	.00033	.00036

The next tables illustrate the approximation for the probability of unique recoverability in the uniform case and in the nonuniform case $p_A = .3544, p_C = .1430, p_G = .1451, p_T = .3575, p = .2949$. The third column reports the difference $f(\lambda) - R$, which serves as a lower bound on the probability of unique recoverability and is thus a "guaranteed value"; in contrast the value $f(\lambda)$ in the fifth column serves as a "prediction" with no guarantee. The values for (m, t) chosen for tables 8 and 9 include some round values of m; some values of m chosen so that there is a guaranteed 95%, 90%, or 50% probability of unique recoverability; some values of m chosen so that approximation $f(\lambda)$ is close to 95%, 90%, or 50%; and finally for the uniform case, $t = 7, m = 180$ and $t = 11, m = 2450$, which were given in [Pevzner *et al.* (1991)] to have a 95% probability of unique recoverability in simulations.

Table 8: **The probability of unique recoverability in the uniform case**

m	$t = l - 1$	$f(\lambda) - R$	λ	$f(\lambda)$	R	R_2
100	7	.9254	.2015	.9936	.0682	.0565
200	7	.4076	.8599	.9077	.5000	.4701
800	9	.7351	.8969	.9008	.1656	.1579
85	7	.9525	.1422	.9967	.0442	.0346
110	7	.9019	.2467	.9906	.0887	.0755
111	7	.8993	.2515	.9903	.0909	.0776
188	7	.5097	.7567	.9260	.4162	.3890
469	9	.9503	.3037	.9861	.0357	.0320
586	9	.9005	.4776	.9676	.0671	.0621
983	9	.5011	1.359	.8053	.3041	.2933
2288	11	.9501	.4638	.9693	.0191	.0179
2814	11	.9002	.7028	.9349	.0347	.0331
4735	11	.5004	1.995	.6613	.1609	.1571
9988	13	.95004	.5560	.9572	.0071	.0068
12208	13	.90005	.8311	.9129	.0129	.0124
20909	13	.5001	2.440	.5634	.0633	.0621
169	7	.6458	.6068	.9499	.3040	.2808
205	7	.3612	.9048	.8993	.5381	.5070
353	7	-2.113	2.753	.4986	2.612	2.518
659	9	.8559	.6059	.9500	.0941	.0883
2615	11	.9218	.6066	.9499	.0281	.0266
3186	11	.8500	.9017	.8999	.0499	.0480
5553	11	.2414	2.746	.4999	.2584	.2532
10430	13	.9418	.6064	.9500	.0081	.0077
180	7	.5707	.6916	.9368	.3660	.3406
2450	11	.9372	.5322	.9605	.0232	.0219

Table 9: **The probability of unique recoverability in the nonuniform case**

m	$t = l - 1$	$f(\lambda) - R$	λ	$f(\lambda)$	R	R_2
50	7	.4498	.1504	.9964	.5465	.5020
100	8	.5311	.1912	.9942	.4630	.4235
200	9	.6360	.2340	.9915	.3554	.3219
400	10	.7310	.2804	.9880	.2570	.2298
800	11	.8032	.3327	.9834	.1902	.1587
1600	12	.8533	.3929	.9774	.1240	.1076
17	7	.9556	.0123	.99997	.0442	.0360
18	7	.9492	.0143	.09996	.0506	.0416
47	7	.5224	.1317	.9972	.4747	.4339
69	9	.9507	.0267	.9998	.0491	.0401
233	9	.5018	.3184	.9848	.4829	.4430
378	11	.9501	.0746	.99909	.0489	.0393
1224	11	.5008	.7776	.9224	.4215	.3877
2173	13	.95002	.2141	.9928	.0428	.0345
5704	13	.5002	1.4685	.7810	.2808	.2579
97	7	-1.523	.6043	.9503	2.474	2.367
321	9	.0093	.6069	.9499	.9405	.8828
1081	11	.6239	.6068	.9499	.3260	.2964
3663	13	.8392	.6066	.9499	.1106	.0963
7804	13	-.0807	2.7468	.4999	.5807	.5485
26470	15	.3041	2.7466	.5000	.1958	.1811
89783	17	.4323	2.7466	.5000	.0676	.0610

5 Final discussion

Our primary motivation has been to give error bounds for approximations to the probability of unique recoverability, and along the way to give careful bounds for Poisson approximations for long repeats, with and without allowing self-overlap, for sequences of i.i.d. letters, both with and without a uniform distribution. The corresponding limit theorem for the probability of unique recoverability for the uniform case, without error bounds, was given in [Dyer, Frieze, Suen (1994)]. For the nonuniform case, bounds on the expected number of self-overlapping repeats, as in our (27) and (28), are essential.

The values we report in tables 8 and 9, including guarantees on the probability of unique recoverability, can be compared to those reported in [Pevzner *et al.* (1991)], which are based on Monte Carlo simulation. For small l, such as $l = 8$, our error bounds are rather large, but our error bounds are rigorous; in contrast simulation values have associated confidence intervals, which are random and vary with the simulation. For only slightly larger l, such as $l = 12$ or 16, our theoretical error bounds are quite satisfactory. From tables 6,7,8, and 9 it is clear that the major source of error is R_2, from the Poisson process approximation. We have treated the Poisson approximation as carefully as possible at present; but as the remark in section 2.1 explains, it is conceivable that our upper bounds could be improved by a factor growing like $\log m$. It is tempting to hope for such an improvement.

A natural extension of the Poisson process analysis in this paper would be to address questions of partial recovery. For example, what is the distribution of the length M of the longest contiguous substring of a target of length m which can be uniquely reconstructed? In this paper, we have approximated only $\mathbf{P}(M = m)$. It should even be feasible to describe approximately the joint distribution of the lengths of all fragments that can be determined from the spectrum. For another approach, consider the random variable N counting the number of sequences of length m which have the same l-spectrum as $A_1 A_2 \cdots A_m$. In this paper, we have approximated only $\mathbf{P}(N = 1)$, but it may be possible to handle the distribution of N.

For applications, the most drastically unrealistic feature of our model is the assumption that the *multiset* of l-tuples can be read from a target sequence; information on multiplicities is not available in the laboratory. Assuming that only the *set* of l-tuples were known, it is plausible, from the structure of the de Bruijn graph, that there would still be a high probability that the multiset could be reconstructed. If this is so, then the above analysis, together with one additional error term, might serve to predict and bound the probability that a random target sequence of length m could be uniquely reconstructed from the set of l-tuples it contains. The issues of partial recovery and set versus multiset are addressed in [Arratia, Reinert (1996)].

For the theoretical understanding of physical sequencing by hybridization, it would be good, but difficult, to analyze some probability model where the given data are generated from the l-spectrum *with errors*. In applications, both false positives and false negatives can occur, for reporting whether or not given l-tuples are present in the target sequence.

Acknowledgement. We would like to again [Arratia, Gordon, Waterman (1986), p. 993] thank S. Karlin for explicitly asking, in the analysis of long matches, not only *how many* occur but also *where* do they occur.

References

[Aldous (1989)] D. Aldous, *Probability approximations via the Poisson clumping heuristic.* Springer, New York.

[Arratia, Goldstein, Gordon (1989)] R. Arratia, L. Goldstein and L. Gordon, Two moments suffice for Poisson approximations: The Chen-Stein method. *Ann. Probab.* **17** (1989), 9–25.

[Arratia, Goldstein, Gordon (1990b)] R. Arratia, L. Goldstein and L. Gordon, Poisson Approximation and the Chen-Stein Method. *Statistical Science* **5** (1990), 403–434.

[Arratia, Gordon, Waterman (1986)] R. Arratia, L. Gordon and M.S. Waterman, An extreme value theory for sequence matching. *Ann. Statist.* **14** (1986), 971–993.

[Arratia, Gordon, Waterman (1990a)] R. Arratia, L. Gordon and M.S. Waterman, The Erdös-Rényi law in distribution. for coin tossing and sequence matching. *Ann. Statist.* **18** (1990), 539–570.

[Arratia, Reinert (1996)] R. Arratia and G. Reinert, Poisson process approximation for repeats in one sequence and its application to sequencing by hybridization. *Combinatorial Pattern Matching; 7th Annual Symposium, CPM 96, Lecture Notes in Computer Science* **1075** (1996), 209–219.

[Arratia, Tavaré (1993)] R. Arratia and S.Tavaré, Reviews of *Probability approximations via the Poisson clumping heuristic* by D. Aldous and *Poisson approximation* by A.D. Barbour, L. Holst, and S. Janson. *Ann. Probab.* **21** (1993), 2269–2279.

[Arratia, Waterman (1985)] R. Arratia and M. S. Waterman, An Erdös-Rényi law with shifts. *Advances in Mathematics* **55** (1985), 13–23.

[Bains, Smith (1988)] W. Bains and G.C. Smith, A novel method for DNA sequence determination. *J. Theor. Biol.* **135**, (1988), 303–307.

[Barbour, Holst, Janson (1992)] A.D. Barbour, L. Holst, S. Janson, *Poisson approximation.* Clarendon, Oxford 1992.

[Chen (1975)] L.H.Y. Chen, Poisson approximation for dependent trials. *Ann. Probab.* **3** (1975), 534–545.

[Drmanac, Crkvenjakov (1987)] R. Drmanac and R. Crkvenjakov, *Yugoslav Patent Application 570.* (1987).

[Dyer, Frieze, Suen (1994)] M. Dyer, A. Frieze and S. Suen, The Probability of Unique Solutions of Sequencing by Hybridization. *J. Comp. Biol.* **1** (1994), 105–110.

[Fodor *et al.* (1991)] S.P.A. Fodor, J.L. Read, M.S. Pirrung, L. Stryer, A.T. Lu and D. Solas, Light-directed spatially addressable parallel chemical synthesis. *Science* **251**, (1991) 767–773.

[Fodor *et al.* (1993)] S.P.A. Fodor, R.P. Rava, X.C. Huang, A.C. Pease, C.P. Holmes and C.L. Adams, Multiplex biochemical assays with biological chips. *Nature* **364** (1993), 555–556.

[Graham, Knuth, Patashnik (1987)] R.L. Graham, D.E. Knuth and O. Patashnik, *Concrete Mathematics.* Addison-Wesley, 1988.

[Karlin, Ost (1987)] S. Karlin and F. Ost, Counts on long aligned word matches along random letter sequences. *Adv. Appl. Prob.* **19** (1987), 293–351.

[van Lint, Wilson (1992)] J.H. van Lint and R.M. Wilson, *A Course in Combinatorics.* Cambridge University Press 1992.

[Lysov *et al.* (1988)] Yu.P. Lysov, V.L. Florent'ev, A.A. Khorlin, Khrapko, V.V. Shik and A.D. Mirzabekov, DNA Sequencing by hybridization with oligonucleotides. A novel method. *Dokl. Acad. Sci USSR* **303** (1988), 1508–1511.

[Macevicz (1989)] S.C. Macevicz, *International Patent Application PS US89 04741* (1989).

[Novak (1995)] S.Yu.Novak, Long match patterns in random sequences. *Siberian Adv. Math.* **5** (1995), 128–140.

[Pease *et al.* (1994)] A.C. Pease, D. Solas, E.J. Sullivan, M.T. Cronin, C.P. Holmes and S.P.A. Fodor, Oligonucleotide arrays for for rapid DNA sequence analysis. *Proc. Natl. Acad. of Sci. USA* **91** (1994), 5022–5026.

[Pevzner (1989)] P.A. Pevzner, l-tuple DNA Sequencing: Computer Analysis. *Journal of Biomolecular Structure and Dynamics* **7** (1989), 63–73.

[Pevzner (1995)] P.A. Pevzner, DNA physical mapping and alternating Eulerian cycles in colored graphs. *Algorithmica* **13** (1995), 77–105.

[Pevzner *et al.* (1991)] P.A. Pevzner, Yu.P. Lysov, K.R. Khrapko, A.V. Belyavsky, V.LO. Florentiev and A.D. Mirzabekov, Improved chips for sequencing by hybridization. *Journal of Biomolecular Structure & Dynamics* **9** (1991), 399–410.

[Pevzner, Lipshutz (1994)] P.A. Pevzner and R.A. Lipshutz, Towards DNA Sequencing Chips. *19th Symposium on Mathematical Foundations of Computer Science, Kosice, Slovakia, Lecture Notes in Computer Science* **841** (1994), 143–158.

[Reinert (1996)] G. Reinert, Probabilistic aspects of sequence repeats and sequencing by hybridization. To appear in *Proceedings of the 20th Annual Meeting of the Gesellschaft für Klassifikation; Studies in Classification, Data Analysis, and Knowledge Organization.* Springer 1996.

[Southern (1988)] E. Southern, *United Kingdom Patent Application GB8810400.* (1988).

[Ukkonen (1992)] E. Ukkonen, Approximate string-matching with q-grams and maximal matches. *Theoretical Computer Science* **92** (1992), 191–211.

[Waterman (1995)] M.S. Waterman, *Introduction to Computational Biology: Maps, Sequences and Genomes.* Chapman Hall 1995.

[Zubkov, Mikhailov (1974)] A.M. Zubkov and V.G. Mikhailov, Limit distributions of random variables associated with long duplications in a sequence of independent trials. *Theory Prob. Applications* **19** (1974), 172–179.