Adv. Appl. Prob. **38,** 116–133 (2006)
Printed in Northern Ireland

EXACT DISTRIBUTION OF WORD COUNTS IN SHUFFLED SEQUENCES

EINAR ANDREAS RØDLAND,* *Rikshospitalet–Radiumhospitalet HF, University of Oslo*

Abstract

In DNA sequences, specific words may take on biological functions as marker or signalling sequences. These may often be identified by frequent-word analyses as being particularly abundant. Accurate statistics is needed to assess the statistical significance of these word frequencies. The set of shuffled sequences – letter sequences having the same k-word composition, for some choice of k, as the sequence being analysed – is considered the most appropriate sample space for analysing word counts. However, little is known about these word counts. Here we present exact formulae for word counts in shuffled sequences.

Keywords: Sequence shuffling; Markov chain; word count; exact distribution; hypergeometric distribution; generalised hypergeometric series; moment generating function; genome sequence analysis; directed graph; Euler path

2000 Mathematics Subject Classification: Primary 60C05
Secondary 05A15; 60J10; 60J20; 62E15

1. Introduction

Analyses of DNA sequence composition, in particular the identification of frequent words, require statistical models against which actual findings may be compared. The model generally applied to this kind of analysis is the Markov chain, in which the DNA sequence is modelled as a sequence over the letters A, C, G, and T corresponding respectively to the four bases adenine, cytosine, guanine, and thymine.

Markov models have been extensively studied both in the statistical literature and, more recently, in bioinformatics. One problem with using a Markov model is that the transition probabilities are unknown. An alternative point of view is that the Markov chain of order $k-1$ produces sequences with a k-word composition that is representative of, though not exactly the same as, the sequence being analysed. This shortcoming may be overcome by restricting the model to sequences which have exactly the same k-word composition, a method introduced in [4]: these are the shuffled sequences, also referred to as the permutation model or constrained (or conditional) Markov chain.

There seems to be a general acceptance that shuffled sequences are more appropriate to analyses of word frequencies than are the representative sequences produced by Markov chains; the importance of this was pointed out in [1]. However, less is known about word counts in shuffled sequences. Expressions for both the expected number of words in shuffled sequences and the variance are known [2], [7]. Theoretical results on exact distributions, however, as presented in [8], refer to Markov chains rather than shuffled sequences, and assessing these is computationally demanding even for lower-order models. Since Bonferroni corrections of

Received 6 July 2004; revision received 23 November 2005.

* Postal address: Institute of Medical Microbiology, Centre for Molecular Biology and Neuroscience, University of Oslo, Rikshospitalet–Radiumhospitalet HF, N-0027 Oslo, Norway. Email address: e.a.rodland@medisin.uio.no

https://doi.org/10.1239/aap/1143936143 Published online by Cambridge University Press

large-scale searches for over-represented words may require the correction of P-values by several powers of 10, accurate P-values are required even in the extreme tails: even if the central part of the distribution looks Gaussian, the tails may deviate significantly.

Previous analyses based on Markov chains have indicated that normal approximations are applicable only for words that are quite frequent (e.g. occurring more than 500 times; see [9]). For less frequent words, the Poisson distribution is recommended. This recommendation seems appropriate for Markov chains or when k is small compared to the length of the words being analysed. However, if k is greater than this, the variance of word counts in shuffled sequences is lower than that of a Poisson distribution [7]. For example, if k is maximal (one less than the word length), a hypergeometric distribution, the variance of which is substantially less than that of a Poisson distribution with the same expectancy, would be more appropriate.

Here we present (in Theorem 2) exact formulae for word counts in shuffled sequences: the distributions, upper tail probabilities, and the momentum generating functions. These distributions generalise the hypergeometric distributions and are expressed in terms of generalised hypergeometric series. We present worked examples to illustrate the computational approach and suggest approximations that may be used in large-scale analyses.

2. The space of thoroughly shuffled sequences

Only cyclic sequences are analysed. Linear sequences and sets of linear sequences may be encoded as cyclic sequences, e.g. using a special character to separate them, so this provides a sufficiently general model.

Definition 1. A *cyclic sequence* of length n over an *alphabet* $\mathcal{A}$ is an element $x \in \mathcal{S}_n = \mathcal{A}^{\mathbb{Z}_n}$, where $\mathbb{Z}_n$ denotes integers modulo n, i.e. $x = (x_i)_{i\in\mathbb{Z}_n}$. The set of all cyclic sequences is $\mathcal{S} = \bigcup_{n=1}^{\infty} \mathcal{S}_n$. A *linear sequence* is an element of $\mathcal{A}^n$, for some n.

For DNA sequences, $\mathcal{A}_{\text{nucl}} = \{\text{A, C, G, T}\}$ may be used; for proteins, the set of amino acids; for protein coding sequences, $\mathcal{A}_{\text{CDS}} = \mathcal{A}_{\text{nucl}} \times \mathbb{Z}_3$; or the alphabet may have separate subsets of letters for different regions such as introns and exons.

Definition 2. A *word* is an element of $\mathcal{A}^* = \bigcup_{p=0}^{\infty} \mathcal{A}^p$. For a sequence $x \in \mathcal{S}_n = \mathcal{A}^{\mathbb{Z}_n}$, the *subword* of length p starting at position $i \in \mathbb{Z}_n$ is the word

$$x_{[i,i+p-1]} := x_i x_{i+1} \cdots x_{i+p-1} \in \mathcal{A}^p.$$

The *word count* is a map $N\colon \mathcal{S} \times \mathcal{A}^* \to \mathbb{N}_0$ defined for $x \in \mathcal{S}_n$ and $w \in \mathcal{A}^p$ by

$$N_x(w) := |\{i \in \mathbb{Z}_n : x_{[i,i+p-1]} = w\}|,$$

where the modulus of a set denotes its cardinality. The *p-word count* $N^{(p)}\colon \mathcal{S} \times \mathcal{A}^p \to \mathbb{N}_0$ is the restriction of N to p-words. For a sequence $x \in \mathcal{S}_n$, we use $N_x^{(p)}\colon \mathcal{A}^p \to \mathbb{N}_0$ to describe the *p-word composition* of x. Similarly, N_w denotes word counts in $w \in \mathcal{A}^p$.

Here let us clarify some of the notation to be used below. For any set U, e.g. $U = \mathcal{A}^p$, the set of maps $U \to \mathbb{N}_0$ is denoted $\mathbb{N}_0^U$. This may equivalently be thought of as the set of lists or vectors $a = (a_u)_{u\in U}$, $a_u \in \mathbb{N}_0$. In some cases, the elements $u \in U$ will be used to represent a vector basis of $\mathbb{N}_0^U$, and a thus written as $\sum_{u\in U} a_u u$. For $a, b \in \mathbb{N}_0^U$ we let $a + b = (a_u + b_u)_{u\in U}$, and for real numbers r and s we let $r + sa = (r + sa_u)_{u\in U}$. Finally, we write $a! = \prod_{u\in U} a_u!$, $a^b = \prod_{u\in U} a_u^{b_u}$ (with $0^0 = 1$), and $a^{\langle b\rangle} = a!/(a-b)!$ (the last of which is referred to as the *falling factorial*).

https://doi.org/10.1239/aap/1143936143 Published online by Cambridge University Press

It may be noted that a $(k-1)$th-order Markov process over $\mathcal{A}$ corresponds to an order-1 Markov process over $\mathcal{A}^{k-1}$; hence, results for order-1 processes generally apply also to higher-order processes. The same applies to sequence shuffling; hence, results on letter permutations preserving transition counts, such as those presented in [2], generalise to shufflings preserving k-word counts, as was pointed out in [11].

Definition 3. For a k-word composition $N_x^{(k)} \in \mathbb{N}_0^{\mathcal{A}^k}$ for some $x \in \mathcal{S}$, the set of *thoroughly shuffled sequences* with this composition is $\mathcal{S}_x \equiv \mathcal{S}(N_x^{(k)}) = \{y \in \mathcal{S} \colon N_y^{(k)} = N_x^{(k)}\}$. As a probability space, the uniform probability distribution is used. We refer to this as *shuffling of order* $k-1$, though it is also commonly referred to as the *permutation model* or *constrained Markov chain*, as it is equivalent to a Markov chain of order $k-1$ constrained to this specific k-word count.

The uniform probability distribution is the natural choice; it may be derived either as the restriction of the Markov chain probabilities, or as the probability produced by the shuffling procedure described in [6].

For a randomly shuffled sequence $X \in \mathcal{S}_x$, the word count $N_X(w)$ of a word w in X is a stochastic variable with values in $\mathbb{N}_0$.

3. Graph formulation of sequence shuffling

Two different languages and sets of theories have been used in analysing shuffled sequences: that of sequences and Markov chains, and that of Euler paths on directed graphs (digraphs). The k-word composition is represented by a directed graph whose edges represent k-words and whose vertices represent $(k-1)$-words. The number of edges corresponding to a given k-word will equal the number of occurrences of that k-word (see Figure 1). For our purposes, graph theory is the most convenient language in which to discuss problems of sequence shuffling; although most results could be translated into a purely sequence-based language, this would require some technical tricks and reformulations that we would rather avoid.

The use of Euler paths on directed graphs to analyse Markov chains goes back to Dawson and Good [3] and Goodman [5], who also related Whittle's formula [14] for counting shuffled sequences to the BEST (de Bruijn–van Aardenne-Ehrenfest–Smith–Tutte) theorem [13] for counting Euler paths; they provided conditional probabilities $\mathrm{P}[N_X^{(p)} \mid N_X^{(k)}]$, but these are not very helpful in analysing individual words. The connection to Euler paths was also made by Fitch [4] and further elaborated upon, with respect to sampling random shufflings, in [1] and [6].

Since graph terminology is not fixed, we briefly review the relevant terms and definitions and refer to the main results used. For a more thorough explanation of Euler paths and the BEST theorem in relation to random sequences, we refer the reader to [15].

Definition 4. A *digraph* G has a set V of *vertices*, a set E of *edges*, and a map

$$\varepsilon \equiv (\varepsilon^-, \varepsilon^+) \colon E \to V \times V$$

indicating that $e \in E$ is an edge from vertex $\varepsilon^-(e)$ to vertex $\varepsilon^+(e)$. A *word digraph* also has maps $\alpha \colon V \to \mathcal{A}^*$ and $\alpha \colon E \to \mathcal{A}^*$, indicating that vertices and edges represent words, such that if e is an edge from v to v' then the word $\alpha(e)$ begins with $\alpha(v)$ and ends with $\alpha(v')$.

We will focus on *balanced digraphs*, in which, for each vertex $v \in V$, the number of edges ending at v equals the number of edges beginning at v. This number is the *degree*, $N_G^V(v)$, of v. The *edge count*, $N_G^E(w)$, of a word w is the number of edges e with $\alpha(e) = w$.

https://doi.org/10.1239/aap/1143936143 Published online by Cambridge University Press

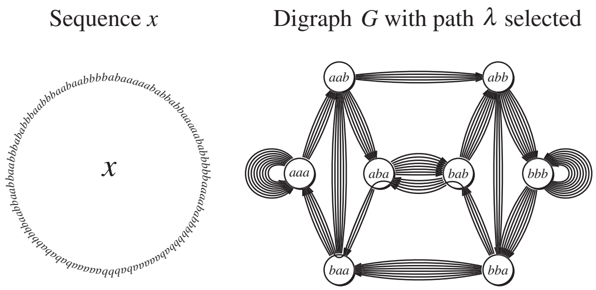

FIGURE 1: All 4-words of x are counted. The digraph G is constructed with vertices representing the 3-words and such that the number of edges from one 3-word to another represents the number of 4-words in x starting and ending with the respective 3-words. A path λ in G, corresponding to the word *bbabaaa*, has been picked. This path is replaced by an edge $\bar{\lambda}$ to form G_λ.

A *path* is a series of distinct edges $\gamma = (\gamma_1, \ldots, \gamma_r)$, $\gamma_i \in E$, such that $\varepsilon^+(\gamma_i) = \varepsilon^-(\gamma_{i+1})$. Let $\varepsilon^-(\gamma) = \varepsilon^-(\gamma_1)$ and $\varepsilon^+(\gamma) = \varepsilon^+(\gamma_r)$ denote the initial and final vertices of the path The word, $\alpha(\gamma)$, formed by the path is made by combining the words $\alpha(\gamma_1), \ldots, \alpha(\gamma_r)$ with overlaps $\alpha(\varepsilon^+(\gamma_i)) = \alpha(\varepsilon^-(\gamma_{i+1}))$.

There is a digraph, unique up to isomorphism, representing each particular k-word composition. In the following, $x \in \mathcal{S}$ will be a fixed sequence and $N_x^{(k)} \in \mathbb{N}_0^{\mathcal{A}^k}$ its k-word composition.

Definition 5. The word digraph *representing* the k-word composition $N_x^{(k)}$ of x, denoted $G_x \equiv G(N_x^{(k)})$, has $V = \{w \in \mathcal{A}^{(k-1)} \colon N_x(w) > 0\} \subset \mathcal{A}^{(k-1)}$ and $\alpha \colon E \to \mathcal{A}^k$ with edge count $N_G^E(w) = N_x(w)$ for $w \in \mathcal{A}^k$. For example, we may use

$$E = \{(w, i) \in \mathcal{A}^k \times \mathbb{N} \colon i \leq N_x(w)\}, \qquad \alpha(w, i) = w.$$

This digraph is balanced and connected in the sense that there is a path between any pair of vertices. The connectedness follows from x being a cyclic sequence. Figure 1 illustrates this construction for the 4-word count of a given sequence.

If we have a set of linear sequences instead of a cyclic sequence, we can add an extra vertex to the set V defined above, and let this special vertex indicate sequence start and end. In addition to the edges defined by the k-word counts, we must then add edges from this special vertex to all the sequence starts, and to it from all sequence ends.

Definition 6. An *Euler path* is a *closed path*, i.e. a path $\gamma = (\gamma_i)_{i \in \mathbb{Z}_n}$ with $\varepsilon^-(\gamma) = \varepsilon^+(\gamma)$, containing all the edges of E exactly once. The map α then maps γ to a cyclic sequence $\alpha(\gamma) \in \mathcal{S}_n = \mathcal{A}^{\mathbb{Z}_n}$. Let $\mathcal{E}_G$ denote the set of Euler paths in G, and $\mathcal{E}_{G,e}$ the set of Euler paths γ with $\gamma_0 = e$.

An *Euler cycle* is the equivalence class of Euler paths modulo the choice of starting point, i.e. $\gamma \sim \gamma'$ if $\gamma_i = \gamma'_{i+j}$ for some $j \in \mathbb{Z}_n$. The set of Euler cycles in G is denoted $\tilde{\mathcal{E}}_G$, and $\tilde{\gamma}$ denotes the equivalence class of γ.

There are exactly n Euler paths for each Euler cycle, one for each choice of starting edge; thus, $\tilde{\mathcal{E}}_G \cong \mathcal{E}_{G,e}$. Using Euler cycles, however, avoids some technicalities related to the enumeration of edges. Note that this use of the term cycle is not standard: closed path and cycle are often used interchangeably.

The correspondence between Euler paths, $\mathcal{E}_G$, and shuffled sequences, $\mathcal{S}_x$, allows us to translate questions concerning random shufflings into questions about random Euler cycles. In particular, the uniform distribution on $\mathcal{S}_x$ is consistent with using the uniform distribution on $\mathcal{E}_G$, which gives the uniform distribution on $\tilde{\mathcal{E}}_G$.

The number of Euler cycles is given by the BEST theorem [13], which in turn relies on the matrix-tree theorem [12]. Let $A \in \mathbb{N}_0^{V \times V}$ be the matrix whose component A_{uv}, for each u and v, is the number of edges from u to v (the *adjacency matrix*), let D be the diagonal matrix with the vertex degrees on the diagonal, and let $L = D - A$ be the *Laplacian* matrix. The *cofactors* of L, i.e. the determinants of the matrices obtained after removing one row and one column from L, are all the same (as $1 \cdot L = 0$ and $L \cdot 1^{\top} = 0$, where 1 here denotes a row vector with unit components and '$\top$' denotes transpose) and are jointly denoted by $\operatorname{cof}(L)$. We can thus write

$$|\tilde{\mathcal{E}}_G| = \tau_G N_G^V! := \tau_G \prod_{v \in V} N_G^V(v)!, \quad \text{where} \quad \tau_G = \frac{\operatorname{cof}(L)}{\det(D)} = \frac{\operatorname{cof}(L)}{\prod_{v \in V} N_G^V(v)}. \tag{1}$$

This is true for any balanced digraph.

When $G = G_x$, the map $\alpha\colon \mathcal{E}_G \to \mathcal{S}_x$ is surjective and $N_G^E!$-to-one, where

$$N_G^E! := \prod_w N_G^E(w)! = \prod_w N_x^{(k)}(w)!.$$

Here the factor $N_G^E(w)!$ is the number of permutations of the edges corresponding to the word w. Thus,

$$|\mathcal{S}_x| = \frac{|\mathcal{E}_G|}{N_G^E!} = n\frac{|\tilde{\mathcal{E}}_G|}{N_G^E!} = n\tau_G\frac{N_G^V!}{N_G^E!} = n\tau_G\frac{\prod_{w \in \mathcal{A}^k} N_x^{(k)}(w)!}{\prod_{w \in \mathcal{A}^{k-1}} N_x^{(k-1)}(w)!},$$

which is the equivalent of Whittle's formula [14] for cyclic sequences.

A naive form of shuffling is to randomly pair incoming edges with outgoing edges at each vertex $v \in V$. This can be done in $N_G^V!$ different ways, of which only $\tau_G N_G^V!$ are Euler cycles; the others are unions of disjoint cycles. Removing τ_G from the formula thus represents a generalised form of shuffling. Whittle also commented on this in [14], saying that this form of shuffling would give a set of sequences rather than just one sequence.

4. Counting sets of words in shuffled sequences

In [2], the expected number of occurrences of a word $w \in \mathcal{A}^p$ in random shufflings of a sequence was analysed. This exploited the fact that the number of ways to replace an occurrence of w with $w_1 w_p$ in any sequence equals the number of occurrences of w in the sequence. Whittle's formula [14] was then used to calculate the number of shuffled sequences and the number of shuffled sequences with one w replaced. This approach was extended in [7] to find the variance. Theorem 1, below, may be seen as a direct extension of this.

The equivalent operation in the digraph representation is to remove a path λ representing the word w, i.e. remove the edges of λ, and insert a new edge $\bar{\lambda}$ from $\varepsilon^-(\lambda)$ to $\varepsilon^+(\lambda)$, representing the corresponding word $\alpha(\bar{\lambda}) = \alpha(\lambda)$ (see Figure 1). Euler cycles containing the path λ will then correspond to Euler cycles in the modified digraph. Counting Euler paths containing specific subpaths thus corresponds to counting shuffled sequences containing specific words.

Definition 7. For a word digraph G and a set of edge-disjoint paths $\Lambda = \{\lambda_1, \ldots, \lambda_p\}$, define the graph G_Λ by removing the paths of Λ from G and adding new edges $\bar{\lambda}_i$ with $\varepsilon(\bar{\lambda}_i) = \varepsilon(\lambda_i)$

and $\alpha(\bar{\lambda}_i) = \alpha(\lambda_i)$. The word count, $\alpha(\Lambda) = \sum_i \alpha(\lambda_i) \in \mathbb{N}_0^{\mathcal{A}^*}$, of Λ counts the words to which the paths λ_i correspond. For $\alpha(\Lambda) = W$, we will often write $G - W$ instead of G_Λ, e.g. in τ_{G-W} and N^E_{G-W}, since the G_Λ associated with such Λ are isomorphic.

Lemma 1. *For a digraph G and a set of edge-disjoint paths Λ in G, there is an injection $\tilde{\mathcal{E}}_{G_\Lambda} \to \tilde{\mathcal{E}}_G$ obtained by replacing occurrences of the edge $\bar{\lambda}$ of G_Λ, $\lambda \in \Lambda$, by the path λ in G. Moreover, this gives a canonical bijection such that $\tilde{\mathcal{E}}_{G_\Lambda} \cong \{\tilde{\gamma} \in \tilde{\mathcal{E}}_G \colon \Lambda \subset \tilde{\gamma}\}$, where $\Lambda \subset \tilde{\gamma}$ means that the paths of Λ are subpaths of $\tilde{\gamma}$.*

Thus, $|\tilde{\mathcal{E}}_{G_\Lambda}| = \tau_{G-W} N^V_{G-W}!$, with $W = \alpha(\Lambda)$, is the number of Euler paths of $\tilde{\mathcal{E}}_G$ that contain Λ as subpaths, and can be calculated using the BEST theorem (1).

Lemma 2. *Let $G = G_x$ be the word digraph representing the k-word composition $N_x^{(k)}$, for some $x \in \mathcal{S}$, let Λ be a set of edge-disjoint paths in G, and let $\alpha(\Lambda) = W = \sum_i r_i w_i$ be its word count, with r_i the number of paths in Λ corresponding to the word w_i. The edge and vertex counts of G_Λ are respectively*

$$N^E_{G-W} = N_x^{(k)} - N_W^{(k)} + W \quad \textit{and} \quad N^V_{G-W} = N_x^{(k-1)} - N_{\Delta W}^{(k-1)},$$

where $N_W^{(k)}(u) = \sum_{i=1}^q r_i N_{w_i}^{(k)}(u)$ counts words $u \in \mathcal{A}^k$ as subwords in W, the term W ensures that $N^E_{G-W}(w_i) = r_i$ for the new edges, and $N_{\Delta W}^{(k-1)}(v) = \sum_{i=1}^q r_i N_{\Delta w_i}^{(k-1)}(v)$ counts words $v \in \mathcal{A}^{k-1}$ as internal subwords of W, with $\Delta w' = w'_2 \cdots w'_{p-1}$ for any word $w' = w'_1 \cdots w'_p \in \mathcal{A}^p$.

Calculating the factorial moments of the distribution requires counting sets of words in the sequence, i.e. sets of subpaths of the Euler cycle that correspond to a specific set of words. Counting words, pairs of words, or sets of words simply amounts to summing over all corresponding choices of Λ.

Definition 8. For a digraph G, let $\mathcal{P}_G$ denote the set of paths in G, and let

$$\mathcal{P}_G^* := \{\Lambda \subset \mathcal{P}_G \colon \Lambda \text{ contains edge-disjoint paths}\}.$$

If γ is a path in G, let $\mathcal{P}_\gamma$ and $\mathcal{P}_\gamma^*$ denote the corresponding restrictions to subpaths of γ. The restrictions of $\mathcal{P}_G^*$ and $\mathcal{P}_\gamma^*$ to sets of type $W = \sum_{i=1}^q r_i w_i \in \mathbb{N}_0^{\mathcal{A}^*}$ are respectively denoted by

$$\mathcal{P}_G^W := \{\Lambda \in \mathcal{P}_G^* \colon \alpha(\Lambda) = W\} \quad \text{and} \quad \mathcal{P}_\gamma^W := \{\Lambda \in \mathcal{P}_\gamma^* \colon \alpha(\Lambda) = W\},$$

and the *edge-disjoint word set count* is written $N_\gamma(W) := |\mathcal{P}_\gamma^W|$.

If $G = G_x$ represents the k-word count of a cyclic sequence x and $\gamma \in \mathcal{E}_G$ is an Euler sequence, then $y = \alpha(x)$ is a shuffling of x. If $W = \sum_i r_i w_i$ then $N_\gamma(W)$ counts the number of ways of picking a set of r_i occurrences of w_i in y, for all i, such that no two words overlap by k or more letters, i.e. such that they are *k-disjoint*. Denoting this *k-disjoint word set count* in y by $N_y(W)$, the above means that $N_y(W) = N_\gamma(W)$.

Combining the above results gives the following theorem.

Theorem 1. *The expected number of k-disjoint word sets of type $W = \sum_{i=1}^q r_i w_i$, where $w_i \neq w_j$ for $i \neq j$, in random shufflings X of a sequence $x \in \mathcal{S}$ is*

$$\mathrm{E}[N_X(W)] = \frac{\tau_{G-W} N_G^E!/N_{G-W}^E!}{\tau_G N_G^V!/N_{G-W}^V!} = \frac{\tau_{G-W} (N_x^{(k)})^{\langle N_W^{(k)} \rangle}}{\tau_G \prod_{i=1}^q r_i! (N_x^{(k-1)})^{\langle N_{\Delta W}^{(k-1)} \rangle}},$$

https://doi.org/10.1239/aap/1143936143 Published online by Cambridge University Press

where the falling factorial $s^{\langle r\rangle} = s!/(s-r)!$ *is applied elementwise, i.e.*

$$(N_x^{(k-1)})^{\langle N_{\Delta W}^{(k-1)}\rangle} = \prod_{v\in\mathcal{A}^{k-1}} (N_x^{(k-1)}(v))^{\langle N_{\Delta W}^{(k-1)}(v)\rangle}.$$

The term $\prod_i r_i! = W!$ *comes from* $N_{G-W}^E! = (N_x^{(k)} - N_W^{(k)})!\,W!$, *and corresponds to the new edges.*

Proof. Since $N_y(W) = N_\gamma(W)$ for $y = \alpha(\gamma)$ and $\mathcal{E}_G \to \mathcal{S}_x$ is $N_G^E!$-to-one, we have

$$\mathrm{E}[N_X(W)] = \frac{1}{|\mathcal{E}_G|}\sum_{\gamma\in\mathcal{E}_G} N_\gamma(W) = \frac{|\mathcal{P}_G^W||\mathcal{E}_{G-W}|}{|\mathcal{E}_G|} = |\mathcal{P}_G^W|\frac{\tau_{G-W}}{\tau_G}\frac{N_{G-W}^V!}{N_G^V!},$$

as

$$\sum_{\gamma\in\mathcal{E}_G} N_\gamma(W) = |\{(\gamma,\Lambda),\ \gamma\in\mathcal{E}_G,\ \Lambda\in\mathcal{P}_\gamma^W\}| = \sum_{\Lambda\in\mathcal{P}_G^W} |\mathcal{E}_{G_\Lambda}|.$$

Picking edge-disjoint paths $\lambda_{i,j}$, $i = 1,\dots,q$, $j = 1,\dots,r_i$, with $\alpha(\lambda_{i,j}) = w_i$ is equivalent to picking an ordered list of $N_W(u)$ edges corresponding to u, for all $u \in \mathcal{A}^k$. This can be done in $(N_G^E)^{\langle N_W^{(k)}\rangle}$ ways: $(N_G^E(u))^{\langle N_W^{(k)}(u)\rangle}$ ways for each u. As the $\lambda_{i,j}$ for each i may be ordered in $r_i!$ different ways and $\mathcal{P}_G^W$ is unordered, we have

$$|\mathcal{P}_G^W| = (N_G^E)^{\langle N_W^{(k)}\rangle}\Big/\prod_i r_i!.$$

5. Distribution of $N_X(w)$

Theorem 1 counts only sets of k-disjoint words. When counting copies of a word w in a sequence, this becomes a problem if the word can form clumps of overlapping occurrences. The effect of clumping in Markov chains was dealt with in [7] and, more extensively, in [10]. Our main attention here is restricted to words that do not clump; we present additional results on clumped words in Appendix A.

Definition 9. Let $w \in \mathcal{A}^p$, $p > k$, where $k-1$ is the order of the shuffling. We say that w is *clumpable* (or *k-clumpable*) if $w_{[1,l]} = w_{[p+1-l,p]}$ for some $l \geq k$; otherwise, it is *nonclumpable*. The *order* of w in x (or relative to $N_x^{(k)}$) is the largest integer R such that $RN_w^{(k)} \leq N_x^{(k)}$.

Given a word w and a composition $N_x^{(k)}$, by piecing together k-words (or edges of the corresponding word digraph) to form longer words without exceeding the available number, $N_x^{(k)}$, we can form at most R simultaneous copies of w, where R is the order of w. However, if $\tau_{G-Rw} = 0$ then there is no way of making a shuffled sequence (an Euler cycle) containing all of these, in which case the largest possible number of occurrences of w in a shuffled sequence is $R-1$.

The keys to describing the distribution for a word w are the generating functions

$$\check{G}(u) = \sum_{r=0}^{\infty}\mathrm{E}[N_X(rw)]u^r \quad\text{and}\quad G(u) = \sum_{r=0}^{\infty}\mathrm{E}\left[\binom{N_X(w)}{r}\right]u^r,$$

https://doi.org/10.1239/aap/1143936143 Published online by Cambridge University Press

which respectively count the sets of r k-disjoint copies of w and the sets of r different, but possibly overlapping, copies of w, and the probability generating function

$$P(t) = \mathrm{E}[t^{N_X(w)}] = \sum_{m=0}^{\infty} \mathrm{P}[N_X(w) = m]t^m = G(t-1).$$

If w is nonclumpable and of order R, these are all polynomials of degree R (or $R-1$ if $\tau_{G-Rw} = 0$), and $G(u) = \check{G}(u)$. They may be expressed in terms of generalised hypergeometric functions (GHFs).

Definition 10. For $p, q \in \mathbb{N}_0$, $a = (a_1, \dots, a_p) \in \mathbb{R}^p$, $A \in \mathbb{R}^q$, $\Delta a \in \mathbb{N}_0^p$, $\Delta A \in \mathbb{N}_0^q$, and a function $h\colon \mathbb{N}_0 \to \mathbb{R}$, define

$$F_h^{(R)}\left(\begin{matrix}\Delta a \times (-a) \\ \Delta A \times (-A)\end{matrix} \,\middle|\, (-1)^{|\Delta a|+|\Delta A|}u\right) := \sum_{r=0}^{R} h(r) \frac{\prod_{i=1}^{p} a_i{}^{\langle r\Delta a_i\rangle}}{\prod_{i=1}^{q} A_i{}^{\langle r\Delta A_i\rangle}} \frac{u^r}{r!}, \tag{2}$$

where

$$|\Delta a| + |\Delta A| = \sum_{i=1}^{p} \Delta a_i + \sum_{i=1}^{q} \Delta A_i$$

(i.e. modulus has a different meaning for vectors than it does for sets). We will generally have $|\Delta a| = |\Delta A| + 1$, in which case the sign in front of u will always be a minus. If no h is specified then implicitly $h = 1$ (thus, $F^{(R)} \equiv F_1^{(R)}$), and if no R is specified then implicitly $R = \infty$. The notation has been chosen so as to make this a generalisation of the *generalised hypergeometric function*

$$F\left(\begin{matrix}a \\ A\end{matrix} \,\middle|\, u\right) = F\left(\begin{matrix}a_1, \dots, a_p \\ A_1, \dots, A_q\end{matrix} \,\middle|\, u\right) = \sum_{r=0}^{\infty} \frac{\prod_{i=1}^{p} a_i{}^{\rangle r\langle}}{\prod_{i=1}^{p} A_i{}^{\rangle r\langle}} \frac{u^r}{r!},$$

where $s^{\rangle r\langle} = (s+r-1)!/(s-1)!$ is the *rising factorial* and $(-s)^{\rangle r\langle} = (-1)^r s^{\langle r\rangle}$.

We use the notation $\Delta a \times a$ merely to represent the list of pairs $(\Delta a_i, a_i)$ of parameters, and will write the parameter pair $1 \times r$ simply as r. Thus, in $\Delta a \times (-a)$, r the negative parameters, $-a_i$, correspond to falling factorial terms and the positive parameter, r, corresponds to a rising factorial term. In the classical generalised hypergeometric function, all the Δa_i equal 1.

If w is a word of order R and $f(r) = \tau_{G-rw}/\tau_G$, Theorem 1 implies that

$$\check{G}(u) = \sum_{r=0}^{R} \mathrm{E}[N_X(rw)] = F_f^{(R)}\left(\begin{matrix}N_w^{(k)} \times (-N_x^{(k)}) \\ N_w^{(k-1)} \times (-N_x^{(k-1)})\end{matrix} \,\middle|\, -u\right).$$

Since GHFs are well studied and implemented in various computer programs, e.g. MAPLE®, and MATHEMATICA®, we may express the modified GHFs defined in (2) in terms of ordinary GHFs.

The terms of the form $\Delta a \times (-a)$ may be rewritten in more common form, using $x^{\langle rm\rangle} = m^{rm}(x/m)^{\langle r\rangle} \cdots ((x-m+1)/m)^{\langle r\rangle}$, as follows:

$$F_h^{(R)}\left(\begin{matrix}\Delta a \times (-a) \\ \Delta A \times (-A)\end{matrix} \,\middle|\, u\right) = F_h^{(R)}\left(\begin{matrix}\dots, -\dfrac{a_i}{\Delta a_i}, \dots, -\dfrac{a_i - \Delta a_i + 1}{\Delta a_i}, \dots \\ \dots, -\dfrac{a_i}{\Delta a_i}, \dots, -\dfrac{a_i - \Delta a_i + 1}{\Delta a_i}, \dots\end{matrix} \,\middle|\, \Delta a^{\Delta a} u\right). \tag{3}$$

In addition, we may express $F_h^{(R)}$ in terms of $F^{(R)}$.

https://doi.org/10.1239/aap/1143936143 Published online by Cambridge University Press

Lemma 3. *With definitions as in Definition 10, by using the difference operator* $\Delta h(r) = h(r+1) - h(r)$, *such that* $(-\Delta)^k h(r) = \sum_{i=0}^{k} (-1)^i \binom{k}{i} h(r+i)$, *and letting* $\pm u$ *denote* $(-1)^{|\Delta a|+|\Delta A|}u$, *we have*

$$F_h^{(R)}\left(\begin{matrix}\Delta a \times (-a)\\ \Delta A \times (-A)\end{matrix}\,\middle|\, \pm u\right) = \sum_{j=0}^{R} \frac{(-\Delta)^j h(0) u^j}{j!} \frac{a^{\langle j\Delta a\rangle}}{A^{\langle j\Delta A\rangle}} \\ \times F^{(R-j)}\left(\begin{matrix}\Delta a \times (j\Delta a - a)\\ \Delta A \times (j\Delta A - A)\end{matrix}\,\middle|\, \pm u\right).$$

Proof. Using $h(r) = \sum_{j=0}^{\infty} (-\Delta)^j h(0) \binom{r}{j}$, the sum $\sum_r h(r) c_r u^r / r!$ can be rewritten

$$\sum_j (-\Delta)^j h(0) \left(\frac{u^j}{j!}\right) \frac{\mathrm{d}^j}{\mathrm{d}u^j} C(u),$$

where $C(u) = \sum_{r=0}^{\infty} c_r u^r / r!$. Setting $c_j = a^{\langle j\Delta a\rangle} / b^{\langle j\Delta b\rangle}$, $j \le R$, gives the desired inequality.

In the common definition of GHFs, parameters $a_i = A_j$ (for any i and j) are eliminated even if this produces '0/0' terms in the coefficients. This is the reason why the order R must be used as upper limit in the sum. However, in most cases this need not be done explicitly.

Lemma 4. *For a word* w *of order* R *in* $x \in \mathcal{S}$, *let*

$$U^- = \{u \in \mathcal{A}^k : (R+1)N_w(u) > N_x(u)\},$$
$$V^- = \{v \in \mathcal{A}^{k-1} : (R+1)N_{\Delta w}(v) > N_x(v)\}.$$

Then $|U^-| > 0$ *and* $|U^-| \ge |V^-|$. *If* $|U^-| > |V^-|$, *we have*

$$F_h^{(R)}\left(\begin{matrix}N_w^{(k)} \times (-N_x^{(k)})\\ N_{\Delta w}^{(k-1)} \times (-N_x^{(k-1)})\end{matrix}\,\middle|\, -u\right) = F_h\left(\begin{matrix}N_w^{(k)} \times (-N_x^{(k)})\\ N_{\Delta w}^{(k-1)} \times (-N_x^{(k-1)})\end{matrix}\,\middle|\, -u\right). \tag{4}$$

Proof. As w is of order R, there is no integer parameter greater than $-R$ in the expansion of (4) defined in (3): there is one $-R$ for each word in U^- and V^-. If $|U^-| > |V^-|$ then removing $|V^-|$ of the $-R$ parameters from both the upper and lower rows leaves at least one remaining in the upper row, causing the series to terminate with degree R.

By the definition of the order R, U^- is nonempty. For each $v \in V^-$, there must be at least one word $u \in U^-$ with $v = u_{[1,k-1]}$ and one with $v = u_{[2,k]}$. When the degree of the $(k-1)$-word v becomes negative after subtracting $R+1$ copies of w, the in-degree to v for at least one k-word must be negative and the out-degree from v for some k-word must be negative. Hence, $|U^-| \ge |V^-|$.

Having $|U^-| = |V^-|$ is unlikely unless the numbers of $N_x^{(k)}$ are very small or w is very long and repetitive; more often, V^- will be empty. Hence, in most cases, the truncated series $F_f^{(R)}$ can be replaced by $F_f \equiv F_f^{(\infty)}$.

The factor $f(r) = \tau_{G-rw}/\tau_G$ for a word w of order R is generally expressible as a rational function except in a special case in which τ_{G-Rw} becomes 0/0 and the limit does not give the appropriate value.

https://doi.org/10.1239/aap/1143936143 Published online by Cambridge University Press

Lemma 5. *For a word w of order R in $x \in \mathcal{S}$, with $f(r) = \tau_{G-rw}/\tau_G$ for integers $r \leq R$, we have*

$$f(r) = \frac{g(r)}{\prod_{w \in \mathcal{A}^{k-1}} (1 - rN_{\Delta w}^{(k-1)}/N_w^{(k-1)})} \quad \textit{for } r < R, \textit{ with} \quad g(r) = \frac{\operatorname{cof}(L - r\Delta L)}{\operatorname{cof}(L)},$$

where $L - r\Delta L$ is the Laplacian matrix of $G - rw$. The function $g(r)$ is a polynomial of degree at most $|w| - k$.

Using this expression for $f(r)$ for real $r < R$, let $\Delta f_R = f(R) - \lim_{r \to R} f(r)$ be the jump at $r = R$. If $\Delta f_R \neq 0$, there must be a cycle $u_0, \dots, u_p = u_0 \in \mathcal{A}^k$, with $v_i = (u_i)_{[1,k-1]} = (u_{i+1})_{[2,k]}$, such that $RN_w(u_i) = N_x(u_i) > 0$ and $RN_{\Delta w}(v_i) = N_x(v_i) > 0$ for all i.

Proof. This result follows from (1) and results in Appendix B, in which an expression for Δf_R is given.

As with Lemma 4, the criterion for $\Delta f_R \neq 0$ is very strong and is rarely satisfied: more often, there will be no $v \in \mathcal{A}^{k-1}$ with $RN_{\Delta w}(v) = N_x(v) > 0$ unless k-word counts are low or w is long and repetitive.

The modifying function $f(r)$ will often be close to 1. However, $g(r)$ will be a polynomial of degree at most $p - k$ for $w \in \mathcal{A}^p$; thus, $\Delta^j g(r) = 0$ for $j > p - k$. In Lemma 3, using $f(r) \approx 1$ ensures that the higher-order differences have little effect. However, if $f(r) = g(r)/\prod(1 - r\Delta A_i/A_i)$ for $r \leq R$ and $\Delta f_R = 0$, then

$$F_f^{(R)}\left(\begin{matrix} \Delta a \times (-a) \\ \Delta A \times (-A) \end{matrix} \,\middle|\, \pm u\right) = F_g^{(R)}\left(\begin{matrix} \Delta a \times (-a) \\ \Delta A \times (1 - A) \end{matrix} \,\middle|\, \pm u\right),$$

where the latter gives a sum over $j = 0, \dots, p - k$ in Lemma 3; if $\Delta f_R \neq 0$, then the corresponding term must be added to the right-hand side.

Corollary 1. *If w is a word whose $(k - 1)$-words are all distinct, then Lemmas 4 and 5 both apply; i.e. $F_h^{(R)} = F_h$ for any h, as in (4), and $\Delta f_R = 0$.*

In the nonclumpable case, $G(u) = \check{G}(u)$ and the probability generating function is $P(t) = G(t - 1)$. We may use this to obtain the probabilities and tail probabilities.

Theorem 2. *Let $w \in \mathcal{A}^*$ be a nonclumpable word of order R (with respect to order-$(k - 1)$ shufflings of x). With $f(r) = \tau_G/\tau_{G-rw}$, $g(r)$ and Δf_R as in Lemma 5, and $f(r + \cdot)$ denoting the function $s \mapsto f(r + s)$, we have*

$$\begin{aligned} \mathrm{P}[N_X(w) = r] &= \frac{(N_x^{(k)})^{\langle rN_w^{(k)} \rangle}}{r!\,(N_x^{(k-1)})^{\langle rN_{\Delta w}^{(k)} \rangle}} F_{f(r+\cdot)}^{(R-r)}\left(\begin{matrix} N_w^{(k)} \times (rN_w^{(k)} - N_x^{(k)}) \\ N_{\Delta w}^{(k-1)} \times (rN_{\Delta w}^{(k-1)} - N_x^{(k-1)}) \end{matrix} \,\middle|\, 1\right) \\ &= \frac{(N_x^{(k)})^{\langle rN_w^{(k)} \rangle}}{r!\,(N_x^{(k-1)} - 1)^{\langle rN_{\Delta w}^{(k)} \rangle}} F_{g(r+\cdot)}^{(R-r)}\left(\begin{matrix} N_w^{(k)} \times (rN_w^{(k)} - N_x^{(k)}) \\ N_{\Delta w}^{(k-1)} \times (rN_{\Delta w}^{(k-1)} - N_x^{(k-1)} + 1) \end{matrix} \,\middle|\, 1\right) \\ &\quad + \Delta f_R \frac{(-1)^{R-r}}{r!\,(R-r)!} \frac{(N_x^{(k)})^{\langle RN_w^{(k)} \rangle}}{(N_x^{(k-1)})^{\langle RN_{\Delta w}^{(k-1)} \rangle}} \end{aligned} \tag{5}$$

and

$$
\begin{aligned}
&\mathrm{P}[N_X(w) \geq r] \\
&\quad = \frac{(N_x^{(k)})^{\langle r N_w^{(k)} \rangle}}{r!\,(N_x^{(k-1)})^{\langle r N_{\Delta w}^{(k)} \rangle}} F_{f(r+\cdot)}^{(R-r)} \left(\begin{matrix} N_w^{(k)} \times (r N_w^{(k)} - N_x^{(k)}), r \\ N_{\Delta w}^{(k-1)} \times (r N_{\Delta w}^{(k-1)} - N_x^{(k-1)}), r+1 \end{matrix} \,\middle|\, 1 \right) \\
&\quad = \frac{(N_x^{(k)})^{\langle r N_w^{(k)} \rangle}}{r!\,(N_x^{(k-1)} - 1)^{\langle r N_{\Delta w}^{(k)} \rangle}} F_{g(r+\cdot)}^{(R-r)} \left(\begin{matrix} N_w^{(k)} \times (r N_w^{(k)} - N_x^{(k)}), r \\ N_{\Delta w}^{(k-1)} \times (r N_{\Delta w}^{(k-1)} - N_x^{(k-1)} + 1), r+1 \end{matrix} \,\middle|\, 1 \right) \\
&\qquad + \Delta f_R \frac{(-1)^{R-r}}{(r-1)!\,(R-r)!\,R} \frac{(N_x^{(k)})^{\langle R N_w^{(k)} \rangle}}{(N_x^{(k-1)})^{\langle R N_{\Delta w}^{(k-1)} \rangle}} \qquad (6)
\end{aligned}
$$

If Lemma 4 applies then $F_{f(r+\cdot)}^{(R-r)}$ and $F_{g(r+\cdot)}^{(R-r)}$ may be replaced by $F_{f(r+\cdot)}$ and $F_{g(r+\cdot)}$, respectively, and if Lemma 5 applies then $\Delta f_R = 0$, causing that term to disappear.

Proof. The probability that $N_X(w) = r$ is the rth term in the Taylor expansion of $P(t) = G(t-1)$, given by

$$
\begin{aligned}
P(t) &= F_f^{(R)} \left(\begin{matrix} N_w^{(k)} \times (-N_x^{(k)}) \\ N_{\Delta x}^{(k-1)} \times (-N_x^{(k-1)}) \end{matrix} \,\middle|\, 1-t \right) \\
&= F_g^{(R)} \left(\begin{matrix} N_w^{(k)} \times (-N_x^{(k)}) \\ N_{\Delta x}^{(k-1)} \times (1-N_x^{(k-1)}) \end{matrix} \,\middle|\, 1-t \right) + \Delta f_R \frac{(N_x^{(k)})^{\langle R N_w^{(k)} \rangle}}{(N_x^{(k-1)})^{\langle R N_{\Delta x}^{(k-1)} \rangle}} \frac{(t-1)^R}{R!}.
\end{aligned}
$$

The upper tail probabilities $\mathrm{P}[N_w(w) > r]$ are the terms of the series $(P(t)-1)/(t-1)$. In terms of $u = t-1$, we have

$$
\frac{G(u)-1}{u} = \frac{(N_x^{(k)})^{\langle N_w^{(k)} \rangle}}{(N_x^{(k-1)})^{\langle N_{\Delta w}^{(k-1)} \rangle}} F_{f(1+\cdot)}^{(R-1)} \left(\begin{matrix} N_w^{(k)} \times (N_w^{(k)} - N_x^{(k)}), 1 \\ N_{\Delta w}^{(k-1)} \times (N_{\Delta w}^{(k-1)} - N_x^{(k-1)}), 2 \end{matrix} \,\middle|\, -u \right),
$$

where the effect of the added parameters 1 and 2 is to replace $u^r/r!$ in the series by $u^r/(r+1)!$. The coefficients may be read off from the Taylor expansion in t.

Setting $f(r) = 1$ corresponds to the generalised shuffling described at the end of Section 3. For a word $w \in \mathcal{A}^p$ with all $(k-1)$-words distinct, $N_X(w)$ may then be expressed as $N_X(w) = U_{p-k}$, where

$$
U_0 = a_0 \qquad \text{and} \qquad U_j \sim \mathrm{Hyp}[U_{j-1}, a_j; A_j], \quad j > 0,
$$

with $a_j = N_x(w_{[j+1,j+k]})$, $A_j = N_x(w_{[j+1,j+k-1]})$, and $\mathrm{Hyp}[u, v; n]$ denoting the hypergeometric distribution with parameters u, v, and n.

6. Computational examples

Computations of upper tail probabilities rely on Theorem 2 combined with Lemmas 3 and 4. As examples, we use the circular sequence of length 100 in Figure 1, and the 1461 intergenic regions of *Haemophilus influenzae* (GenBank® sequence NC000907) with a total length of 221 046, both using $k = 4$. The computations were performed on a laptop with a 1.4 GHz Pentium® M processor and 768 MB of RAM. The SAS® program was used for sequence parsing,

https://doi.org/10.1239/aap/1143936143 Published online by Cambridge University Press

word counting, and large matrix calculations, and the hypergeom() function of MAPLE 9 was used to calculate GHFs.

The approach to analysing a sequence is as follows.

(i) Count the k-words of the sequence.

(ii) Construct the Laplacian matrix.

(iii) Calculate the generalised inverse $L^{\dagger}$ (or $\check{L} = (L + kE)^{-1}$, as in Appendix B).

(iv) For each word w to be analysed, construct the matrix ΔL based on the k-word composition of w, and calculate $g(s) = \det(I - sL^{\dagger}\Delta L)$.

(v) Taking r to be the number of occurrences of w in the sequence, expand either $F^{(R-r)}_{f(r+\cdot)}$ or $F^{(R-r)}_{g(r+\cdot)}$ (of (5) or (6)) using Lemma 3, and calculate the values of the GHFs. Use the latter for exact calculations, as the expansion gives a finite sum, and the former for approximations, as f tends to be almost constant. If Lemma 4 applies then the bounds $R - r$ can be ignored.

Example 1. For the cyclic sequence of Figure 1, with the 3-words ordered alphabetically, the Laplacian matrix and its generalised inverse are

$$L = \begin{bmatrix} 5 & -5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 11 & -6 & -5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 11 & 0 & -3 & -8 & 0 & 0 \\ 0 & 0 & 0 & 12 & 0 & 0 & -4 & -8 \\ -5 & -6 & 0 & 0 & 11 & 0 & 0 & 0 \\ 0 & 0 & -5 & -7 & 0 & 12 & 0 & 0 \\ 0 & 0 & 0 & 0 & -8 & -4 & 12 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -8 & 8 \end{bmatrix},$$

$$L^{\dagger} = 10^{-3} \begin{bmatrix} 135 & 28 & -9 & -16 & -40 & -29 & -36 & -31 \\ -40 & 53 & 16 & 9 & -16 & -4 & -12 & -6 \\ -32 & -30 & 79 & -3 & -7 & 34 & -24 & -18 \\ -26 & -23 & -35 & 49 & -1 & -25 & 28 & 33 \\ 28 & 30 & -7 & -13 & 53 & -27 & -34 & -29 \\ -39 & -36 & 2 & 17 & -14 & 73 & -4 & 1 \\ -5 & -3 & -14 & -14 & 20 & -4 & 49 & -29 \\ -20 & -18 & -30 & -29 & 5 & -20 & 33 & 80 \end{bmatrix}.$$

The word $w = bbabaaa$ occurs twice in the sequence. In terms of the $(bba, bab, aba, baa, aaa)$-submatrices of $L^{\dagger}$ and ΔL (i.e. the matrices of components whose indices are restricted to the given list),

$$\begin{aligned} g(r) &= \det\left(I - \frac{r}{1000} \begin{bmatrix} 49 & -4 & -14 & 20 & -5 \\ -4 & 73 & 2 & -14 & -39 \\ -24 & 34 & 79 & -7 & -32 \\ -34 & -27 & -7 & 53 & 28 \\ -36 & -29 & -9 & -40 & 135 \end{bmatrix} \begin{bmatrix} 0 & -1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right) \\ &= 1 - 0.185r + 0.0148r^2 - 0.000\,571r^3. \end{aligned}$$

https://doi.org/10.1239/aap/1143936143 Published online by Cambridge University Press

The corresponding 4- and 3-word counts are $N_w^{(4)} \times N_x^{(4)} \sim a = (4, 5, 3, 5)$ and $N_{\Delta w}^{(3)} \times N_x^{(3)} \sim A = (12, 11, 11)$, respectively. Theorem 2 with simplifications due to Corollary 1 then gives

$$\mathrm{P}[N_X(w) = r] = \sum_{k=0}^{R-r} \frac{\Delta^k g(r) a^{\langle r+k \rangle}}{r!\, k!\, (A-1)^{\langle r+k \rangle}} F\left(\begin{matrix} r+k-a \\ r+k+1-A \end{matrix} \,\middle|\, 1 \right),$$

which, with $r = 2$ and $R = 3$, equals

$$\frac{0.685 \times 28\,800}{2 \times 1 \times 891\,000} F\left(\begin{matrix} -2, -3, -1, -3 \\ -9, -8, -8 \end{matrix} \,\middle|\, 1 \right) + \frac{0.1217 \times 518\,400}{2 \times 1 \times 513\,216\,000} F\left(\begin{matrix} -1, -2, -0, -2 \\ -8, -7, -7 \end{matrix} \,\middle|\, 1 \right) = 0.010\,78.$$

Using MAPLE, e.g. the function hypergeom([−2, −3, −1, −3], [−9, −8, −8], 1.), gives the values 0.969 and 1 for the two GHFs above. The corresponding upper tail probability is

$$\mathrm{P}[N_X(w) \geq 2] = \frac{0.685 \times 28\,800}{2 \times 1 \times 891\,000} F\left(\begin{matrix} -2, -3, -1, -3, 2 \\ -9, -8, -8, 3 \end{matrix} \,\middle|\, 1 \right) + \frac{0.1217 \times 518\,400}{2 \times 1 \times 513\,216\,000} F\left(\begin{matrix} -1, -2, -0, -2, 3 \\ -8, -7, -7, 4 \end{matrix} \,\middle|\, 1 \right) = 0.010\,87.$$

In the above, Corollary 1 ensured that R and Δf_R could be ignored. However, even in the more general case, explicit testing is not needed. Whether R or Δf_R need to be taken into account will be apparent when writing down the expression. When expanding $F_f^{(R)}$, this happens if either $f(R) = 0/0$ or if the parameter $r - R$, which terminates the series, occurs as many times in the lower set of parameters as it does in the upper set. When expanding $F_g^{(R)}$ there will then be a parameter $1 + r - R$ in the lower set of parameters that makes the unterminated series invalid.

The GHFs used here are alternating series with terms that at first increase exponentially. Thus, direct evaluation of the sum is normally not feasible. However, the GHFs used to express the point probabilities $\mathrm{P}[N_X(w) = r]$ are polynomials $F(u)$ with roots $u \in [1, \infty)$: there is a root at 1 of order

$$\sum_{u \in \mathcal{A}^k} N_w^{(k)}(u) N_x^{(k)}(u) - \sum_{v \in \mathcal{A}^{k-1}} N_{\Delta w}^{(k-1)}(v) N_x^{(k-1)}(v)$$

(if this is positive) that corresponds to the minimal number of occurrences of w in any shuffling; the other roots are simple. When there is no root at 1, the Taylor series of $\log F(u)$ will converge for $u = 1$, often relatively fast. This is not generally true of the GHFs used for the upper tail probabilities; however, for values of r well above the expected number of occurrences of w, this approach still tends to work, as these GHFs may be seen as perturbations of those used in the corresponding point probabilities, thus having roots on or close to the line $(1, \infty)$ in the complex plane.

Example 2. The word $w =$ AAGTGCGGT is a strongly preserved part of the DNA uptake signalling sequence of *H. influenzae*. Counting w in intergenic regions reveals its occurrence 232 times on one of the DNA strands. The word counts corresponding to 4-words of w and 3-words of Δw are

$$a = (1333, 823, 650, 691, 608, 644) \quad \text{and} \quad A = (3115, 2209, 2464, 1705, 1448),$$

respectively. Using the alphabet {A, C, G, T, ?, –}, where '?' represents an unknown letter and '–' is used to replace each gene, the Laplacian is a 155×155 matrix. The generalised inverse $L^\dagger$ is computed and the submatrix corresponding to the 3-words of w used to compute

$$g(1000s) = 1 - 2.46s + 2.37s^2 - 1.106s^3 + 0.244s^4 - 0.0201s^5.$$

The point probability is $\mathrm{P}[N_X(w) = 232] = 1.922 \times 10^{-374}$ and the upper tail probability is $\mathrm{P}[N_X(w) \geq 232] = 1.929 \times 10^{-374}$.

Parsing the files of GenBank (the genetic sequence database of the National Institutes of Health) to count intergenic 4-words and generate the Laplacian matrix took approximately 2 seconds. Calculating the generalised inverse took approximately 0.1 seconds. Computing the point probability and upper tail probability, including the GHFs, took approximately 0.9 seconds.

In the above example, the GHFs take values close to 0.43, and a second-degree Taylor series approximation of $\log F(u)$ for each of them would give sufficient accuracy; the modifying function f takes the value $f(232) = 0.9987$, thus having only a very minor effect. Hence, a good approximation is

$$\mathrm{P}[N_X(w) \geq r] \approx \frac{a^{\langle r\rangle}}{r!\,A^{\langle r\rangle}} \exp(\nu_r + \nu_r\nu_{r+1} - \nu_r^2), \quad \text{where} \quad \nu_k = \frac{k\prod_i(a_i - k)}{(k+1)\prod_i(A_i - k)},$$

giving 1.927×10^{-374} for $r = 232$.

An alternative approximation relies on deviance residuals, which are normally used with exponential families and generalised linear models. The deviance residual R_r for $N_x(w) = r$ is calculated from the cumulant generating function $C(s) = \log \mathrm{E}[\exp(sN_X(w))] = \log G(\mathrm{e}^s - 1)$ as

$$R_r = \operatorname{sgn}(r - \mu)\sqrt{D_r} \qquad \text{where} \qquad D_r = 2[r\hat{\theta}_r - C(\hat{\theta}_r)], \quad \mu = C'(0), \tag{7}$$

and $\hat{\theta}_r$ is the unique solution to $C'(\hat{\theta}_r) = r$. The assumption is that R_r will be approximately normally distributed. The computational advantage of this approach is that $\hat{\theta}_r$ will usually be close to 0, and Taylor series approximations of both $\ln G(u)$ and $C(s)$ converge quickly for u and s close to 0. If we also assume that $f(r) \approx 1$, then $G(u)$ is a GHF. This approximation is shown in Figure 2, together with the exact upper tail probabilities.

DNA consists of two complementary strands. There are 222 occurrences of w on the other strand, each corresponding to the word $w' =$ ACCGCACTT on the first. Counting w on either strand is thus equivalent to counting w and w' on the first strand. A more general problem is that of counting several different words. This is a difficult problem, as it involves a multiple sum that is not easily evaluated. When the words have no common $(k-1)$-words, as is the case for w and w', this sum may be separated into GHFs for each word.

Example 3. Here the word counts are $b = [641, 627, 687, 685, 837, 1339]$ for the 4-words of w' and $B = [1450, 1714, 2441, 2292, 3085]$ for the 3-words of $\Delta w'$. The probability that $N_X(w) = r$ and $N_X(w') = s$ is

$$\sum_{k,l} \frac{\Delta_r^k \Delta_s^l g(r,s) a^{\langle r+k\rangle} b^{\langle s+l\rangle}}{k!\,l!\,r!\,s!\,A^{\langle r+k\rangle} B^{\langle s+l\rangle}} F\left(\begin{matrix} r+k-a \\ r+k+1-A \end{matrix}\,\middle|\, 1\right) F\left(\begin{matrix} s+l-b \\ s+l+1-B \end{matrix}\,\middle|\, 1\right)$$

with $g(r,s) = \det(I - L^\dagger(r\Delta L_w - s\Delta L_{w'}))$, where ΔL_w and $\Delta L_{w'}$ are matrices such that $L - r\Delta L_w - s\Delta L_{w'}$ is the Laplacian matrix of $G - rw - sw'$ (see Lemma 5).

https://doi.org/10.1239/aap/1143936143 Published online by Cambridge University Press

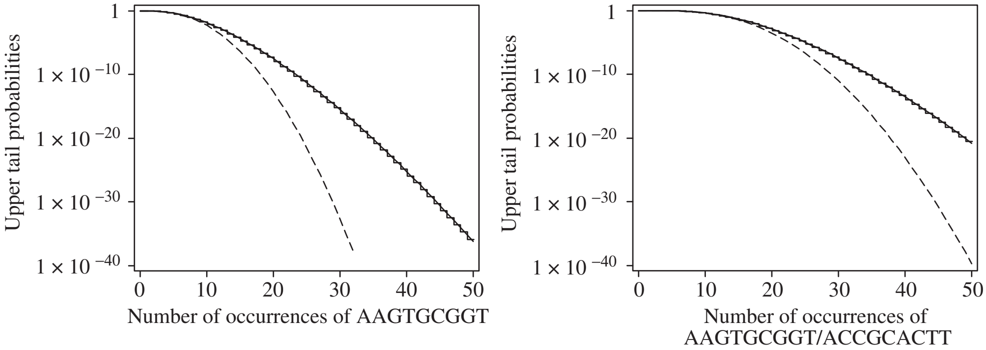

FIGURE 2: Upper tail probabilities for the DNA uptake signalling sequence AAGTGCGGT on one strand (*left*) and on both strands (*right*), the latter corresponding to counting both the uptake sequence and its reverse complement ACCGCACTT. Exact probabilities are shown as stepwise curves, approximations based on deviance residuals as smooth curves, and the normal distribution approximations as dashed curves. The effect of using the true modifying function f, rather than the approximation $f \approx 1$, is too small to see.

The above computation, though feasible in this and similar cases, is both impractical and hard to generalise. However, the approximation (7) is still computable with relative ease: setting $f(r, s) \approx 1$ gives the approximation

$$C_{N_X(w)+N_X(w')}(s) \approx \log F\left(\begin{matrix}-a\\-A\end{matrix}\,\middle|\, 1-\mathrm{e}^s\right) + \log F\left(\begin{matrix}-b\\-B\end{matrix}\,\middle|\, 1-\mathrm{e}^s\right)$$

in the above example, giving the approximate upper tail probabilities shown in Figure 2. If a Taylor series approximation of $C(s)$ is used, this relies only on $\mathrm{E}[\binom{N}{k}]$, for k up to the desired order; these expectations may be evaluated even in the presence of clumping or when counting several words, using Theorem 3, below.

Appendix A. Clumping

If all occurrences of a word $w \in \mathcal{A}^*$ in a sequence y are k-disjoint, then the number of ways of picking r of these is $N_y(rw) = \binom{N_y(w)}{r}$. If, on the other hand, w can overlap by k or more letters, $N_y(rw)$ will not count all sets of r occurrences of w, but only those in which the words do not overlap. This problem of clumping was dealt with in [7], in determining $\mathrm{var}[N_y(w)]$, and further in [10], with respect to clumping in Markov chains.

Definition 11. A *segment* of a sequence is specified by a pair (i, l), where i is the starting position and l is the length; thus, if x is a sequence, the segment (i, l) points to the word $x_i \cdots x_{i+l-1}$.

A *clump* is either a linear sequence $v \in \mathcal{A}^p$ or a cyclic sequence $v \in \mathcal{S}$ together with a set $\{(i_j, l_j)\}$ of segments of v which is such that every segment of length $k+1$ in v is contained in at least one of the segments (i_j, l_j), i.e. the segments cover all of v, overlapping by at least k letters. Let $\mathcal{C}$ denote the set of clumps in which two circular clumps are considered equivalent if one is a rotation of the other.

For a clump c defined as above, let $\alpha(c) = v$ denote the corresponding sequence and let $N_c = \sum_j v_{[i_j, i_j+l_j-1]} \in \mathbb{N}_0^{\mathcal{A}*}$, where the words $v_{[i_j, i_j+l_j-1]}$ represent unit basis vectors in $\mathbb{N}_0^{\mathcal{A}*}$, count the words forming the clump.

https://doi.org/10.1239/aap/1143936143 Published online by Cambridge University Press

The simplest clump is that corresponding to a single word: $c = (w, \{(1, p)\})$ for $w \in \mathcal{A}^p$. Thus, the set of words is naturally embedded in the set of clumps. If $w \in \mathcal{A}^p$ has $w_{[1,q]} = w_{[p-q+1,p]}$, it can form clumps c, of r copies, with $\alpha(c) = w_{[1,q]}^{\circ(r-1)} w$ (where '$\circ(r-1)$' denotes concatenation $r-1$ times), for which $N_c = rw$. Note that some words may overlap in several different ways.

When picking an arbitrary set of words in a sequence x, those overlapping by k or more letters may be combined into clumps. This will result in k-disjoint clumps. Theorem 1 may be used to count k-disjoint words and, hence, also k-disjoint clumps. Since different clumps may produce the same sequence, there is a difference between counting the number of ways of picking sets of clumps and the number of ways of picking sets of words. In fact,

$$N_x(\rho) = \frac{\alpha(\rho)!}{\rho!} N_x(\alpha(\rho)),$$

where $\rho = \sum s_i c_i \in \mathbb{N}_0^{\mathbb{C}}$ is the clump count and $\alpha(\rho) = \sum s_i \alpha(c_i) = \sum r_j w_j \in \mathbb{N}^{\mathcal{A}*}$ the corresponding word count, such that $\rho! = \prod s_i!$ and $\alpha(\rho)! = \prod r_j!$.

Theorem 3. *For any circular sequence $x \in \mathcal{S}$, distinct words $w_1, \dots, w_s$, and $W = \sum_i r_i w_i$ with $r_i \in \mathbb{N}_0$, we have*

$$\prod_i \binom{N_x(w_i)}{r_i} = \sum_{\substack{\rho \in \mathbb{N}_0^{\mathbb{C}} \\ N_\rho = W}} N_y(\rho) = \sum_{\substack{\rho \in \mathbb{N}_0^{\mathbb{C}} \\ N_\rho = W}} \frac{\alpha(\rho)!}{\rho!} N_x(\alpha(\rho)),$$

where $N_\rho = \sum_{c \in \mathbb{C}} \rho_c N_c$ counts the words that make up the clumps. The expectation for shuffled sequences $X \in \mathcal{S}_x$ is then given by

$$\mathrm{E}\Bigg[\prod_i \binom{N_X(w_i)}{r_i}\Bigg] = \sum_{\substack{\rho \in \mathbb{N}_0^{\mathbb{C}} \\ N_\rho = W}} \mathrm{E}[N_y(\rho)] = \sum_{\substack{\rho \in \mathbb{N}_0^{\mathbb{C}} \\ N_\rho = W}} \frac{\tau_{G-\alpha(\rho)}}{\tau_G} \frac{(N_x^{(k)})^{\langle N_{\alpha(\rho)}^{(k)} \rangle}}{\rho!\, (N_x^{(k-1)})^{\langle N_{\Delta\alpha(\rho)}^{(k-1)} \rangle}},$$

where $N_{\alpha(\rho)}^{(k)}$ counts the k-words in the words formed by the clumps and $N_{\Delta\alpha(\rho)}^{(k-1)}$ counts the number of $(k-1)$-words in their interiors.

Appendix B. Calculating τ

The expectations and probabilities found in Theorem 1 are primarily determined by the factorial and falling factorial terms, but with τ_{G-W}/τ_G as a correction factor encoding the effect of the global composition of the sequence. In many cases, this factor will be close to 1 and will hence have little effect.

Recall that τ_G is expressed in terms of a cofactor of the Laplacian matrix, given in (1). If $W = \sum r_i w_i$, where w_i are distinct words, and L is the Laplacian of $G \equiv G_x$, then the modified word digraph $G - W$ has Laplacian $L - \sum r_i \Delta L_i$ and diagonal matrix $D - \sum r_i \Delta D_i$ counting vertex degrees, i.e. ΔD_i has $N_{\Delta w_i}^{(k-1)}$ on the diagonal, counting the internal $(k-1)$-words of w_i, and $\Delta L_i = \Delta A_i - \Delta D_i$, where ΔA_i is the adjacency matrix, counting the edges of a path corresponding to the word w_i minus the new edge corresponding to w_i. From this we find that

$$\tau_{G-W} = \frac{\operatorname{cof}(L - \sum r_i \Delta L_i)}{\det(D - \sum r_i \Delta D_i)} = \frac{\operatorname{cof}(L) \det(I - \hat{L} \sum r_i \Delta L_i)}{\det(D) \prod_{v \in \mathcal{A}^{k-1}} (1 - \sum r_i N_{\Delta w_i}(v)/N_x(v))},$$

where either $\hat{L}$ is the Moore–Penrose generalised inverse $L^\dagger$ or $\hat{L} = (L + kE)^{-1} = L^\dagger + E/k$, with $E = 1^\top 1/m$, $m = |V|$ being the size of the matrices. This follows from Jacobi's rule, which gives $\mathrm{d}[\det(L + tE)]/\,\mathrm{d}t|_{t=0} = m \operatorname{cof}(L)$, and

$$\det(L - \Delta L + tE) = \det(L + tE) \det(I - L^\dagger \Delta L)$$

as $E\Delta L = 0$.

Computational advantages of this formula are that $\hat{L}$ may be computed once per sequence and then used in analysing different word frequencies, and that τ_{G-W} depends only on the submatrix of $\hat{L}$ corresponding to the $(k-1)$-subwords used in W. Thus, for $W = rw$,

$$f(r) = \frac{\tau_{G-W}}{\tau_G} = \frac{\det(I - r\hat{L}\Delta L)}{\det(I - rD^{-1}\Delta D)},$$

given that $D - r\Delta D$ has positive elements on the diagonal.

For $r = R$, where R is the order of w, the above expression for $f(r)$ may become $0/0$ in the case in which all edges to and from some vertices $V' \subset V$ have been used up, i.e. $RN_w^{(k-1)}(v) = N_x^{(k-1)}(v)$ for $v \in V'$. When this happens, τ_{G-Rw} is defined by removing from $L - R\Delta L$ and $D - R\Delta D$ the rows and columns corresponding to V' before calculating the cofactor and the determinant. If $V'' = V \setminus V'$ are the remaining vertices, then

$$\lim_{r \to R} \frac{\operatorname{cof}(L - r\Delta L)}{\det(D - r\Delta D)} = \frac{\det(\Delta L)_{V'}}{\det(\Delta D)_{V'}} \frac{\operatorname{cof}(L - R\Delta L)_{V''}}{\det(D - R\Delta D)_{V''}},$$

where the subscripts indicate the corresponding submatrices. This may be seen by expressing τ_{G-rw} in terms of cofactors of $Q_r = (D - r\Delta D)^{-1}(L - r\Delta L)$, for which the limit Q_R becomes well defined, with the components of the $V' \times V''$ and $V'' \times V'$ submatrices all vanishing. Thus,

$$f(R) = \frac{\det(\Delta D)_{V'}}{\det(\Delta L)_{V'}} \lim_{r \to R} f(r),$$

where the correction factor differs from 1 only if $(\Delta A)_{V'}$ contains a cycle, in the sense that there exist vertices $v_0, \dots, v_p = v_0 \in V'$ such that $\Delta A_{v_i v_{i+1}} \neq 0$.

Acknowledgement

This work was supported by the Research Council of Norway.

References

[1] Altschul, S. and Erickson, B. (1985). Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Molec. Biol. Evol.* **2,** 526–538.

[2] Cowan, R. (1991). Expected frequencies of DNA patterns using Whittle's formula. *J. Appl. Prob.* **28,** 886–892.

[3] Dawson, R. and Good, I. (1957). Exact Markov probabilities from oriented linear graphs. *Ann. Math. Statist.* **28,** 946–956.

[4] Fitch, W. (1983). Random sequences. *J. Molec. Biol.* **163,** 171–176.

[5] Goodman, L. (1958). Exact probabilities and asymptotic relationships for some statistics from m-th order Markov chains. *Ann. Math. Statist.* **29,** 476–490.

[6] Kandel, D., Matias, Y., Unger, R. and Winkler, P. (1996). Shuffling biological sequences. *Discrete Appl. Math.* **71,** 171–185.

[7] Prum, B., Rodolphe, F. and de Turckheim, E. (1995). Finding words with unexpected frequencies in deoxyribonucleic acid sequences. *J. R. Statist. Soc. B* **57,** 205–220.

[8] Robin, S. and Daudin, J. (1999). Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.* **36,** 179–193.

https://doi.org/10.1239/aap/1143936143 Published online by Cambridge University Press

[9] Robin, S. and Schbath, S. (2001). Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comput. Biol.* **8,** 349–359.

[10] Schbath, S. (1995). Compound Poisson approximation of word counts in DNA sequences. *ESAIM Prob. Statist.* **1,** 1–16.

[11] Schbath, S. (1995). Étude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN. Doctoral Thesis, Université René Descartes, Paris V.

[12] Tutte, W. (1948). The dissection of equilateral triangles into equilateral triangles. *Proc. Camb. Philos. Soc.* **44,** 463–482.

[13] Van Aardenne-Ehrenfest, T. and de Bruijn, N. (1951). Circuits and trees in oriented linear graphs. *Simon Stevin* **28,** 203–217.

[14] Whittle, P. (1955). Some distribution and moment formulae for the Markov chain. *J. R. Statist. Soc. B* **17,** 235–242.

[15] Zaman, A. (1984). Urn models for Markov exchangeability. *Ann. Prob.* **12,** 223–229.

https://doi.org/10.1239/aap/1143936143 Published online by Cambridge University Press