Published in final edited form as:

J Math Biol. 2014 August ; 69(2): 469–500. doi:10.1007/s00285-013-0705-3.

Robust k -mer frequency estimation using gapped k -mers

Mahmoud Ghandi,

Department of Biomedical Engineering and McKusick-Nathans Institute of Genetic Medicine,
Johns Hopkins University, Baltimore, MD 21205, USA

Morteza Mohammad-Noori, and

School of Mathematics, Statistics and Computer Science, University of Tehran. Iran

School of Computer Science, Institute for Research in Fundamental Sciences, P.O. Box:
19395-5746, Tehran, Iran

Michael A. Beer

Department of Biomedical Engineering and McKusick-Nathans Institute of Genetic Medicine,
Johns Hopkins University, Baltimore, MD 21205, USA

Morteza Mohammad-Noori: mnoori@khayam.ut.ac.ir; Michael A. Beer: mbeer@jhu.edu

Abstract

Oligomers of fixed length, k , commonly known as k -mers, are often used as fundamental elements in the description of DNA sequence features of diverse biological function, or as intermediate elements in the construction of more complex descriptors of sequence features such as position weight matrices. k -mers are very useful as general sequence features because they constitute a complete and unbiased feature set, and do not require parameterization based on incomplete knowledge of biological mechanisms. However, a fundamental limitation in the use of k -mers as sequence features is that as k is increased, larger spatial correlations in DNA sequence elements can be described, but the frequency of observing any specific k -mer becomes very small, and rapidly approaches a sparse matrix of binary counts. Thus any statistical learning approach using k -mers will be susceptible to noisy estimation of k -mer frequencies once k becomes large. Because all molecular DNA interactions have limited spatial extent, gapped k -mers often carry the relevant biological signal. Here we use gapped k -mer counts to more robustly estimate the ungapped k -mer frequencies, by deriving an equation for the minimum norm estimate of k -mer frequencies given an observed set of gapped k -mer frequencies. We demonstrate that this approach provides a more accurate estimate of the k -mer frequencies in real biological sequences using a sample of CTCF binding sites in the human genome.

Keywords

DNA sequence; Oligomer; k -mer; Frequency estimation; Statistical learning

© Springer-Verlag Berlin Heidelberg 2013

Correspondence to: Michael A. Beer, mbeer@jhu.edu.

Present Address: M. Ghandi Broad Institute, Cambridge, MA 02142, USA: ghandi@broadinstitute.org

Electronic supplementary material The online version of this article (doi:10.1007/s00285-013-0705-3) contains supplementary material, which is available to authorized users.

1 Introduction

Most analysis of polymeric biomolecules (e.g. protein, RNA, or DNA) at some point requires a model mapping polymer sequence features to functional molecular structures. In the context of DNA sequence properties, sequence similarity via sequence alignment, alignment to known repeat elements, and CpG islands are well known examples. In protein functional and structural studies, amino acid patterns are frequently mapped to structural or functional motifs (such as leucine zippers or phosphorylation sites). These descriptions have been fairly successful. However, when modeling DNA-protein interactions, a key molecular event in the process of transcriptional regulation, there is less consensus on what should be used as the best description of a DNA sequence (a binding site) bound by a protein (a transcription factor). Many models of this process use a position weight matrix (PWM) to describe the DNA binding site Stormo (2000); Beer and Tavazoie (2004). Other approaches use oligomers of fixed length, k , commonly known as k -mers, to describe the DNA binding sites. Using k -mers has the distinct advantage that they completely cover the discrete space of all possible DNA molecules of length k , while the space of all possible PWMs is continuous and large. We were initially motivated to generalize k -mer methods when our machine learning algorithms based on k -mer frequencies were found to be effective at predicting enhancers and modeling transcription factor binding sites Lee et al. (2011). Other examples of successful bioinformatic applications based on k -mers as sequence features include sequence homology Göke et al. (2012); Helden (2004), protein homology Leslie et al. (2004), and predictions of cis-regulatory modules Kantorovitz et al. (2009), transcription factor binding sites Xie et al. (2005); Elemento and Tavazoie (2005), transcription initiation sites Meinicke et al. (2004); Sonnenburg et al. (2006), and splice sites Ben-Hur et al. (2008); Sonnenburg et al. (2007). However, all of these k -mer based approaches suffer from the fundamental difficulty of estimating long k -mer frequencies in a finite set of biological samples. This paper presents a general solution to this problem, and our method can be applied to improve the statistical robustness of any of the aforementioned k -mer based approaches or others which use k -mer frequencies as direct features or as an intermediate step in the construction of more complex sequence descriptors.

When using k -mers, larger k 's will resolve larger binding sites and more accurately reflect biological function. For example, some transcription factors (such as ABF1 or CTCF) have relatively long binding sites that cannot be completely represented by short k -mers. So longer k -mers capture more relevant information; however, there is a limitation on the maximum length k which can be effectively used in statistical algorithms. Because longer k -mers are more sparsely populated in any finite training sequence set, there is a maximum length k for which the k -mer frequencies can be robustly estimated. Thus in practice, a k is chosen which is a tradeoff between resolving features and robust estimation of their frequencies. To overcome the finite training set size problem, one approach is to employ gapped k -mer frequencies. A gapped k -mer has a length ℓ , and a number of informative columns within that ℓ -mer, k , which reflects the base pairs which actually affect the strength of the TF-DNA binding interaction. We have found that using gapped k -mers can substantially improve the reliability of the ℓ -mer frequency estimation for a finite genomic training set, because while ℓ -mers become sparsely populated, gapped k -mers will still have

many instances in the training set, and thus their frequencies can be more reliably estimated. We use the observed gapped k -mer frequency distribution for all gapped k -mers to estimate the ungapped ℓ -mer frequencies, which are sparsely populated. Mathematically this turns out to be the minimum norm estimate for the ℓ -mer frequencies given the frequencies for all gapped k -mers. We derive the matrix, W , mapping between these two spaces. A closed form for this matrix is obtained by studying the combinatorial properties of the incidence matrix.

This paper is organized as follows: in Sect. 2 we formulate the ℓ -mer frequency estimation as a linear optimization problem and show that it can be solved by finding an eigendecomposition for AA^T , where A is the incidence matrix. The main result of the paper is an expression for the entries of the matrix W , presented in Theorem 2. We then give a brief outline of the proof of this expression. In Sect. 3, we provide the necessary preliminary information and some definitions needed for the eigendecomposition and the subsequent proof. In Sect. 3.4 we define the function v and prove some identities that enables us to find an eigendecomposition for matrix AA^T in Sect. 4. Then in Sect. 5 we show that matrix W is the Moore–Penrose pseudoinverse of matrix A and we prove the formula which gives the entries of W . The combinatorial arguments are continued in Sect. 6 to find a basis for the row space of matrix A that can further explain the rank and some other features of this matrix. In Sect. 7 we provide an explicit numerical example for A , the eigendecomposition of AA^T , and the derivation of W . It may aid the reader to have this specific example in mind while reading the proof. Then, in Sect. 8 we demonstrate the utility of our method on sequences bound by the transcription factor CTCF in the human genome. Finally, in Sect. 9, we give a summary of the work and discuss future directions.

2 Problem statement

In any given sequence sample, there are observed counts of gapped k -mers and ungapped ℓ -mers. Our fundamental assumption is that the counts of *gapped* k -mers in this sample are sufficient to define its biological function, and are also more robustly estimated from the sequence sample. Instead of using the *observed* counts of *ungapped* ℓ -mers, we want to find the most robust set of ungapped ℓ -mers that is also consistent with the gapped k -mer distribution. By *robust*, we mean that this estimate is resilient to small changes in the input or training set, even for large ℓ for which the actual ℓ -mer counts are very sparse. A classifier based on these more robustly estimated ℓ -mer counts is consequently also more robust in the sense that it is less sensitive to small changes in the input or training set, and is therefore a more stable predictor than a classifier based on exact counts.

Because the mapping from ungapped ℓ -mers to gapped k -mers is underdetermined, there are many sets of ℓ -mer counts that could have produced the observed set of gapped k -mer counts. We propose as the best set of ℓ -mer counts the minimum norm ℓ -mer distribution consistent with the gapped k -mer distribution. This is also the solution that minimizes the mean-squared error from a constant (flat) ℓ -mer distribution. The observed set of ungapped ℓ -mers is only one set of sequences which would have produced the given gapped k -mer counts. Our minimum norm distribution is in some sense the most likely of all sequences which could have produced the observed gapped k -mer counts.

For the case of DNA sequences, the alphabet is $\{A, C, G, T\}$, so the length of the alphabet is $b = 4$, but our solution for the optimal ℓ -mer count distribution presented below is valid for any b . We will consider ungapped ℓ -mers, and gapped k -mers of length ℓ with k ungapped (informative) positions.

Definition 1 The set of ungapped ℓ -mers is $U = \{u_j\}, 1 \leq j \leq N = b^\ell$, the set of all different sequences of length ℓ over the alphabet $\{0, 1, \dots, b-1\}$.

Definition 2 The set of gapped k -mers is $V = \{v_i\}, 1 \leq i \leq M = \binom{\ell}{k} b^k$, the set of all gapped k -mers of length ℓ with k known bits and $\ell - k$ gaps, where $k < \ell$.

Definition 3 The matrix which maps ungapped ℓ -mers to gapped k -mers is the matrix $A_{M \times N} = [a_{i,j}]$, a binary matrix defined as the following:

$$a_{i,j} = \begin{cases} 1 & \text{if } v_i \text{ matches } u_j \\ 0 & \text{otherwise} \end{cases}$$

Here “ v_i matches u_j ” means that all ungapped positions in the gapped k -mer v_i have the same letter of the alphabet as the corresponding position in the ungapped ℓ -mer u_j .

The ungapped count vector is defined as follows:

Definition 4 $\mathbf{x}$ is a vector of length N , where x_j is the count for u_j , and the gapped count vector is:

Definition 5 $\mathbf{y}$ is a vector of length M , where y_i is the count for v_i .

Given the above definitions, we can write the mapping from ℓ -mer counts to gapped k -mer counts as:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \quad (1)$$

As shown in Sect. 6 below, while it is usually but not always the case that $M < N$, since the rank of the matrix A , $\text{rank}(A) = \sum_{n=0}^k \binom{\ell}{n} (b-1)^n < N = b^\ell = \sum_{n=0}^\ell \binom{\ell}{n} (b-1)^n$ for $k < \ell$, this system is always underdetermined. Therefore, there are many possible ℓ -mer count vectors $\mathbf{x}$ that would produce the same gapped k -mer count vector $\mathbf{y}$. While the maximum entropy $\mathbf{x}$ is probably the most robust estimate to use, its solution is nonlinear and would likely require prohibitive numerical computation. As a reasonable and tractable alternative, we choose as the next best alternative the minimum L2-norm solution to Eq. (1), $\hat{\mathbf{x}}$.

Theorem 1 Suppose that the matrices A , Q and Λ are defined as above. Then the minimum norm estimate for $\mathbf{x}$ is given by $\hat{\mathbf{x}} = W\mathbf{y}$, where $W_{N \times M}$ can be written as the following:

$$W = A^\top Q \Lambda^{-1} Q^\top \quad (2)$$

Proof To find the $\mathbf{x}$ which minimizes the L2-norm $\mathbf{x}^\top \mathbf{x}$, let

$$F = \frac{1}{2} \mathbf{x}^\top \mathbf{x} - \lambda \mathbf{f}, \quad (3)$$

where $\mathbf{f} = \mathbf{A}\mathbf{x} - \mathbf{y} = 0$ is the constraint that satisfies (1), and λ is a vector of Lagrange multipliers. Minimizing yields

$$\frac{\partial F}{\partial \mathbf{x}} = \mathbf{x} - \mathbf{A}^\top \lambda = 0 \text{ and } \frac{\partial F}{\partial \lambda} = -(\mathbf{A}\mathbf{x} - \mathbf{y}) = 0. \quad (4)$$

Since $\mathbf{A}^\top$ is not invertible we solve for $\mathbf{x}$ from $\mathbf{x} = \mathbf{A}^\top \lambda$ and use this in $\mathbf{A}\mathbf{x} - \mathbf{y} = 0$ to get

$$\mathbf{y} = \mathbf{A}\mathbf{A}^\top \lambda. \quad (5)$$

Since $\mathbf{A}\mathbf{A}^\top$ is a positive semidefinite matrix, it admits the eigendecomposition $\mathbf{A}\mathbf{A}^\top = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ where the matrix $\mathbf{\Lambda}$ is a diagonal matrix having nonzero eigenvalues of $\mathbf{A}\mathbf{A}^\top$ on its diagonal and the columns of $\mathbf{Q}$ are normalized orthogonal eigenvectors ordered similarly. It is obvious that $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$ and it is not hard to prove that $\mathbf{Q}\mathbf{Q}^\top \mathbf{A} = \mathbf{A}$ (The proof is similar to that of Proposition 5(iii) of this paper). Multiplying on the left by $\mathbf{A}^\top \mathbf{Q}\mathbf{\Lambda}^{-1} \mathbf{Q}^\top$ yields the minimum norm solution:

$$\mathbf{W}\mathbf{y} = \mathbf{A}^\top \mathbf{Q}\mathbf{\Lambda}^{-1} \mathbf{Q}^\top \mathbf{y} = \mathbf{A}^\top \lambda = \hat{\mathbf{x}}. \quad (6)$$

The derivation of a simple form for the matrix $\mathbf{W} = \mathbf{A}^\top \mathbf{Q}\mathbf{\Lambda}^{-1} \mathbf{Q}^\top$, which generates the minimum norm $\mathbf{x}$ from the observed gapped k-mer counts $\mathbf{y}$, by $\hat{\mathbf{x}} = \mathbf{W}\mathbf{y}$, is the main result of this paper.

It is worth noting that the minimum L2-norm $\mathbf{x}$ can also be thought of as the minimum mean square error or the most likely distribution under the assumption that the x_j 's are independent and have a joint normal prior probability distribution with equal variances and expected values for all the ℓ -mer counts:

$$F(\mathbf{x}) = (2\pi)^{-\frac{N}{2}} |\Sigma_{\mathbf{x}}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^\top \Sigma_{\mathbf{x}}^{-1}(\mathbf{x} - \bar{\mathbf{x}})} \quad (7)$$

where $\Sigma_{\mathbf{x}} = \sigma^2 \mathbf{I}_N$ is a diagonal matrix with constant elements on the diagonal and $\bar{\mathbf{x}}$ is a constant vector, i.e. $x_i = c$. The $\mathbf{x}$ that maximizes (7) subject to the constraints given by (1) turns out to be the minimum norm solution. The proof for this is very similar to the proof of Theorem 1, as follows.

Proof Applying the Lagrange multipliers technique to find the $\mathbf{x}$ that maximizes the logarithm of $F(\mathbf{x})$ subject to the constraints given in (1) we have:

$$\mathbf{x} - \bar{\mathbf{x}} = \mathbf{A}^\top \lambda. \quad (8)$$

Reordering (8) and applying (1) we obtain

$$\mathbf{y} = A A^\top \lambda + A \bar{\mathbf{x}}. \quad (9)$$

Now, consider the following eigendecomposition for $A A^\top$:

$$A A^\top = Q \Lambda Q^\top. \quad (10)$$

Multiplying both sides of (9) by $A^\top Q \Lambda^{-1} Q^\top$ and applying (10) we obtain

$$A^\top Q \Lambda^{-1} Q^\top (\mathbf{y} - A \bar{\mathbf{x}}) = A^\top Q \Lambda^{-1} Q^\top A A^\top \lambda = A^\top \lambda. \quad (11)$$

Reordering (11) and applying (8) we obtain

$$\hat{\mathbf{x}} = W(\mathbf{y} - A \bar{\mathbf{x}}) + \bar{\mathbf{x}}. \quad (12)$$

In our case there is no difference between the expected values for different ℓ -mer counts, i.e. $x_i^- = c$. Also it can be shown that the sum of the elements in rows of matrix $W A$ is equal to 1 (see Proposition 6), therefore, in this case $W A \bar{\mathbf{x}} = I \bar{\mathbf{x}}$, and equation (12) can be simplified to

$$\hat{\mathbf{x}} = W \mathbf{y} = A^\top Q \Lambda^{-1} Q^\top \mathbf{y}. \quad (13)$$

hence $W = A^\top Q \Lambda^{-1} Q^\top$, as required. Note that $\bar{\mathbf{x}}$ is independent of $\mathbf{x}^-$ only if all the ℓ -mers have equal expected counts.

The derivation of an explicit form for the matrix W , which is shown to be the Moore–Penrose pseudoinverse of A and maps gapped k -mer counts to ungapped ℓ -mer count estimates, is the central result of this paper. Although Eq. (2) gives a method to obtain the weight matrix W from the eigendecomposition of matrix $A A^\top$ given in (10), numerical calculation of the eigenvectors for $A A^\top$ would be computationally expensive as the size of matrix A grows very rapidly for large values of ℓ , k , and b . For example, for $(\ell = 15, k = 7, b = 4)$, we have $N \approx 10^9$, and $M \approx 10^8$. However, considering the symmetry of matrix A , we will show in the detailed proof that follows that the matrix W has a simple structure. In this matrix, the entry $w_{i,j}$ only depends on the number of mismatches between the ℓ -mer u_i and the gapped- k -mer v_j . So there exists a finite sequence of only $k + 1$ values $w_0, w_1, \dots, w_k$ such that $w_{i,j} = w_m$ if u_i and v_j have exactly m mismatches. A *mismatch* is defined to be a difference between a gapped k -mer and an ℓ -mer in an ungapped position. So, for example, if N denotes a gap, $N A N G$ and $A C G G$ have one mismatch, and $A N G G$ and $A C G G$ have zero mismatches. w_m is largest for $m = 0$ and typically becomes negative for $m = 1$, and then oscillates about zero. See Sect. 7 for a concrete example. Thus, the entries of matrix W are limited to a small set of values $\{w_0, \dots, w_k\}$ and these values are specified by the following theorem:

Theorem 2 *The values of the elements of matrix W are given by the following equation, in which, ℓ is the sequence length, b is the size of the alphabet, k is the number of known bits, and m is the number of mismatches between the corresponding ℓ -mer u_i and the gapped- k mer v_j :*

$$w_{\ell,k,m} = \frac{\binom{k-\ell}{m}}{b^\ell \binom{\ell}{k} \binom{k}{m}} \sum_{n=0}^{k-m} \binom{\ell}{n} (b-1)^n \quad (14)$$

Note that matrix W clearly depends on ℓ, k and b but for fixed ℓ, k and b , the entry on row i and column j of this matrix only depends on m , the number of mismatches between v_i and u_j , i.e. differences between ungapped positions in the gapped k -mer and the ungapped ℓ -mer. Hence elements of W are limited to a small set of $k+1$ values, as specified by the above theorem, and is very simple and easy to compute.

2.1 Outline of Proof of Theorem 2

In the following sections we will prove Theorem 2 by finding an eigendecomposition for matrix $AA^\top$ and then applying (2). Although the concepts and propositions used in this proof begin from Sect. 3, the reader could accept several identities and combinatorial arguments and go directly to the main statement of the proof in Sect. 5.

A brief outline of the proof follows. To obtain an eigendecomposition for $AA^\top$, we define a new function v in Sect. 3.4 on pairs of letters; each letter is an element of the alphabet consisting of the symbol g (denoting a gap) and nonnegative integers. The function v is then naturally extended to the product set, that is to the set of pairs of words of the same length. Propositions 2 and 3 study some related identities. The reader can skip Proposition 3 and move to Sect. 4, where the desired eigendecomposition is given in Proposition 4. The proof of this proposition contains two main points: The eigenvectors given in Proposition 2 (iii) are pairwise orthogonal (this is concluded from the recursive structure of the proposed matrices) and the given eigenvalues are all of the nonzero eigenvalues of $AA^\top$. At this point, the reader may have a look at $W_{3,2}$ in Sect. 7 as a useful example. After passing rapidly through Propositions 5 and 6, the essential deduction appears in the proof of Proposition 7, whose statement is just an equivalent form of Theorem 2. This proof takes advantage of several other identities developed earlier in the paper. The reader who does not require these details may focus on the illustrative example given in Sect. 7 for the simple matrix $W_{3,2}(2)$.

3 Preliminaries and notation

In this section, first we review some linear algebra definitions and theorems in Sect. 3.1, then provide some binomial identities in Sect. 3.2, then define some alphabets and word sets and matrix A in Sect. 3.3. Finally in Sect. 3.4 we define function v and prove some identities that enable us to find an eigendecomposition for matrix $AA^\top$.

3.1 Review of some linear algebra concepts

Let F be a field and $A \in F^{m \times n}$ be a matrix. The row space of A is denoted as $\text{row}(A)$, the column space of A is denoted as $\text{col}(A)$, and the dimension of the row space (which is the same as the dimension of the column space) of A is denoted as $\text{rank}(A)$. The kernel of A , denoted as $\ker(A)$, is the space of all column vectors x satisfying $Ax = \mathbf{0}$ and the dimension of this space is called the nullity of A and denoted as $\text{null}(A)$. It is known that $\text{null}(A) + \text{rank}(A) = n$. Let $B \in F^{n \times n}$. The characteristic polynomial of B is defined as $p_B(z) = \det(zI - B)$. An element $\lambda \in F$ is an eigenvalue of B if there exists a nonzero column vector x satisfying $Bx = \lambda x$; The vector x is called an eigenvector of B . It is observed that λ is an eigenvalue of B if and only if it is a root of the characteristic equation $p_B(z)$. For an eigenvalue λ , the space $\ker(B - \lambda I)$ is called the eigenspace of B corresponding to λ . The algebraic multiplicity of an eigenvalue λ , denoted as $\alpha(\lambda)$, is the multiplicity of the root λ of $p_B(z)$. The geometric multiplicity of an eigenvalue λ , denoted by $\gamma(\lambda)$, is the dimension of $\ker(B - \lambda I)$. The matrix B is called diagonalizable if there exists a nonsingular matrix $P \in F^{n \times n}$ such that $B = P\Lambda_0 P^{-1}$ for some diagonal matrix $\Lambda_0 \in F^{n \times n}$. A matrix $B \in \mathbb{R}^{n \times n}$ is known to be diagonalizable by a nonsingular matrix P if and only if the eigenvalues of B are real and $\alpha(\lambda) = \gamma(\lambda)$ for all eigenvalues λ of B . It is deduced that a diagonalizable matrix B has a diagonal decomposition $B = P\Lambda_0 P^T$ where P is a unitary matrix whose columns contain normalized orthogonal eigenvectors of B and Λ_0 is a diagonal matrix having eigenvalues of B on its diagonal, ordered according to the columns of P . From $B = P\Lambda_0 P^T$ by simple matrix calculations, we conclude that $B = Q\Lambda Q^T$, where the matrix Q is obtained by deleting the columns of P which are in $\ker(B)$, and Λ is obtained by deleting the zero columns of Λ_0 .

For a matrix $A \in \mathbb{C}^{m \times n}$ its Hermitian adjoint, A^* , is its conjugate transpose, i.e. A^* is an $n \times m$ matrix with $A^*(i, j) = \overline{A(j, i)}$. A matrix A is Hermitian if $A = A^*$, thus a real matrix A is Hermitian if and only if it is symmetric. A Hermitian matrix A is positive definite (resp. positive semidefinite) if $x^* Ax > 0$ (resp. $x^* Ax \geq 0$) for all nonzero $x \in \mathbb{C}^n$. It is concluded that a square real symmetric matrix A of order n is positive definite (resp. positive semidefinite) if $x^T Ax > 0$ (resp. $x^T Ax \geq 0$) for all nonzero $x \in \mathbb{R}^n$. For any matrix A , the matrix $A^T A$ is positive semidefinite, and $\text{rank}(A) = \text{rank}(AA^T)$. Conversely, any Hermitian positive semidefinite matrix M can be written as $M = A^* A$; this is the Cholesky decomposition.

The Moore–Penrose pseudoinverse of a matrix A , denoted by A^+ , is defined as a matrix that satisfies all the following four conditions:

$$AA^+A = A \quad A^+AA^+ = A^+ \quad (AA^+)^* = AA^+ \quad (A^+A)^* = A^+A.$$

The Moore–Penrose pseudoinverse exists and is unique for any given matrix A . Also we have $A^+ = (A^*A)^+A^* = A^*(AA^*)^+$. For further properties of the Moore–Penrose pseudoinverse see for instance Albert (1972).

3.2 Some binomial identities

For real number x and nonnegative integer n , the *falling factorial* $(x)_n$ is defined as $x(x-1) \cdots$

$(x-n+1)$ and the *binomial coefficient* $\binom{x}{n}$ is defined as $\frac{(x)_n}{n!}$. The binomial coefficients satisfy many interesting identities, some of which are discussed in this section [for further information see Graham et al. (1994)]. The following examples are among the well-known identities used frequently in this paper:

$$\binom{r}{s} \binom{s}{t} = \binom{r}{t} \binom{r-t}{s-t}, \quad \sum_{i=0}^k \binom{i+n}{i} = \binom{k+n+1}{k}, \quad \sum_{i=0}^k (-1)^i \binom{n}{i} = \binom{k-n}{k}.$$

Two more identities are given in the following Proposition.

Proposition 1 (i) Let $0 \leq t \leq n \leq p \leq k$ be integers. Then the following identity holds

$$\binom{k}{p} \binom{p}{t} \binom{k-p}{n-t} = \binom{k}{n} \binom{n}{t} \binom{k-n}{p-t} \quad (15)$$

(ii) Let $0 \leq p \leq k \leq \ell$ be integers. Then the following identity holds

$$\sum_{n=0}^k \sum_{t=0}^n (-1)^{n-t} \binom{\ell}{n} \binom{n}{t} \binom{k-n}{p-t} x^t = \binom{k-\ell}{k-p} \sum_{n=0}^p \binom{\ell}{n} x^n. \quad (16)$$

Proof See Appendix.

3.3 Some more definitions and notations

Since we work with sequences of given symbols (or words over a given alphabet), some definitions in this respect are mentioned. Let $b > 1$ be an integer; as usual, we define the b -ary alphabet by $\Sigma_b = \{0, 1, \dots, b-1\}$. In addition to this alphabet, two other alphabets, defined as follows, are frequently used in this work:

$$\Delta_b = \Sigma_b \cup \{g\} \quad \Gamma_b = \Delta_b \setminus \{0\}$$

The symbol g used above represents “gap” (or wildcard) symbol. Hence, a gapped k -mer v of length $|\ell|$ is a sequence v of length ℓ over $\Sigma_b \cup \{g\}$ with exactly $\ell - k$ occurrences of symbol g . We add some useful definitions from the literature to simplify the representation of our problem here. A word w on a given alphabet $\mathcal{A}$ is a finite sequence $w = w_1 w_2 \cdots w_n$ with $w_i \in \mathcal{A}$. The length of w , the integer n , is denoted as $|w|$. The occurrences of a symbol i in a word w is denoted as $|w|_i$. Five important sets of words are defined below:

$$U_\ell = \Sigma_b^\ell \quad V_{\ell k} = \mathcal{A} \{w \in \Delta_b^\ell : |w|_g = \ell - k\} \quad V'_{\ell k} = \{w \in \Gamma_b^\ell : |w|_g = \ell - k\} \quad V_{\ell, \leq k} = \bigcup_{m=0}^k V_{\ell m} \quad V'_{\ell, \leq k} = \bigcup_{m=0}^k V'_{\ell m}$$

For a set X and a nonnegative integer n , by $\binom{X}{n}$, we mean the set of all n -element subsets of X . For a word $v' \in \Gamma_b^*$, we let $G_{v'} = \{i: 1 \leq i \leq |v'|, v'_i = g\}$ and $G_{v'}^- = \{1, \dots, |v'|\} \setminus G_{v'}$. Thus if $v' \in \Gamma_b^\ell$, then we have $|G_{v'}^-| = n$ if and only if $v' \in V_{\ell n}'$. We say elements $u \in U_\ell$ and $v \in V_{\ell k}$ match if for any $1 \leq i \leq \ell$ with $v_i = g$ we have $u_i = v_i$. This is simply the formal statement of Definition 3. The set of the elements $v \in V_{\ell k}$ which are matchable with a given element $u \in U_\ell$ is denoted by $M_{\ell k}(u)$. The set of elements $u \in U_\ell$ which are matchable with a given element $v \in V_{\ell k}$ is denoted by $M_{\ell k}'(v)$.

We define the matrix $A_{\ell k}$ as a $(0, 1)$ -matrix whose rows (columns) are indexed by elements of $V_{\ell k}$ (elements of U_ℓ) and with $A_{\ell k}(v, u) = 1$ if and only if u and v match. Thus we have $A_{\ell k}(v, u) = 1$ if and only if $v \in M_{\ell k}(u)$ if and only if $u \in M_{\ell k}'(v)$. If we interpret each element $v \in V_{\ell k}$ as the set of all elements $u \in U_\ell$ which are matchable with v , then an incidence structure with the point-set U_ℓ and the block-set $V_{\ell k}$ is obtained in which every block consists of $b^{\ell-k}$ points and every point occurs in $\binom{\ell}{n}$ blocks. The matrix $A_{\ell k}^\top$ then appears as the incidence matrix of this incidence structure which is simultaneously the inclusion matrix of elements of U_ℓ vs. the elements of $V_{\ell k}$.

In this paper we prove that $A_{\ell k} A_{\ell k}^\top$ is diagonalizable and based on an explicit eigendecomposition. We find the Moore–Penrose pseudo-inverse of $A_{\ell k}$ and we provide a formula for the entries of this matrix.

3.4 The function ν and related identities

For many applications it is useful to study the matrix $A_{\ell i} A_{\ell j}^\top$. The rows and the columns of this matrix are respectively indexed by elements of $V_{\ell k}$ and the entry $(A_{\ell k} A_{\ell k}^\top)(v, w)$ has a simple combinatorial interpretation: the value $(A_{\ell k} A_{\ell k}^\top)(v, w)$ counts the number of elements of U_ℓ which match both v and w . In this section we present an orthogonal basis for the eigenspace of $A_{\ell k} A_{\ell k}^\top$. For this purpose, first we introduce the definition of the function $\nu: b \times b \rightarrow \mathbb{Z}$ which has an essential role in the rest of this section. Then by using some identities we find the eigenvalues and the corresponding eigenvectors. We extend the natural order on (nonnegative) integers to $\mathbb{Z} \cup \{g\}$ using the convention $i < g$ for any integer i . Then we define the function $\nu: b \times b \rightarrow \mathbb{Z}$ as follows:

$$\nu(x, y) = \begin{cases} 1 & \text{if } x < y \text{ or } x = y = g, \\ 0 & \text{if } x > y, \\ -y & \text{if } x = y \neq g. \end{cases}$$

For any positive integer ℓ , this definition is naturally extended to $\nu: \Delta_b^\ell \times \Delta_b^\ell \rightarrow \mathbb{Z}$ by a simple product rule as follows:

$$\nu(w_1 \cdots w_\ell, v'_1 \cdots v'_\ell) = \prod_{i=1}^{\ell} \nu(w_i, v'_i).$$

Let n, k and ℓ be integers with $0 \leq n \leq k \leq \ell$. Recall that $V'_{\ell n}$ is the set of all words over the alphabet $\Gamma_b = \{1, 2, \dots, b-1\} \cup \{g\}$ with exactly $\ell - n$ symbols g . Below, we define two column vectors which are very important in this section.

Definition 6 Let $v' \in V'_{\ell n}$. The vector $x_{v'}^{\ell kn}$ is a column vector whose rows are indexed by the elements of $V_{\ell k}$ with entries $x_{v'}^{\ell kn}(w) = \nu(w, v')$. The column vector $z_{v'}^{\ell n}$ is defined by $z_{v'}^{\ell n} = x_{v'}^{\ell \ell n}$. In other words, $z_{v'}^{\ell n}$ is a column vector whose rows are indexed by elements u of U_{ℓ} with entries $z_{v'}^{\ell n}(u) = \nu(u, v')$. When there is no need to emphasize the parameters ℓ, k and n , we simply write $x_{v'}$ and $z_{v'}$.

Proposition 2 Let $n \leq k \leq \ell$ and $v' \in V'_{\ell n}$. The following matrix identities hold.

- i. $A_{\ell k}^{\top} x_{v'}^{\ell kn} = \binom{\ell - n}{\ell - k} z_{v'}^{\ell n}.$
- ii. $A_{\ell k} z_{v'}^{\ell n} = b^{\ell - k} x_{v'}^{\ell kn}.$
- iii. $A_{\ell i} A_{\ell j}^{\top} x_{v'}^{\ell j n} = b^{\ell - i} \binom{\ell - n}{\ell - j} x_{v'}^{\ell i n}.$
- iv. $A_{\ell k}^{\top} A_{\ell k} z_{v'}^{\ell n} = b^{\ell - k} \binom{\ell - n}{\ell - k} z_{v'}^{\ell n}.$

Proof See Appendix.

Remark 1 The definition of the function ν and its extension and the inverse type formulas of parts (i) and (ii) of the previous proposition show similarities with properties of the Möbius function (see Chapter 6 of Cameron 2003 for instance).

Proposition 3 Let $u \in U_{\ell}$ and $v \in V_{\ell k}$. Let $P = \{i : 1 \leq i \leq \ell, v_i = g, v_i = u_i\}$, $Q = \{i : 1 \leq i \leq \ell, v_i = g, v_i = u_i\}$, $|P| = p$ (and consequently $|Q| = k - p$).

- i. For $1 \leq i \leq \ell$ let

$$\phi_i(u, v) = \sum_{j=\max\{1, v_i\}}^{b-1} \frac{\nu(v_i, j) \nu(u_i, j)}{j(j+1)}.$$

Then we have

$$\phi_i(u, v) = \begin{cases} \frac{b-1}{b}, & \text{if } i \in P; \\ \frac{-1}{b}, & \text{otherwise, i.e. if } i \in Q. \end{cases}$$

- ii. The following identity holds

$$\sum_{v' \in V'_{\ell n}} \frac{\nu(v, v') \nu(u, v')}{\prod_{i \in \bar{G}_{v'}} v'_i (v'_i + 1)} = \frac{1}{b^n} \sum_{t=0}^n (-1)^{n-t} \binom{p}{t} \binom{k-p}{n-t} (b-1)^t, \quad (17)$$

Proof See Appendix.

4 The eigendecomposition of $A_{\ell k} A_{\ell k}^\top$

Definition 7 Given b , we define ℓ_k as a matrix whose rows and columns are indexed by the elements of $V_{\ell k}$ and $V'_{\ell, \leq k}$ respectively and whose entries are given by $\ell_k(w, v') = v(w, v')$. In other words, the columns of ℓ_k are exactly the vectors $x_{v'}$ for $v' \in V'_{\ell, \leq k}$. For $0 \leq m \leq k$, we define the matrix $\ell_{k;m}$ as the submatrix of ℓ_k which is obtained by retaining the columns indexed by elements $v' \in V'_{\ell m}$ and omitting the other columns. Consequently, we have the following block decomposition:

$$\Delta_{\ell, k} = [\Delta_{\ell, k; 0}, \Delta_{\ell, k; 1}, \dots, \Delta_{\ell, k; k}]$$

The block decomposition of $\ell_{k;m}$ is given in the proof of the following proposition.

Proposition 4 The matrix $A_{\ell k} A_{\ell k}^\top$ has exactly $\sum_{n=0}^k \binom{\ell}{n} (b-1)^n$ nonzero eigenvalues as follows: For any n , $0 \leq n \leq k$, there are $\binom{\ell}{n} (b-1)^n$ eigenvalues equal to $\lambda_n = \binom{\ell-n}{k-n} b^{\ell-k}$ and the eigenvectors corresponding λ_n are of the form $x_{v'}^{\ell, k, n}$ with $v' \in V'_{\ell n}$. These eigenvectors are pairwise orthogonal and $A_{\ell k} A_{\ell k}^\top$ is diagonalizable.

Proof By using Proposition 2 (iii), each $x_{v'}^{\ell, k, n}$ is an eigenvector of $A_{\ell k} A_{\ell k}^\top$ corresponding to the eigenvalue $\binom{\ell-n}{k-n} b^{\ell-k}$. Moreover, for different words $u' \in V'_{\ell, n_1}$ and $u' \in V'_{\ell, n_2}$, the vectors $x_{u'}^{\ell, k, n_1}$ and $x_{u'}^{\ell, k, n_2}$ are orthogonal as it is proved in the sequel. To prove the orthogonality of the eigenvectors, it is enough to prove that for any integer ℓ , for all integers k, m, n with $0 \leq m, n \leq k$ the matrix products $\Delta_{\ell k; n}^\top \Delta_{\ell k; m}$ are diagonal if $n = m$ and $\mathbf{0}$ otherwise. We prove this by induction on ℓ . For $\ell = 0$ there exists just one matrix $\Delta_{0, k; n}$, so there is nothing to prove. For $\ell = 1$, either $k = 0$ or $k = 1$. In the first case the matrix $\Delta_{1, 0; 0}$ has just one entry, so there is nothing to prove. In the case $k = 1$ we have the following matrices and it is straightforward to check that the matrices $\Delta_{1, 1; n}^\top \Delta_{1, 1; m}$ are diagonal if $n = m$ and $\mathbf{0}$ otherwise.

$$\Delta_{1,1;0} = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ b-1 \end{pmatrix} \begin{pmatrix} g \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \Delta_{1,1;1} = \begin{pmatrix} 0 \\ 1 \\ i-1 \\ i \\ \vdots \\ b-1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & \cdots & i & \cdots & \cdots & b-1 \\ 1 & \cdots & \cdots & 1 & \cdots & \cdots & 1 \\ -1 & \ddots & \ddots & 1 & \ddots & \ddots & 1 \\ 0 & \ddots & \ddots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -i & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & \cdots & 0 & -(b-1) \end{pmatrix}$$

Before continuing the proof, note that from the definition of $\ell, k; m$ the block decomposition of this matrix is concluded as follows, where $0 \dots$, for instance, stands for the set of all

indexes commencing with 0. Note that the matrix $\mathbf{0}_{\ell, k; n}$ is an $\binom{\ell}{k} b^k \times \binom{\ell}{n} (b-1)^n$ matrix with all zero entries. Thus $\ell, k; m$ has the following recursive structure:

$$\begin{pmatrix} 0 \dots & g \dots & 1 \dots & \dots & i \dots & \dots & \dots & b-1 \dots \\ \Delta_{\ell-1, k-1; m} & \Delta_{\ell-1, k-1; m-1} & \cdots & \Delta_{\ell-1, k-1; m-1} & \cdots & \cdots & \cdots & \Delta_{\ell-1, k-1; m-1} \\ 1 \dots & \Delta_{\ell-1, k-1; m} & -\Delta_{\ell-1, k-1; m-1} & \ddots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \mathbf{0}_{\ell-1, k-1; m-1} & \Delta_{\ell-1, k-1; m-1} & \ddots & \ddots & \ddots & \vdots \\ i \dots & \vdots & \vdots & -i \Delta_{\ell-1, k-1; m-1} & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \mathbf{0}_{\ell-1, k-1; m-1} & \ddots & \ddots & \ddots & \vdots \\ b-1 \dots & \Delta_{\ell-1, k-1; m} & \mathbf{0}_{\ell-1, k-1; m-1} & \cdots & \mathbf{0}_{\ell-1, k-1; m-1} & \cdots & \mathbf{0}_{\ell-1, k-1; m-1} & -(b-1) \Delta_{\ell-1, k-1; m-1} \\ g \dots & \Delta_{\ell-1, k; m} & \mathbf{0}_{\ell-1, k; m-1} & \cdots & \mathbf{0}_{\ell-1, k; m-1} & \cdots & \cdots & \mathbf{0}_{\ell-1, k; m-1} \end{pmatrix}$$

Now we suppose that the aforementioned property holds for $\ell-1$, which means that the matrices $\Delta_{\ell-1, k; n}^\top \Delta_{\ell-1, k; m}$ are diagonal if $n = m$ and $\mathbf{0}$ otherwise. To prove this property for ℓ , note that using the above block decomposition and some calculation, the matrix $\Delta_{\ell, k; n}^\top \Delta_{\ell, k; m}$ consists of $b \times b$ blocks, all of which are $\mathbf{0}$ when $m \neq n$; but the matrices on the diagonal are themselves diagonal when $m = n$. Considering the dimensions, this fact simply implies the required proposition for ℓ and proves the result.

Now note that the matrix $A_{\ell, k} A_{\ell, k}^\top$ has entries $b^{\ell-k}$ on its main diagonal so

$\text{trace}(A_{\ell, k} A_{\ell, k}^\top) = b^{\ell-k} b^k \binom{\ell}{k} = b^\ell \binom{\ell}{k}$ and this equals the summation of previously obtained eigenvalues, considering their multiplicities, as follows:

$$\sum_{n=0}^k \binom{\ell}{n} (b-1)^n \lambda_n = \sum_{n=0}^k \binom{\ell}{n} \binom{\ell-n}{k-n} b^{\ell-k} (b-1)^n = b^{\ell-k} \binom{\ell}{k} \sum_{n=0}^k \binom{k}{n} (b-1)^n = b^\ell \binom{\ell}{k}.$$

But the matrix $A_{\ell, k} A_{\ell, k}^\top$ is symmetric positive semi-definite, so it has no negative eigenvalues; hence, all the remaining eigenvalues are 0. We have already shown that for any nonzero

eigenvalue λ of $A_{\ell k} A_{\ell k}^\top$, the algebraic multiplicity and the geometric multiplicity of λ are the same. For $\lambda = 0$, observe that $A_{\ell k} A_{\ell k}^\top$ is a square matrix of order $\binom{\ell}{k} b^k$ and the algebraic multiplicity of 0 is $\alpha(0) = \binom{\ell}{k} b^k - \sum_{n=0}^k \binom{\ell}{n} (b-1)^n$. On the other hand, the geometric multiplicity of 0 is the nullity of $A_{\ell k} A_{\ell k}^\top$ which equals $\gamma(0) = \binom{\ell}{k} b^k - \text{rank}(A_{\ell k} A_{\ell k}^\top)$, hence $\gamma(0) = \alpha(0)$. Since for any eigenvalue λ of $A_{\ell k} A_{\ell k}^\top$, the equation $\gamma(\lambda) = \alpha(\lambda)$ holds, we conclude that this matrix is diagonalizable.

Remark 2 Again consider equation (2) and suppose that the matrix $A_{\ell k} A_{\ell k}^\top$ has an eigendecomposition of the form $A_{\ell k} A_{\ell k}^\top = P_{\ell k} \Lambda_0 P_{\ell k}^\top$ in which the columns of $P_{\ell k}$ are normalized orthogonal eigenvectors and the diagonal matrix Λ_0 has corresponding eigenvalues on its diagonal. Moreover, we may suppose that the nonzero eigenvalues appear first, thus the matrix $P_{\ell k}$ has a decomposition $P_{\ell k} = [Q_{\ell k} N_{\ell k}]$ where the columns of $Q_{\ell k}$ are the eigenvectors which do not lie in the null space of $A_{\ell k} A_{\ell k}^\top$, and the columns of $N_{\ell k}$ are the eigenvectors which lie in the null space of $A_{\ell k} A_{\ell k}^\top$.

More precisely, we present the matrix $Q_{\ell k}$ as below.

Definition 8 Based on the informal description of $Q_{\ell k}$ given in Remark 2 and the results of

Proposition 4, we define $Q_{\ell k}$ as a matrix whose columns are vectors $\frac{1}{\|x_{v'}\|} x_{v'}$, where $v' \in V'_{\ell, \leq k}$.

Proposition 5 With the above definitions, the matrix $A_{\ell k} A_{\ell k}^\top$ admits an eigendecomposition of the form $A_{\ell k} A_{\ell k}^\top = Q_{\ell k} \Lambda Q_{\ell k}^\top$. Moreover, the following properties hold:

i.
$$Q_{\ell k} = \Delta_{\ell k} E, \quad (18)$$

where $E = \text{diag}\left(\frac{1}{\|x_{v'}\|}\right)_{v' \in V'_{\ell, \leq k}}$. Moreover for $v' \in V'_{\ell n}$ we have

$$\|x_{v'}^{\ell k n}\| = \sqrt{\binom{\ell-n}{\ell-k} b^{k-n} \prod_{i \in \bar{G}_{v'}} (v'_i + v_i'^2)}. \quad (19)$$

ii. We have

$$Q_{\ell k} \Lambda Q_{\ell k}^\top = A_{\ell k} A_{\ell k}^\top, \quad (20)$$

$$Q_{\ell k} \Lambda^{-1} Q_{\ell k}^{\top} = \Delta_{\ell k} D \Delta_{\ell k}^{\top}, \quad (21)$$

$$\text{where } D = \text{diag} \left((d_{v'})_{v' \in V'_{\ell, \leq k}} \right) \text{ and}$$

$$d_{v'} = \frac{1}{\left(\frac{\ell - |\bar{G}_{v'}|}{\ell - k} \right)^2 b^{\ell - |\bar{G}_{v'}|} \prod_{i \in \bar{G}_{v'}} (v'_i + v_i'^2)}. \quad (22)$$

iii. We have

$$Q_{\ell k}^{\top} Q_{\ell k} = I, \quad (23)$$

$$Q_{\ell k} Q_{\ell k}^{\top} A_{\ell k} = A_{\ell k}. \quad (24)$$

Proof See Appendix.

5 The Moore–Penrose pseudo-inverse of $A_{\ell k}$

In this section we prove that the matrix $W_{\ell k}$, defined in Equation (2), is the Moore–Penrose pseudo-inverse of $A_{\ell k}$. We calculate the entries of this matrix:

$$W_{\ell k} = A_{\ell k}^{\top} Q_{\ell k} \Lambda^{-1} Q_{\ell k}^{\top}. \quad (25)$$

Proposition 6 *The matrix $W_{\ell k} = A_{\ell k}^{\top} Q_{\ell k} \Lambda^{-1} Q_{\ell k}^{\top}$ is the Moore–Penrose pseudoinverse of $A_{\ell k}$. Moreover, the sum of entries of any given row (/column) of the matrix $W_{\ell k} A_{\ell k}$ equals 1.*

Proof See Appendix.

Proposition 7 *Let $u \in U_{\ell}$ and $v \in V_{\ell k}$. Let $P = \{i : 1 \leq i \leq \ell, v_i = g, v_i = u_i\}$, $Q = \{i : 1 \leq i \leq \ell, v_i = g, v_i = u_i\}$, $|P| = p$ (and consequently $|Q| = k - p$). Then the value $W_{\ell k}(u, v)$ is obtained as follows*

$$W_{\ell k}(u, v) = \frac{\binom{k - \ell}{k - p}}{\binom{\ell}{k} \binom{k}{p}} \sum_{n=0}^p \binom{\ell}{n} (b - 1)^n. \quad (26)$$

Equation 26 shows that the matrix entries $W_{\ell k}(u, v)$ only depend on the number of mismatches $(k - p)$ between the gapped k -mer, v , and the ℓ -mer, u .

Proof Define the matrix $B_{\ell k}$ by

$$B_{\ell k} = Q_{\ell k} \Lambda^{-1} Q_{\ell k}^{\top}, \quad (27)$$

so $W_{\ell k} = A_{\ell k}^{\top} B_{\ell k}$. First we calculate the entries of $B_{\ell k}$. For this, note that by (21) and (22), for any $y, v \in V_{\ell k}$ we have

$$B_{\ell k}(y, v) = \sum_{v' \in V'_{\ell}, \leq k} \nu(y, v') \nu(v, v') d_{v'} = \sum_{v' \in V'_{\ell}, \leq k} \frac{\nu(y, v') \nu(v, v')}{\binom{\ell - |\bar{G}_{v'}|}{\ell - k}^2 b^{\ell - |\bar{G}_{v'}|} \prod_{i \in \bar{G}_{v'}} v'_i(v'_i + 1)}.$$

Now, fixing $u \in U_{\ell}$ and $v \in V_{\ell k}$, by using $W_{\ell k} = A_{\ell k}^{\top} B_{\ell k}$, we have

$$W_{\ell k}(u, v) = \sum_{y \in M_{\ell k}(u)} B_{\ell k}(y, v),$$

which yields

$$\begin{aligned} W_{\ell k}(u, v) &= \sum_{y \in M_{\ell k}(u)} \sum_{v' \in V'_{\ell}, \leq k} \frac{\nu(y, v') \nu(v, v')}{\binom{\ell - |\bar{G}_{v'}|}{\ell - k}^2 b^{\ell - |\bar{G}_{v'}|} \prod_{i \in \bar{G}_{v'}} v'_i(v'_i + 1)} \\ &= \sum_{n=0}^k \frac{1}{\binom{\ell - n}{\ell - k}^2 b^{\ell - n}} \sum_{v' \in V'_{\ell n}} \frac{\nu(v, v')}{\prod_{i \in \bar{G}_{v'}} v'_i(v'_i + 1)} \sum_{y \in M_{\ell k}(u)} \nu(y, v') \\ &= \sum_{n=0}^k \frac{1}{\binom{\ell - n}{\ell - k}^2 b^{\ell - n}} \sum_{v' \in V'_{\ell n}} \frac{\nu(v, v') \nu(u, v')}{\prod_{i \in \bar{G}_{v'}} v'_i(v'_i + 1)} \text{ (by (35))} \\ &= \frac{1}{b^{\ell}} \sum_{n=0}^k \frac{1}{\binom{\ell - n}{\ell - k}} \sum_{t=0}^n (-1)^{n-t} \binom{p}{t} \binom{k-p}{n-t} (b-1)^t \text{ (by (17))} \\ &= \frac{1}{b^{\ell}} \binom{k}{p} \sum_{n=0}^k \frac{1}{\binom{\ell - n}{\ell - k}} \sum_{t=0}^n (-1)^{n-t} \binom{k}{p} \binom{p}{t} \binom{k-p}{n-t} (b-1)^t \\ &= \frac{1}{b^{\ell}} \binom{k}{p} \sum_{n=0}^k \frac{\binom{k}{n}}{\binom{\ell - n}{\ell - k}} \sum_{t=0}^n (-1)^{n-t} \binom{n}{t} \binom{k-n}{p-t} (b-1)^t \text{ (by (15))} \\ &= \frac{1}{b^{\ell}} \binom{k}{p} \sum_{n=0}^k \frac{\binom{\ell}{n}}{\binom{\ell}{k}} \sum_{t=0}^n (-1)^{n-t} \binom{n}{t} \binom{k-n}{p-t} (b-1)^t \\ &= \frac{1}{b^{\ell}} \binom{k}{p} \binom{\ell}{k} \sum_{n=0}^k \sum_{t=0}^n (-1)^{n-t} \binom{\ell}{n} \binom{n}{t} \binom{k-n}{p-t} (b-1)^t \\ &= \frac{\binom{k-\ell}{k-p}}{b^{\ell} \binom{k}{p} \binom{\ell}{k}} \sum_{n=0}^p \binom{\ell}{n} (b-1)^n, \text{ (by (16))}, \end{aligned}$$

as required.

Remark 3 Alternatively, one can use an eigendecomposition for $A^\top A$, given by $A^\top A = Q' \Lambda Q'^\top$ and obtain $W = Q' \Lambda^{-1} Q'^\top A^\top$. Proposition 2 can be used to show that $z_{v'}^{\ell n}$, $(0 \leq n \leq k)$ defined in Definition 6 are eigenvectors for the matrix $A^\top A$, where $x_{v'}^{\ell kn}$ are eigenvectors for $AA^\top$ (corresponding nonzero eigenvalues).

Now we are at the point to prove Theorem 2:

Proof of Theorem 2 Considering Proposition 7 and setting $p = k - m$, the result is obtained.

Corollary 1 With notations of the previous Proposition, suppose that $u \in U_\ell$ and $v \in V_{\ell k}$

have exactly m mismatches and define $w'_{\ell km}$ by $w'_{\ell km} = b^\ell \binom{\ell}{k} W_{\ell k}(u, v)$. Then the values $w'_{\ell km}$ satisfy the following recursive equations.

$$w'_{\ell, k, 0} = w'_{\ell-1, k, 0} + (b-1)w'_{\ell-1, k-1, 0} \quad (28)$$

$$w'_{\ell, k, m} = \frac{k-\ell}{k} w'_{\ell, k-1, m-1}, \quad 0 < m \leq k. \quad (29)$$

Proof Since u and v have m mismatches, using notations of Proposition 7, we have $|Q| = k - p = m$, so $p = k - m$. Hence,

$$w'_{\ell, k, m} = \frac{\binom{k-\ell}{m}}{\binom{k}{m}} \sum_{n=0}^{k-m} \binom{\ell}{n} (b-1)^n. \quad (30)$$

Now, if $m = 0$ then $w'_{\ell, k, 0} = \sum_{n=0}^k \binom{\ell}{n} (b-1)^n$ and (28) is easily concluded. To prove the other recurrence relation, suppose that $m > 0$ and observe that by (30),

$$\frac{w'_{\ell, k, m}}{w'_{\ell, k-1, m-1}} = \frac{\binom{k-\ell}{m}}{\binom{k-1-\ell}{m-1}} \frac{\binom{k-1}{m-1}}{\binom{k}{m}} = \frac{k-\ell}{k}.$$

This concludes the derivation of the matrix $W_{\ell k}$.

6 A basis for $\text{row}(A_{\ell k})$

This section gives some additional insight about the rank of matrix A and the dependencies between rows of A .

It is easily proved that the matrix $A_{\ell k}$ admits the following block decomposition, where $0 \cdots$, for instance, stands for the set of all indexes commencing with 0.

$$A_{\ell,k} = \begin{array}{c} 0 \cdots \quad 1 \cdots \quad \cdots \quad b-1 \cdots \\ \begin{pmatrix} A_{\ell-1,k-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & A_{\ell-1,k-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & A_{\ell-1,k-1} \end{pmatrix} \\ b-1 \cdots \\ g \cdots \end{array}$$

This decomposition gives some information about the row space of $A_{\ell,k}$. Here we propose an explicit basis for this space, but before this we need to define another matrix.

Definition 9 The matrix $A_{\ell,k}$ is a $(0, 1)$ matrix with rows and columns indexed by the elements of U_ℓ and $V_{\ell,k}$, respectively and with entries satisfying $A_{\ell,k}(u, v) = 1$ if and only if u is matchable with v . In other words, $A_{\ell,k}$ is the matrix obtained by combining the matrices $A_{\ell,i}$, $i = 0, \dots, k$, vertically. The row of $A_{\ell,k}$ indexed by a word $w \in V_{\ell,k}$, is denoted by $\mathbf{r}_w$.

Proposition 8 The matrix $A_{\ell,k}$ has the same row space as $A_{\ell,k}$, hence

$$\text{rank}(A_{\ell, \leq k}) = \sum_{n=0}^k \binom{\ell}{n} (b-1)^n. \text{ Moreover, the set } \{\mathbf{r}_{w'}\}_{w' \in V'_{\ell,k}} \text{ constitutes a basis for } \text{row}(A_{\ell,k}).$$

Proof First we prove that any row of $A_{\ell,n-1}$ is expressible as a summation of some rows of $A_{\ell,n}$. For this, consider a word $w \in V_{\ell,n-1}$ and with no loss of generality, suppose that w begins with g , say $w = gv$ for some $v \in V_{\ell-1,n-1}$. It is easy to observe that

$$\mathbf{r}_{gv} = \sum_{t=0}^{b-1} \mathbf{r}_{tv}. \quad (31)$$

Thus, the row space of $A_{\ell,n-1}$ is contained in the row space of $A_{\ell,n}$, and we provide $\text{row}(A_{\ell,k}) = \text{row}(A_{\ell,k})$, so $\text{rank}(A_{\ell,k}) = \text{rank}(A_{\ell,k})$. Based on this, using the fact $\text{rank}(A_{\ell,k}) = \text{rank}(A_{\ell,k}^T A_{\ell,k})$ and Proposition 4, we obtain

$$\text{rank}(A_{\ell, \leq k}) = \sum_{n=0}^k \binom{\ell}{n} (b-1)^n.$$

Now observe that (31) can also be written as

$$\mathbf{r}_{0v} = \mathbf{r}_{gv} - \sum_{t=1}^{b-1} \mathbf{r}_{tv} \quad (32)$$

and by consecutive applications of this equation, we conclude that for any word $w = 0^m v$ with $m > 0$ and $v \in V'_{\ell-m, k-m}$, the following equation holds

$$\mathbf{r}_{0^m v} = \sum_{x \in \Gamma_b^m} (-1)^{|x|_g} \mathbf{r}_{xv}. \quad (33)$$

Note that the words xv in the right, all belong to $V'_{\ell k}$. We conclude that for any $w \in V_{\ell k} \setminus V'_{\ell k}$, $\mathbf{r}_w$ can be written as a linear combination of some rows $\mathbf{r}_{v'}$ with $v' \in V'_{\ell k}$. Now the set of vectors $(\mathbf{r}_{v'})_{v' \in V'_{\ell k}}$ generates $\text{row}(A_{\ell, k})$ and has the same number of elements as the dimension of this space, so it is a basis for $\text{row}(A_{\ell, k})$, as required.

Remark 4 Some properties of our matrix $A_{\ell k}$ are similar to the set inclusion matrix of t -subsets vs. k -subsets; see for instance Wilson (1990).

7 Numerical example

In this section we give a numerical example for matrices A and W for a specific simple choice of b , ℓ , and k suitably small so that it is possible to provide the entire matrix. We also give the column vectors $z_{v'}^{\ell n}$ and $x_{v'}^{\ell kn}$ (the eigenvectors of $A^\top A$ and $AA^\top$ respectively).

For simplicity let $b = 2$, $\ell = 3$, and $k = 2$. Then the matrix $A_{\ell k}$ is:

$$A_{3,2} = \begin{matrix} & \begin{pmatrix} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \end{pmatrix} \\ \begin{matrix} 00g \\ 01g \\ 0g0 \\ 0g1 \\ 10g \\ 11g \\ 1g0 \\ 1g1 \\ g00 \\ g01 \\ g10 \\ g11 \end{matrix} & \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}$$

Here the rows and columns orders are chosen to match the block decomposition given in Sect. 6, and g represents a gap. The rank of this matrix is $\text{rank}(A) = 7$ which is less than $N = b^\ell = 8$ and the system $\mathbf{y} = A\mathbf{x}$ is underdetermined.

The integer n satisfies $0 \leq n \leq k = 2$, so it takes values 0, 1 and 2.

- i. Let $n = 0$. Then the only possible word for v' is ggg and we have

$$z_{ggg}^{3,0} = \begin{matrix} 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{matrix} \begin{pmatrix} ggg \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad x_{ggg}^{3,2,0} = \begin{matrix} 00g \\ 01g \\ 0g0 \\ 0g1 \\ 10g \\ 11g \\ 1g0 \\ 1g1 \\ g00 \\ g01 \\ g10 \\ g11 \end{matrix} \begin{pmatrix} ggg \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

It can be verified that $A_{3,2}^\top x_{ggg}^{3,2,0} = 3z_{ggg}^{3,0}$ and $A_{3,2} z_{ggg}^{3,0} = 2x_{ggg}^{3,2,0}$. Thus $A_{3,2} A_{3,2}^\top x_{ggg}^{3,2,0} = 6x_{ggg}^{3,2,0}$.

- ii. Now, let $n = 1$. The word v' can be one of $1gg$, $g1g$ or $gg1$. Consider the case $v' = 1gg$ for instance. The column vectors $z_{v'}^{\ell n}$ and $x_{v'}^{\ell kn}$ are as follow

$$z_{1gg}^{3,1} = \begin{matrix} 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{matrix} \begin{pmatrix} 1gg \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{pmatrix} \quad x_{1gg}^{3,2,1} = \begin{matrix} 00g \\ 01g \\ 0g0 \\ 0g1 \\ 10g \\ 11g \\ 1g0 \\ 1g1 \\ g00 \\ g01 \\ g10 \\ g11 \end{matrix} \begin{pmatrix} 1gg \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

We have $A_{3,2}^\top x_{1gg}^{3,2,1} = 2z_{1gg}^{3,1}$ and $A_{3,2} z_{1gg}^{3,1} = 2x_{1gg}^{3,2,1}$. Thus $A_{3,2} A_{3,2}^\top x_{1gg}^{3,2,1} = 4x_{1gg}^{3,2,1}$.

- iii. Let $n = 2$. The word v' can be one of $11g$, $1g1$ or $g11$. Consider $v' = 11g$ for instance. The column vectors $z_{v'}^{\ell n}$ and $x_{v'}^{\ell kn}$ are as

$$z_{11g}^{3,2} = \begin{matrix} 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{matrix} \begin{pmatrix} 11g \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ 1 \\ 1 \end{pmatrix} \quad x_{11g}^{3,2,2} = \begin{matrix} 00g \\ 01g \\ 0g0 \\ 0g1 \\ 10g \\ 11g \\ 1g0 \\ 1g1 \\ g00 \\ g01 \\ g10 \\ g11 \end{matrix} \begin{pmatrix} 11g \\ 1 \\ -1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

It is easily verified that $A_{3,2}^\top x_{11g}^{3,2,2} = z_{11g}^{3,2}$ and $A_{3,2} z_{11g}^{3,2} = 2x_{11g}^{3,2,2}$. Thus $A_{3,2} A_{3,2}^\top x_{11g}^{3,2,2} = 2x_{11g}^{3,2,2}$.

The complete matrix of eigenvectors of $AA^\top$ (corresponding to nonzero eigenvalues) is given by

$$\Delta_{3,2} = \begin{pmatrix} \text{ggg} & \text{gg1} & \text{g1g} & \text{1gg} & \text{g11} & \text{1g1} & \text{11g} \\ 00\text{g} & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 01\text{g} & 1 & 0 & -1 & 1 & 0 & 0 & -1 \\ 0\text{g}0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0\text{g}1 & 1 & -1 & 0 & 1 & 0 & -1 & 0 \\ 10\text{g} & 1 & 0 & 1 & -1 & 0 & 0 & -1 \\ 11\text{g} & 1 & 0 & -1 & -1 & 0 & 0 & 1 \\ 1\text{g}0 & 1 & 1 & 0 & -1 & 0 & -1 & 0 \\ 1\text{g}1 & 1 & -1 & 0 & -1 & 0 & 1 & 0 \\ \text{g}00 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ \text{g}01 & 1 & -1 & 1 & 0 & -1 & 0 & 0 \\ \text{g}10 & 1 & 1 & -1 & 0 & -1 & 0 & 0 \\ \text{g}11 & 1 & -1 & -1 & 0 & 1 & 0 & 0 \end{pmatrix}$$

and $Q_{3,2} = \frac{1}{\sqrt{12}}, \frac{1}{\sqrt{8}}, \frac{1}{\sqrt{8}}, \frac{1}{\sqrt{8}}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}$.

The diagonal matrix Λ consisting of the nonzero eigenvalues is $\Lambda = \text{diag}(6, 4, 4, 4, 2, 2, 2)$.

So we have the following eigendecomposition: $AA^T = Q\Lambda Q^T$.

For the pseudo-inverse matrix W , we have $W = A^T Q \Lambda^{-1} Q^T = A^T D^{-1}$. Where

$D = E\Lambda^{-1}E^T = \text{diag}(\frac{1}{72}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$ and W is

$$W_{\ell=3,k=2} = \frac{1}{24} \begin{pmatrix} 00\text{g} & 01\text{g} & 0\text{g}0 & 0\text{g}1 & 10\text{g} & 11\text{g} & 1\text{g}0 & 1\text{g}1 & \text{g}00 & \text{g}01 & \text{g}10 & \text{g}11 \\ 000 & 7 & -2 & 7 & -2 & -2 & 1 & -2 & 1 & 7 & -2 & -2 & 1 \\ 001 & 7 & -2 & -2 & 7 & -2 & 1 & 1 & -2 & -2 & 7 & 1 & -2 \\ 010 & -2 & 7 & 7 & -2 & 1 & -2 & -2 & 1 & -2 & 1 & 7 & -2 \\ 011 & -2 & 7 & -2 & 7 & 1 & -2 & 1 & -2 & 1 & -2 & -2 & 7 \\ 100 & -2 & 1 & -2 & 1 & 7 & -2 & 7 & -2 & 7 & -2 & -2 & 1 \\ 101 & -2 & 1 & 1 & -2 & 7 & -2 & -2 & 7 & -2 & 7 & 1 & -2 \\ 110 & 1 & -2 & -2 & 1 & -2 & 7 & 7 & -2 & -2 & 1 & 7 & -2 \\ 111 & 1 & -2 & 1 & -2 & -2 & 7 & -2 & 7 & 1 & -2 & -2 & 7 \end{pmatrix}$$

It can be observed that the element w_{ij} corresponding to the ℓ -mer u_i and the gapped k -mer v_j only depends on the number of mismatches between u_i and v_j and is given by Eq. (14).

8 Biological example

We now present a simple application to real biological sequences, applying our method to estimate k -mer frequencies in human sequences determined to be bound by the well studied transcription factor CTCF in state-of-the-art Chip-seq experiments. In many cases, transcription factor binding is determined by local sequence features flanking the TFs direct binding sequence. Hence even for Chip-seq experiments on a TF which directly binds DNA, it is often necessary to describe the local sequence features in the surrounding sequence, which can modulate the TF's cell-specific occupancy. In other cases, sets of functional DNA sequences, such as enhancers, can be defined by combinations of cooperating sets of binding

factors. In these situations, the optimal description (length and number of gaps) can vary considerably among the various factors involved. Therefore, we have had most success with modeling binding sites in enhancers with the full set of k -mers. However, for longer binding sites, the number of functional sequences can be too small for reliable k -mer frequency estimation.

CTCF is somewhat unique in that it has a very large number of binding sites in the genome as a consequence of its ubiquitous biological role as an insulator. The Chip-seq data for CTCF binding in the human genome consists of 32055 binding regions, each 300bp long. Boyle et al. (2011) Within these regions, the direct CTCF binding site is fairly long, with 13–16 informative bases. CTCF is also fairly unique in that it seems to bind the same regions with little combinatorial modulation from neighboring accessory factor binding sites. Thus for CTCF, we have a large set of functional regions and we can estimate the k -mer counts up to $k=10$ fairly robustly. In contrast, in most other cases the number of sequences of a specific functional class is much less than 32,055, and is typically several hundred or a few thousand. This is the case for cell-type specific enhancers, where Chip-seq experiments usually find smaller sets of sequences. For example, in previous work, we have used k -mer sequence features to predict enhancers bound by P300 in 500–2,500 sequences in mouse forebrain, midbrain, and limb. Lee et al. (2011) This approach suffers from overfitting when the binding sites involved get longer than 6–8 bp.

Using the CTCF data we can ask how closely the k -mer frequency distribution in a subset of the actual biological sequences matches the much more robustly known k -mer frequency distribution in the full set of 32,055 CTCF bound sequences. We can measure the difference between the estimated frequency distribution and the known distribution by using the root

mean squared error between the two distributions, $\text{RMSE}(p) = \sqrt{\sum_i (p_{\text{est},i} - p_{\text{known},i})^2}$. We find $p_{\text{known},i}$ from the full set of 32,055 sequences excluding the smaller test set of sequences used in the estimation. In Fig. 1, we show the $\text{RMSE}(p)$ from 10 independent random samples ranging in size from 500 to 2,500 sequences, using our minimum norm ℓ -mer estimate using $\ell = 10$ and $k = 6$ (*black*), and we compare it to the directly observed ℓ -mer distribution using ungapped 10-mers on the same subsets of sequences (*red*). In blue is shown a single sample varying continuously in size by single binding regions. For larger samples of the data, the error of these two methods becomes comparable, but for smaller sets (500 or 1,000 sequences), the ℓ -mer frequency estimate using gapped k -mers is much closer to the actual distribution in the 32,055 sequences (63% higher error using ungapped ℓ -mers for $N = 500$). This demonstrates the utility of our assumption that matching the gapped k -mer distribution with the minimum norm ungapped ℓ -mer distribution more robustly estimates frequencies in actual biological sequence samples.

9 Conclusions

We have derived a simple form for $W_{\ell k}$, the matrix which maps gapped k -mer counts to ℓ -mer count estimates. We have shown that this matrix is the Moore–Penrose pseudo-inverse of $A_{\ell k}$ and yields the minimum norm estimate for $\mathbf{x}$. This approach can be used to find robust estimates of ℓ -mer frequencies from limited training data, which should reduce overfitting in

statistical model learning from genomic sequence data. We have shown that this method more reliably estimates the frequency distribution of k -mers in a sample of CTCF binding sites from the human genome than the actual k -mer frequencies observed in the small sample.

One caveat which should be mentioned is that our minimum norm estimate of $W_{\ell k}$ has the property that it does not constrain count estimates of $\mathbf{x}$ in Eq. 13 to be positive. In practice we can circumvent this difficulty by thresholding the filter that efficiently generates the estimated counts from the training sequences, but this was intentionally not done in the example in Sect. 8, to make the evaluation of our method as conservative as possible. In addition to this detail, we note that matrix W , could be computed numerically in $O(MN)$ to find the estimated frequencies. However, having an analytical solution for W , we can develop fast algorithms that are applicable for any large ℓ and k to calculate the count estimates. From the analytical solution for W , we can find an analytical solution for the matrix $G = WA$ and use that to directly calculate the count estimates from the training set without the need to calculate the gapped kmer counts. This eliminates the need for numerical computation of W . Also, by having an analytical solution for W , we can expand the k -mer SVM algorithm Lee et al. (2011) to compute distances in the ℓ -mer count estimate space. Efficient calculation of the distance between two sequences in the larger ℓ -mer space requires some further mathematical development. In a future work, we will compare the relative advantages of directly using gapped k -mers as features instead of using estimated ungapped ℓ -mer frequencies. Although the gapped k -mer distribution is used as a fundamental feature in generating the minimum norm solution $\hat{\mathbf{x}}$, the ungapped ℓ -mer distribution may have a correlation structure that is more favorable than the gapped k -mer distribution for some machine learning applications. In the forthcoming application paper we will develop a fast algorithm to compute the ℓ -mer count estimates and will explore the sensitivity of sequence classification on the parameters ℓ and k , and show how using gapped k -mers dramatically improves the accuracy our SVM predictions of cell specific mammalian enhancers.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the reviewers for their comments and suggestions which significantly improved the manuscript. We also thank users of <http://math.stackexchange.com> online community, specifically users Joriki and Siva for their useful comments which helped us in the development of the proof. Dongwon Lee graciously provided the processed CTCF sequence data. The research of M.M. was in part supported by a grant from IPM (No. CS1390-4-07), and M.B. was supported by the Searle Scholars Program and in part by NIH grant NS062972.

Appendix: Proofs of some of the Propositions

Proof of Proposition 1 The first identity is proved as follows

$$\begin{aligned}
& \binom{k}{p} \binom{p}{t} \binom{k-p}{n-t} \\
&= \binom{k}{t} \binom{k-t}{p-t} \binom{k-p}{n-t} \\
&= \binom{k}{t} \binom{k-t}{k-p} \binom{k-p}{k-n-p+t} = \binom{k}{t} \binom{k-t}{k-n} \binom{k-n}{p-t} \\
&= \binom{k}{t} \binom{k-t}{n-t} \binom{k-n}{p-t} \\
&= \binom{k}{n} \binom{n}{t} \binom{k-n}{p-t}.
\end{aligned}$$

To prove (16), denote the left side by $\tau_{p,k,\ell}$, then using

$$\binom{k-n}{p-t} = \binom{k-n-1}{p-t-1} + \binom{k-n-1}{p-t}, \text{ we obtain}$$

$$\begin{aligned}
& \tau_{p,k,\ell} = \tau_{p-1,k-1,\ell} \\
& + \sum_{t=0}^k (-1)^{k-t} \binom{\ell}{k} \binom{k}{t} \binom{-1}{p-t-1} x^t \\
& + \tau_{p,k-1,\ell} + \sum_{t=0}^k (-1)^{k-t} \binom{\ell}{k} \binom{k}{t} \binom{-1}{p-t} x^t \\
& = \tau_{p-1,k-1,\ell} \\
& + \tau_{p,k-1,\ell} + \sum_{t=0}^k (-1)^{k-t} \binom{\ell}{k} \binom{k}{t} \left(\binom{-1}{p-t} + \binom{-1}{p-t-1} \right) x^t \\
& = \tau_{p-1,k-1,\ell} \\
& + \tau_{p,k-1,\ell} + \sum_{t=0}^k (-1)^{k-t} \binom{\ell}{k} \binom{k}{t} \binom{0}{p-t} x^t \\
& = \tau_{p-1,k-1,\ell} \\
& + \tau_{p,k-1,\ell} + (-1)^{k-p} \binom{\ell}{k} \binom{k}{p} x^p.
\end{aligned}$$

Then by replacing k by k' in $\tau_{p,k,\ell} - \tau_{p,k-1,\ell} = (-1)^{k-p} \binom{\ell}{p} \binom{\ell-p}{k-p} x^p + \tau_{p-1,k-1,\ell}$ and summing up over $k', 0 \leq k' \leq k$, we get

$$\tau_{p,k,\ell} = \sum_{k'=0}^k (-1)^{k'-p} \binom{\ell}{p} \binom{\ell-p}{k'-p} x^p + \sum_{k'=0}^k \tau_{p-1,k'-1,\ell},$$

whence

$$\tau_{p,k,\ell} = \binom{\ell}{p} \binom{k-\ell}{k-p} x^p + \sum_{k'=0}^k \tau_{p-1,k'-1,\ell}. \quad (34)$$

Now we prove the identity $\tau_{p,k,\ell} = \binom{k-\ell}{k-p} \sum_{n=0}^p \binom{\ell}{n} x^n$ for $p = 0, 1, \dots, k$ by

bounded induction on p . For $p = 0$, using (34) we obtain $\tau_{0,k,\ell} = \binom{k-\ell}{k} x^0$ which proves the required identity in this case. Now let $0 < p < k$ and suppose that the result is true for $p-1$. We prove the validity of the identity for p by using the induction hypothesis and (34), as follows:

$$\begin{aligned} \tau_{p,k,\ell} &= \binom{k-\ell}{k-p} \binom{\ell}{p} x^p \\ &\quad + \sum_{k'=0}^k \tau_{p-1,k'-1,\ell} \\ &= \binom{k-\ell}{k-p} \binom{\ell}{p} x^p \\ &\quad + \sum_{k'=0}^k \binom{k'-\ell-1}{k'-p} \sum_{n=0}^{p-1} \binom{\ell}{n} x^n = \binom{k-\ell}{k-p} \binom{\ell}{p} x^p \\ &\quad + \binom{k-\ell}{k-p} \sum_{n=0}^{p-1} \binom{\ell}{n} x^n \\ &= \binom{k-\ell}{k-p} \sum_{n=0}^p \binom{\ell}{n} x^n. \end{aligned}$$

Proof of Proposition 2 (i) It is enough to prove that the identity

$$\sum_{y \in M_{\ell k}(u)} \nu(y, v') = \binom{\ell-n}{\ell-k} \nu(u, v') \quad (35)$$

holds for any $u \in U_\ell$. The nonzero summands of the summation on the left are obtained from the words $y \in V_{\ell k}$ in which all symbols g appear in positions Gv' (Using $|Gv'| = \ell - n$,

we conclude that there are $\binom{\ell-n}{\ell-k}$ such summands). On the other hand, since $y \in M_{\ell k}(u)$, it is easily seen that for any such y we have $\nu(y, v') = \nu(u, v')$. Thus the summation is

simplified to $\binom{\ell-n}{\ell-k} \nu(u, v')$ as required.

(ii) It is enough to prove that the identity

$$\sum_{u \in M'_{\ell k}(v)} \nu(u, v') = b^{\ell-k} \nu(v, v') \quad (36)$$

holds for any $v \in V_{\ell k}$. We prove this in two cases:

Case (a) Suppose that $Gv \subseteq Gv'$. Then for any $u \in M'_{\ell k}(v)$ we have

$$\nu(u, v') = \prod_{i \in \tilde{G}_{v'}} \nu(u_i, v'_i) = \prod_{i \in \tilde{G}_{v'}} \nu(v_i, v'_i) = \nu(v, v')$$

and since there are $b^{\ell-k}$ such words u , the equation (36) thus follows.

Case (b) Suppose that $Gv \not\subseteq Gv'$, consequently $Gv \setminus Gv' \neq \emptyset$. Now for any $i \in Gv \setminus Gv'$ we have $\nu(v_i, v'_i) = 0$, thus the right side of (36) is 0. We prove that the left side is also 0 as follows. The nonzero summands in the summation are obtained from elements $u \in X$ where the subset $X \in U_\ell$ is given by

$$X = \{u \in U_\ell : u_i \leq v_i \text{ for } i \in Gv \setminus Gv' \text{ and } u_i = v_i \text{ for } i \in \tilde{G}_v\}.$$

Thus we obtain

$$\begin{aligned} & \sum_{u \in M'_{\ell k}(v)} \nu(u, v') \\ &= \sum_{u \in X} \nu(u, v') \\ &= \sum_{u \in X} \prod_{i=1}^{\ell} \nu(u_i, v'_i) \\ &= \sum_{u \in X} \left(\prod_{i \in \tilde{G}_v} \nu(u_i, v'_i) \prod_{i \in Gv \setminus Gv'} \nu(u_i, v'_i) \prod_{i \in Gv \cap Gv'} \nu(u_i, v'_i) \right) \\ &= \sum_{u \in X} \left(\prod_{i \in \tilde{G}_v} \nu(v_i, v'_i) \prod_{i \in Gv \setminus Gv'} \nu(u_i, v'_i) \prod_{i \in Gv \cap Gv'} \nu(u_i, g) \right) \\ &= \prod_{i \in \tilde{G}_v} \nu(v_i, v'_i) \prod_{i \in Gv \setminus Gv'} \sum_{u_i=0}^{v'_i} \nu(u_i, v'_i) \prod_{i \in Gv \cap Gv'} \sum_{u_i=0}^{b-1} \nu(u_i, g) \\ &= b^{|Gv \cap Gv'|} \prod_{i \in \tilde{G}_v} \nu(v_i, v'_i) \prod_{i \in Gv \setminus Gv'} \sum_{u_i=0}^{v'_i} \nu(u_i, v'_i) = 0. \end{aligned}$$

The last identity holds because for any $i \in Gv \setminus Gv'$ we have

$$\sum_{u_i=0}^{v'_i} \nu(u_i, v'_i) = \sum_{u_i=0}^{v'_i-1} 1 - v'_i = 0.$$

(iii), (iv) These are immediate consequences of parts (i) and (ii).

Proof of Proposition 3 (i) If $i \in P$ and $v_i > 0$ then

$$\phi_i(u, v) = \sum_{j=v_i}^{b-1} \frac{\nu(v_i, j)\nu(u_i, j)}{j(j+1)} = \sum_{j=v_i}^{b-1} \frac{\nu(v_i, j)^2}{j(j+1)} = \frac{v_i^2}{v_i(v_i+1)} + \sum_{j=v_i+1}^{b-1} \frac{1}{j(j+1)} = \frac{v_i}{v_i+1} + \left(\frac{1}{v_i+1} - \frac{1}{b} \right) = \frac{b-1}{b}.$$

If $i \in P$ and $v_i = 0$, then given $j \geq 1$, we have $\nu(u_i, j) = \nu(v_i, j) = 1$, hence,

$$\phi_i(u, v) = \sum_{j=1}^{b-1} \frac{1}{j(j+1)} = \frac{b-1}{b}.$$

The case $i \in Q$ is done similarly.

(ii) Without loss of generality suppose that $v = v_1 \cdots v_k g^{\ell-k}$. Let $P = \{1, \dots, p\}$ and $Q = \{p +$

$1, \dots, k\}$ and for a subset $\bar{G} \in \binom{\{1, \dots, \ell\}}{\ell-n}$ let $X'_{\ell n}(\bar{G}) = \{v' \in V'_{\ell n} : \bar{G}_{v'} = \bar{G}\}$. Denote the left side of (17) by S . Since the nonzero summands in (17) are obtained from the words

$v' \in V'_{\ell n}$ which have g 's at the $\ell - k$ rightmost positions, we may just consider words

$v' = v'_1 \cdots v'_k g^{\ell-k}$ with $|v'_1 \cdots v'_k|_g = k - n$. Then we have

$$\begin{aligned}
S &= \sum_{t=0}^n \sum_{T_1 \in \binom{P}{t}} \sum_{T_2 \in \binom{Q}{n-t}} \sum_{v' \in X'_{\ell_n}(T_1 \cup T_2)} \frac{\nu(v, v') \nu(u, v')}{\prod_{i \in \bar{G}_{v'}} v'_i(v'_i + 1)} \\
&= \sum_{t=0}^n \sum_{T_1 \in \binom{P}{t}} \sum_{T_2 \in \binom{Q}{n-t}} \sum_{v' \in X'_{\ell_n}(T_1 \cup T_2)} \frac{\prod_{i \in \bar{G}_{v'}} \nu(v_i, v'_i) \nu(u_i, v'_i)}{\prod_{i \in \bar{G}_{v'}} v'_i(v'_i + 1)} \\
&= \sum_{t=0}^n \sum_{T_1 \in \binom{P}{t}} \sum_{T_2 \in \binom{Q}{n-t}} \sum_{v' \in X'_{\ell_n}(T_1 \cup T_2)} \prod_{i \in T_1 \cup T_2} \frac{\nu(v_i, v'_i) \nu(u_i, v'_i)}{v'_i(v'_i + 1)} \\
&= \sum_{t=0}^n \sum_{T_1 \in \binom{P}{t}} \sum_{T_2 \in \binom{Q}{n-t}} \prod_{i \in T_1 \cup T_2} \left(\sum_{v'_i = \max\{1, v_i\}}^{b-1} \frac{\nu(v_i, v'_i) \nu(u_i, v'_i)}{v'_i(v'_i + 1)} \right) \\
&= \sum_{t=0}^n \sum_{T_1 \in \binom{P}{t}} \sum_{T_2 \in \binom{Q}{n-t}} \prod_{i \in T_1 \cup T_2} \phi_i(u, v) \\
&= \sum_{t=0}^n \sum_{T_1 \in \binom{P}{t}} \sum_{T_2 \in \binom{Q}{n-t}} \left(\frac{b-1}{b} \right)^t \left(\frac{-1}{b} \right)^{n-t} \text{ (by part(i))} \\
&= \frac{1}{b^n} \sum_{t=0}^n (-1)^{n-t} \binom{p}{t} \binom{k-p}{n-t} (b-1)^t.
\end{aligned}$$

Proof of Proposition 5 (i) These are immediate consequences of Definition 3, except the last equation. To calculate the norm of $x_{v'}$, without loss of generality let $G_{v'} = \{1, \dots, n\}$, hence $v' = v'_1 \dots v'_n g^{\ell-n}$ and it is evident that the nonzero elements $v(w, v')$ of $x_{v'}$ are obtained from the words $w = rs \in V_{\ell k}$ with $|s| = \ell - n$, $|s|_g = \ell - k$, $r = r_1 \dots r_n$ and $0 \leq r_i \leq v'_i$ for $i = 1, \dots, n$.

Moreover, the value $v(w, v')$ is independent of s and there are $\binom{\ell-n}{\ell-k} b^{k-n}$ choices for s . We thus provide

$$\|x_{v'}\|^2 = \binom{\ell-n}{\ell-k} b^{k-n} \prod_{i=1}^n \sum_{x_i=0}^{v'_i} \nu(w_i, v'_i)^2 = \binom{\ell-n}{\ell-k} b^{k-n} \prod_{i=1}^n (v'_i + v_i'^2),$$

as required.

(ii) The first equation is proved easily considering Remark 2. To prove the second one, observe that

$$Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top = \Delta_{\ell k} E \Lambda^{-1} E^\top \Delta_{\ell k}^\top,$$

in which the matrix $D = E \Lambda^{-1} E^\top$ is a diagonal matrix of the form $D = \text{diag}((d_{v'})_{v' \in V'_{\ell, \leq k}})$

with entries $d_{v'} = \frac{1}{\lambda_{|\bar{G}_{v'}|} \|x_{v'}\|^2}$ or

$$d_{v'} = \frac{1}{\left(\begin{matrix} \ell - |\bar{G}_{v'}| \\ \ell - k \end{matrix} \right)^2 b^{\ell - |\bar{G}_{v'}|} \prod_{i \in \bar{G}_{v'}} (v'_i + v_i'^2)}.$$

(iii) Since the columns of $Q_{\ell k}$ are normalized orthogonal eigenvectors, the first identity holds. To prove the second one, using the notation of Remark 2, we begin by claiming

$$A_{\ell k}^\top N_{\ell k} = 0.$$

In fact if $y \in \ker(A_{\ell k} A_{\ell k}^\top)$ then from $A_{\ell k} A_{\ell k}^\top y = 0$ we obtain $y^\top A_{\ell k} A_{\ell k}^\top y = 0$ and $\|A_{\ell k}^\top y\| = 0$, thus $A_{\ell k}^\top y = 0$, which proves the claim. Now, by using the decomposition $P_{\ell k} = [Q_{\ell k} R_{\ell k}]$ and the equation $P_{\ell k} P_{\ell k}^\top = I$, we obtain $Q_{\ell k} Q_{\ell k}^\top + N_{\ell k} N_{\ell k}^\top = I$, hence $A_{\ell k}^\top Q_{\ell k} Q_{\ell k}^\top = A_{\ell k}^\top - A_{\ell k}^\top N_{\ell k} N_{\ell k}^\top$ which yields $A_{\ell k}^\top Q_{\ell k} Q_{\ell k}^\top = A_{\ell k}^\top$ using the above claim. Transposing both sides yields the result.

Proof of Proposition 6 First note that

$$\begin{aligned} W_{\ell k} A_{\ell k} W_{\ell k} &= (A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top) A_{\ell k} (A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top) \\ &= A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top (A_{\ell k} A_{\ell k}^\top) Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top = A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top (Q_{\ell k} \Lambda Q_{\ell k}^\top) Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top \\ &= A_{\ell k}^\top Q_{\ell k} (\Lambda^{-1} Q_{\ell k}^\top Q_{\ell k} \Lambda) (Q_{\ell k}^\top Q_{\ell k}) \Lambda^{-1} Q_{\ell k}^\top \\ &= A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top = W_{\ell k}. \end{aligned}$$

Second, we have

$$\begin{aligned} A_{\ell k} W_{\ell k} A_{\ell k} &= A_{\ell k} (A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top) A_{\ell k} \\ &= (A_{\ell k} A_{\ell k}^\top) Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top A_{\ell k} \\ &= Q_{\ell k} (\Lambda Q_{\ell k}^\top Q_{\ell k} \Lambda^{-1}) Q_{\ell k}^\top A_{\ell k} \\ &= Q_{\ell k} Q_{\ell k}^\top A_{\ell k} \\ &= A_{\ell k} \text{ (by Proposition 5(iii))}. \end{aligned}$$

Finally, observe that $W_{\ell k} A_{\ell k}$ and $A_{\ell k} W_{\ell k}$ are real symmetric matrices. It is concluded that $W_{\ell k}$ is the Moore–Penrose pseudo-inverse of $A_{\ell k}$.

To prove the last statement, we show that the sum of the elements in each row of matrix $W_{\ell k} A_{\ell k}$ is equal to 1. The same result for columns is concluded from the symmetry of $W_{\ell k} A_{\ell k}$.

Let $J = [1, 1, \dots, 1]^\top$ and $F = [1, 0, \dots, 0]^\top$. Note that $J = z_{gg\dots g}^{\ell 0}$ and $F = x_{gg\dots g}^{\ell k 0}$ and note that

the row sums of $W_{\ell k} A_{\ell k}$ are the entries of $W_{\ell k} A_{\ell k} J$. Given $M = \binom{\ell}{k} b^k$ the number of all possible gapped k -mers, we have

$$\begin{aligned}
 W_{\ell k} A_{\ell k} J &= A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top A_{\ell k} z_{gg\dots g}^{\ell 0} \\
 &= A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top b^{\ell-k} x_{gg\dots g}^{\ell k 0} \text{ (by Proposition 2(ii))}, \\
 &= A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} \frac{M}{\sqrt{M}} b^{\ell-k} F \text{ (using } Q_{\ell k}^\top x_{gg\dots g}^{\ell k 0} \\
 &= \sqrt{M} [1, 0, 0, \dots]^\top = A_{\ell k}^\top Q_{\ell k} \frac{1}{\binom{\ell}{k} b^{\ell-k}} \frac{M}{\sqrt{M}} b^{\ell-k} F = A_{\ell k}^\top \frac{1}{\sqrt{M}} \frac{1}{\binom{\ell}{k} b^{\ell-k}} \frac{M}{\sqrt{M}} b^{\ell-k} x_{gg\dots g}^{\ell k 0} = \binom{\ell}{k} \frac{1}{\sqrt{M}} \frac{1}{\binom{\ell}{k} b^{\ell-k}} \frac{M}{\sqrt{M}} \\
 &= z_{gg\dots g}^{\ell 0} = [1, 1, \dots, 1]^\top.
 \end{aligned}$$

References

- Albert, AE. Regression and the Moore-penrose Pseudoinverse. New York: Academic Press; 1972.
- Beer MA, Tavazoie S. Predicting gene expression from sequence. *Cell*. 2004; 117:185–198. [PubMed: 15084257]
- Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support vector machines and kernels for computational biology. *PLoS Comput Biol*. 2008; 4:e1000173. [PubMed: 18974822]
- Boyle AP, Song L, Lee B-K, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res*. 2011; 21:456–464. [PubMed: 21106903]
- Cameron PJ. Notes on Counting. 2003 <http://www.maths.qmul.ac.uk/pjc/notes/counting.pdf>.
- Elemento O, Tavazoie S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol*. 2005; 6:R18. [PubMed: 15693947]
- Göke J, Schulz MH, Lasserre J, Vingron M. Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts. *Bioinformatics (Oxford, England)*. 2012
- Graham, RL.; Knuth, DE.; Patashnik, O. Concrete mathematics: a foundation for computer science. 2nd edn.. Boston: Addison Wesley Publishing Company; 1994.
- van Helden J. Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics*. 2004; 20:399–406. [PubMed: 14764560]
- Kantorovitz MR, Kazemian M, Kinston S, Miranda-Saavedra D, Zhu Q, Robinson GE, Göttgens B, Halfon MS, Sinha S. Motif-blind, genome-wide discovery of cis-regulatory modules in *Drosophila* and mouse. *Dev Cell*. 2009; 17:568–579. [PubMed: 19853570]
- Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res*. 2011; 21:2167–2180. [PubMed: 21875935]
- Leslie CS, Eskin E, Cohen A, Weston J, Noble WS. Mismatch string kernels for discriminative protein classification. *Bioinformatics*. 2004; 20:467–476. [PubMed: 14990442]

- Meinicke P, Tech M, Morgenstern B, Merkl R. Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites. *BMC Bioinformatics*. 2004; 5:169. [PubMed: 15511290]
- Sonnenburg S, Schweikert G, Philips P, Behr J, Rätsch G. Accurate splice site prediction using support vector machines. *BMC Bioinformatics*. 2007; 8:S7. [PubMed: 18269701]
- Sonnenburg S, Zien A, Rätsch G. ARTS: accurate recognition of transcription starts in human. *Bioinformatics*. 2006; 22:e472–e480. [PubMed: 16873509]
- Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics*. 2000; 16:16–23. [PubMed: 10812473]
- Wilson RM. A diagonal form for the incidence matrices of t -subsets vs. k -subsets. *Eur J Combin*. 1990; 11:609–615.
- Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*. 2005; 434:338–345. [PubMed: 15735639]

Fig. 1.

Minimum norm k -mer distribution has lower $\text{RMSE}(p)$ error than the actual observed k -mer distribution in small samples of CTCF binding sites. We show the root mean squared error, $\text{RMSE}(p)$, between two distributions, the observed k -mer distribution (*red*) and the minimum norm estimate (*black*) using $\ell = 10$ and $k = 6$ on a subset of N sequences, compared to the known distribution from the full set of sequences. In *blue* is shown a sample varying in size

by single elements. For smaller subsets typical of enhancer training sets, the k -mer frequency estimate using gapped k -mers is much closer to the actual distribution