Density estimation for ordinal biological sequences and its applications

Wei-Chia Chen,[1,*] Juannan Zhou,[2] and David M. McCandlish[3]
[1]*Department of Physics, National Chung Cheng University, Chiayi 62102, Taiwan, Republic of China*
[2]*Department of Biology, University of Florida, Gainesville, Florida 32611, USA*
[3]*Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA*

(Received 17 April 2024; revised 5 August 2024; accepted 3 October 2024; published 30 October 2024)

Biological sequences do not come at random. Instead, they appear with particular frequencies that reflect properties of the associated system or phenomenon. Knowing how biological sequences are distributed in sequence space is thus a natural first step toward understanding the underlying mechanisms. Here we propose a method for inferring the probability distribution from which a sample of biological sequences were drawn for the case where the sequences are composed of elements that admit a natural ordering. Our method is based on Bayesian field theory, a physics-based machine learning approach, and can be regarded as a nonparametric extension of the traditional maximum entropy estimate. As an example, we use it to analyze the aneuploidy data pertaining to gliomas from The Cancer Genome Atlas project. In addition, we demonstrate two follow-up analyses that can be performed with the resulting probability distribution. One of them is to investigate the associations among the sequence sites. This provides a way to infer the governing biological grammar. The other is to study the global geometry of the probability landscape, which allows us to look at the problem from an evolutionary point of view. It can be seen that this methodology enables us to learn from a sample of sequences about how a biological system or phenomenon in the real world works.

DOI: 10.1103/PhysRevE.110.044408

I. INTRODUCTION

The technology of DNA sequencing has evolved immensely since its invention by Sanger in 1977 [1]. Particularly, in the past 20 years we have seen the technology transforming from the first-generation electrophoretic sequencing into the second-generation massively parallel sequencing and the third-generation real-time, single-molecule sequencing. For a review, see Ref. [2]. Accompanying those advances are applications of DNA sequencing in new ways or to new areas, such as ChIP-seq [3] and RNA-seq [4–8]. Needless to say, a natural product of these sequencing experiments is a wealth of biological sequences.

It is now a well-known fact that biological sequences do not come at random. For instance, DNAs in a segment of genome are usually arranged in some order, encoding a piece of genetic information. Thus, given a sample of biological sequences, a fundamental question is *what the underlying probability distribution from which the sequences were drawn looks like* [9–13]. In a recent work [14] we developed a novel density estimation method for biological sequences based on Bayesian field theory [15–21]. This method, called SeqDEFT (standing for Sequence Density Estimation using Field Theory), was aimed for *categorical* sequences, that is, sequences made up of elements that are distinct from each other, but do not have a natural ordering. Typical examples of categorical sequence data include DNA sequences (consisting of the four nucleotides) and protein sequences (consisting of the 20 amino acids). The SeqDEFT method can be regarded as a nonparametric extension of the position weight matrix model [22,23] or the Potts model [24–33], depending on the chosen value of a hyperparameter. We showed that SeqDEFT was able to provide a more detailed view of the probability landscape than these commonly used models in two different biological contexts.

In this work we propose a method, again within the framework of Bayesian field theory, for the density estimation on a different type of sequence data, whose elements do have a natural ordering. Such *ordinal* sequence data are not uncommon in biology. An example is the karyotype of a cell or an organism, which contains the copy number of each chromosome thereof. (In biology, a karyotype includes the sizes, numbers, and shapes of the chromosomes. Here we are adopting a simplified version of karyotype and look only at the copy numbers.) Due to the advancement of DNA sequencing technology, today a karyotype can even record the copy number of each gene in the whole genome [2]. The ability to obtain karyotypes either in the chromosome level or in the gene level has a huge impact on medicine, as it has been shown that patients having certain diseases tend to possess some particular patterns of chromosome or gene copy number variations (e.g., [34]). Such molecular information is playing a more and more important role in clinical studies. For an overview, see Ref. [35].

For instance, in the most recent WHO (World Health Organization) classification, diffuse gliomas, which is the most common malignant tumor of the central nerve system in adults, have been classified into three subtypes, oligodendroglioma, astrocytoma, and glioblastoma, according to

*Contact author: wcchen@ccu.edu.tw

their molecular features as well as traditional histology [36]. Both oligodendroglioma (WHO Grade 2, 3) and astrocytoma (WHO Grade 2, 3, 4) have the IDH gene mutant, but oligodendroglioma has the additional codeletion of chromosome arms 1p and 19q, whereas astrocytoma does not. On the other hand, glioblastoma (WHO Grade 4) does not have the IDH gene mutant, but many patients have gain in chromosome 7 and loss in chromosome 10, as well as other gene mutations. The five-year survival rate of low-grade glioma is about 80%, and for high-grade glioma it is below 5% [36]. This refined classification, based on both molecular features and traditional histology, not only can better predict disease progression but may also suggest novel treatments for the disease [36].

The density estimation method proposed here shares many properties with its categorical analog (i.e., SeqDEFT). In particular, it includes the position weight matrix model and the Potts model as limiting cases. Also, being a nonparametric method, it can be expected to provide a better estimate of how ordinal sequences are distributed in sequence space than those traditional methods. Moreover, we propose two follow-up analyses that can be conducted with a probability distribution to extract more information from data. First, we show that by examining the associations among the sequence sites, we can find out which patterns dominate in a certain background. These "rules" are analogous to the language grammar that one has to obey in order to make valid words, and thus can be regarded as the biological grammar governing the system or phenomenon. Second, we can visualize a probability landscape using a dimensionality reduction technique based on an evolutionary model. Doing so allows us to imagine what a population is likely going to experience as it wanders around the sequence space.

This article is organized as follows. In Sec. II we develop the density estimation method and point out the difference between the method and its categorical analog. Then we perform density estimation, along with the two follow-up analyses, on the aneuploidy data pertaining to gliomas collected by The Cancer Genome Atlas (TCGA) project [37] in Sec. III. Finally, we conclude with a summary and discussion in Sec. IV.

II. FORMALISM

In this section we develop our method for density estimation on ordinal sequence data within the framework of Bayesian field theory. We call this method ordinal SeqDEFT. Although some of the results have been presented in the literature, especially [14] and [20], we include all the essential material to make the content of this section self-contained.

Our goal here is to estimate a probability distribution with a sample of ordinal biological sequences. The strategy is based on maximum likelihood estimation, subject to an appropriate smoothness constraint. The purpose of this constraint is to prevent the distribution, which is going to adopt a nonparametric representation, from overfitting. Thus, the resulting distribution will be one that reaches a balance between the data and the smoothness measure. As it turns out, how strong the smoothness constraint should be can also be determined by the data. Below we will look into the technicalities of how to implement this strategy.

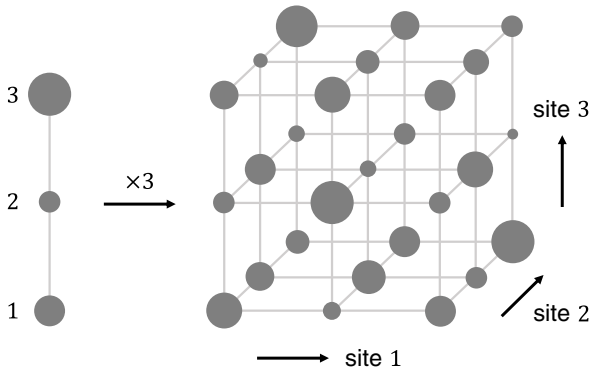

FIG. 1. Representation of an $\alpha^\ell = 3^3$ ordinal sequence space. In this example, each sequence consists of three sites, and each site can be in one of three states, say, 1, 2, 3. The order of the states is represented by a path graph of length 3 (left). The sequence space, in turn, is represented as a grid graph obtained by a threefold Cartesian product of the path graph (right). Each direction of the grid graph corresponds to a particular site. This graph has 27 nodes in total, each node represents a sequence, and the size of the node is proportional to the probability of the sequence.

Assume that the sequences have length ℓ and each site of the sequences has α possible elements. The number of all possible combinations is equal to α^ℓ. Each combination is a possible sequence. The major difficulty in analyzing sequence data arises from the lack of relationships among the sequences. Without knowing the sequences' relationships, it is not clear how to define meaningful quantities for the problem. Thus, in order to proceed, we need to introduce some kind of relationships among the sequences, and such relationships should be chosen according to the nature of the data in question. For instance, for categorical sequence data, each site is allowed to mutate from one element to all the other elements. Such a relationship can be represented by a *complete graph*, and the whole sequence space can then be represented by a Cartesian product of ℓ complete graphs, namely, a *Hamming graph* [14]. For ordinal sequence data considered here, each site can only jump from one element to the element(s) next to it in order. This relationship can be represented by a *path graph*, and thus we can use an ℓ-fold Cartesian product of path graphs, which results in a *grid graph*, to represent the sequence space. Figure 1 shows an example of this procedure with $\alpha^\ell = 3^3$.

Having represented the sequence space by a grid graph, we denote the probability for a sequence by $Q_{i_1\cdots i_\ell}$ where the indices $i_1, \ldots, i_\ell$, each corresponding to a site, are from 1 to α. We use this notation to emphasize that we are treating the probability distribution Q as an ℓ-dimensional array with α elements in each direction. Then we parametrize the distribution Q by a field ϕ in the following manner:

$$Q_{i_1\cdots i_\ell} = \frac{e^{-\phi_{i_1\cdots i_\ell}}}{\sum_{j_1\cdots j_\ell} e^{-\phi_{j_1\cdots j_\ell}}}. \tag{1}$$

Doing so, Q automatically satisfies the conditions, $Q_{i_1\cdots i_\ell} \geqslant 0$ and $\sum_{i_1\cdots i_\ell} Q_{i_1\cdots i_\ell} = 1$, required in order for it to be a valid probability distribution, whereas the field ϕ can take on any form or any value.

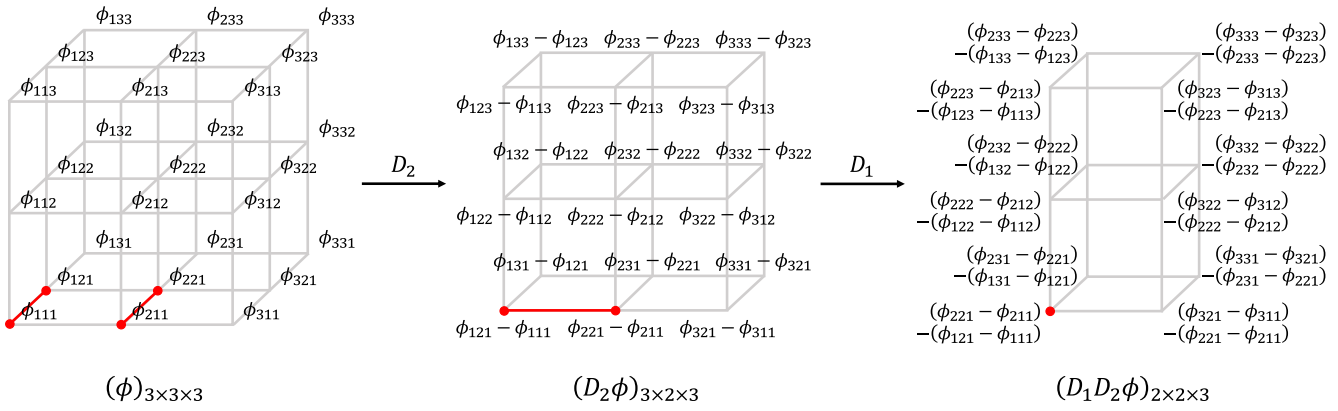

FIG. 2. Illustration of taking discrete differences of a field ϕ defined on an $\alpha^\ell = 3^3$ grid graph, as the one shown in Fig. 1. By construction, ϕ is a three-dimensional array with three elements in each direction. Upon taking two consecutive discrete differences, first along the second direction (D_2) and then along the first direction (D_1), we obtain an array $D_1D_2\phi$ with dimension $2 \times 2 \times 3$. In each step, the components of the resulting array are the differences of the adjacent values of the previous array in that direction. Each component of the array $D_1D_2\phi$ involves four sequences that form a face in the sequence space.

To control the "smoothness" of ϕ, that is, how ϕ changes from one sequence to another, we impose the following prior distribution on it:

$$p(\phi|a) \propto e^{-S_a^0[\phi]}, \tag{2}$$

where the functional $S_a^0[\phi]$, called the prior action, is given by

$$S_a^0[\phi] = \frac{a}{2s}\phi^T \Delta^{(P)}\phi. \tag{3}$$

Note that we have used the vector notation $\phi^T \Delta^{(P)}\phi$ to represent the operation acting on ϕ in order to make the formalism here resemble its categorical analog [14] as much as possible. In Eq. (3), a is a hyperparameter that ranges from zero to infinity and controls the overall smoothness of the field ϕ. Smaller values of a correspond to more rugged ϕ, and larger values of a correspond to more uniform ϕ. The constant s is arbitrary, but we will see that there is a natural choice for it in a moment. Finally, the parameter P represents the order of the operator Δ employed to compute the smoothness measure of ϕ.

Our treatment of ordinal sequence data here is similar to what we do with continuous variables in numerical calculations, in which we first put the problem on a grid and then proceed to compute derivatives of some quantities. However, in our previous work [14] we found that for functions defined in sequence space it was more meaningful to measure their smoothness by a statistic called "association," such as odds, odds ratio, ratio of odds ratios, etc. [38], than by the ordinary derivatives. Association provides an estimate of how strongly sequence sites are correlated. Thus, we define the operator Δ in the prior action as follows:

$$\phi^T \Delta^{(P)}\varphi = \frac{1}{P!}\sum_{i_1,\ldots,i_P} \langle D_{i_1}\cdots D_{i_P}\phi, D_{i_1}\cdots D_{i_P}\varphi\rangle, \tag{4}$$

where the indices $i_1, \ldots, i_P$ must satisfy the conditions $1 \leqslant i_1, \ldots, i_P \leqslant \ell$ and $i_1 \neq \cdots \neq i_P$. These conditions immediately imply that $P = 1, 2, \ldots, \ell$. Notice that ϕ and φ are both an ℓ-dimensional array with α elements in each direction. The symbol D_i denotes taking the discrete difference along the ith direction of an array, and the symbol $\langle u, v\rangle$ denotes taking the "inner product" of the two arrays u and v, that is, multiplying u and v element by element, and then summing up the results.

So what this operator $\Delta^{(P)}$ does in the prior is that, given a value of P, $1 \leqslant P \leqslant \ell$, it first calculates the discrete difference of the field ϕ along all possible combinations of P different directions, and then sums up the squares of all the differences. As an example, consider the toy model in Fig. 1, in which ϕ is a three-dimensional array with three elements in each direction. Let $P = 2$, for instance, then

$$\phi^T \Delta^{(2)}\phi = \|D_1D_2\phi\|^2 + \|D_1D_3\phi\|^2 + \|D_2D_3\phi\|^2, \tag{5}$$

where $\|u\|^2 \equiv \langle u, u\rangle$ and we have used the fact that the difference operation is commutative, $D_iD_j = D_jD_i$. Fig. 2 illustrates the procedure of computing the first term in Eq. (5), $\|D_1D_2\phi\|^2$. As shown in the figure, ϕ is by construction a $3 \times 3 \times 3$ array. Taking a discrete difference along the second direction, we obtain an array $D_2\phi$ of dimension $3 \times 2 \times 3$, whose components are differences of adjacent values of ϕ in the direction. Taking another discrete difference along the first direction, we obtain another array $D_1D_2\phi$ of dimension $2 \times 2 \times 3$, whose components are differences of adjacent values of $D_2\phi$ in the first direction. The quantity $\|D_1D_2\phi\|^2$ is then equal to the sum of the squares of all the components in the array $D_1D_2\phi$. The second and third terms in Eq. (5) can be computed in a similar way, and the grand total of the three terms gives the value of $\phi^T \Delta^{(2)}\phi$.

In the example above, each component of the array $D_1D_2\phi$ is a hierarchical difference of ϕ involving four sequences that form a face in the sequence space; see Fig. 2. Since ϕ and Q are related via Eq. (1), hierarchical differences of ϕ translate into (negative logarithm of) hierarchical ratios of Q. For instance, $(\phi_{221} - \phi_{211}) - (\phi_{121} - \phi_{111}) = -\log\{(Q_{221}/Q_{211})/(Q_{121}/Q_{111})\}$. In this case, the hierarchical ratio of Q represents a second-order association (i.e., odds ratio) between the first and second sites, both involving elements 1 and 2, conditional on the third site being fixed at 1. In general, the operator $\Delta^{(P)}$ requires taking discrete differences of ϕ in P different directions with the remaining directions being the background, and the 2^P sequences involved in this procedure form a hypercube of dimension P. For example,

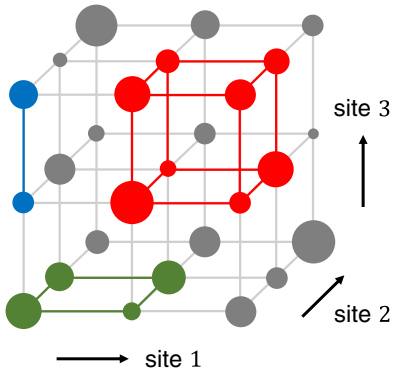

FIG. 3. Same grid graph as in Figs. 1 and 2. The computation of Pth-order associations involves probabilities of the sequences that form a P-dimensional hypercube. For $P = 1, 2, 3$, the hypercube is edge (blue), face (green), and cube (red), respectively. In this graph, there are 54 edges, 36 faces, and 8 cubes. The edge highlighted in blue represents a first-order association (i.e., odds) involving elements 2 and 3 of the third site and conditional on the first and second sites both being fixed at 1: Q_{113}/Q_{112}. The face highlighted in green represents a second-order association (i.e., odds ratio) involving elements 1 and 2 of both the first and second sites and conditional on the third site being fixed at 1: $(Q_{221}/Q_{211})/(Q_{121}/Q_{111})$. The cube highlighted in red represents a third-order association (i.e., ratio of odds ratios) that involves elements 2 and 3 of the first site, elements 1 and 2 of the second site, and elements 2 and 3 of the third site: $\{(Q_{323}/Q_{313})/(Q_{223}/Q_{213})\}/\{(Q_{322}/Q_{312})/(Q_{222}/Q_{212})\}$.

for $P = 1, 2, 3$, the hypercubes are edges, faces, and cubes, respectively, as shown in Fig. 3. Each P-dimensional hypercube gives a value of Pth-order association, which tells us how strongly the P sites are correlated with respect to the elements involved at each site and a certain background. By definition, the value of an association is positive and can be very large or very small. We can look at the logarithmic value of an association instead. When doing so, the value ranges from negative infinity to positive infinity. The larger the magnitude of a logarithmic association, the stronger the correlation. A zero logarithmic association means there is no correlation.

We denote a Pth-order association by $\mathcal{A}_k^{(P)}$, where k is the index of the hypercube. This suggests that we can choose the constant s in the prior action to be the number of P-dimensional hypercubes embedded in the grid graph, and it is equal to $s = \binom{\ell}{P}(\alpha-1)^P\alpha^{\ell-P}$. Equation (4), with $\phi = \varphi$, can then be rewritten as

$$\phi^T \Delta^{(P)} \phi = \sum_{k=1}^{s} \left(\log \mathcal{A}_k^{(P)}\right)^2, \tag{6}$$

which is a sum of squared log P-associations over all P-dimensional hypercubes. The prior distribution of ϕ, in turn, can be expressed in terms of associations as

$$p(\phi|a) \propto \prod_{k=1}^{s} \exp\left[-\frac{1}{2}\frac{(\log \mathcal{A}_k^{(P)})^2}{s/a}\right]. \tag{7}$$

As we will see below, this expression is very informative and can help us understand some general properties of the solutions.

Having defined the prior distribution of ϕ, we can obtain its posterior distribution by multiplying the prior by the data likelihood. The resulting posterior distribution of ϕ can be expressed as

$$p(\phi|\text{data}, a) \propto e^{-S_a[\phi]}, \tag{8}$$

where $S_a[\phi]$ is called the posterior action and is given by

$$S_a[\phi] = \frac{a}{2s}\phi^T \Delta^{(P)}\phi + NR\phi + Ne^{-\phi}. \tag{9}$$

The derivation can be found in Appendix A. Here N is the number of sequences in the data set, and R is an array having the same shape as ϕ and consisting of the observed frequencies. Again, we have used vector notation in Eq. (9) in order for it to resemble its categorical counterpart [14]. In the notation introduced in Eq. (4), $R\phi = \langle R, \phi\rangle$ and $e^{-\phi} = \langle \mathbf{1}, e^{-\phi}\rangle$, where $\mathbf{1}$ is an array having the same shape as $e^{-\phi}$ and full of ones. Given a value of a, the *maximum a posteriori* (MAP) estimate of ϕ, which corresponds to the mode of the posterior distribution, can be obtained by minimizing the action $S_a[\phi]$. It can be shown that the MAP estimate, denoted ϕ_a, satisfies the following equation of motion (EOM):

$$\frac{a}{s}\Delta^{(P)}\phi_a + NR - Ne^{-\phi_a} = 0. \tag{10}$$

Below we use the equation of motion to derive some general properties of the MAP estimates.

From Eq. (4), it is clear that the ℓ-dimensional array $\mathbf{1}$ consisting of ones is in the kernel of $\Delta^{(P)}$ for any P. This is because $\mathbf{1}$ is uniform, and so a single difference along any direction will result in another array full of zeros. Taking the "inner product" of $\mathbf{1}$ from left with Eq. (10) leads to $\sum_{i_1\cdots i_\ell} e^{-(\phi_a)_{i_1\cdots i_\ell}} = \sum_{i_1\cdots i_\ell} R_{i_1\cdots i_\ell} = 1$. Thus, we have $Q_a = e^{-\phi_a}$ from Eq. (1). In the limit $a = 0$, the EOM implies that $Q_0 = e^{-\phi_0} = R$, the observed frequency. This can be understood from Eq. (7). We see that when $a = 0$, all the values of $\log \mathcal{A}^{(P)}$ have an infinite variance, meaning that they are totally unconstrained. That is, the prior has no effect at all, and the result is completely determined by the data. On the other hand, in the limit $a \to \infty$, the EOM requires $\Delta^{(P)}\phi_\infty = 0$, meaning that ϕ_∞ is restricted to the kernel of $\Delta^{(P)}$. From Eq. (7), we see that in this limit all the values of $\log \mathcal{A}^{(P)}$ are centered at zero with a zero variance, which means that the MAP estimate has vanishing Pth-order association. In other words, ϕ_∞ can have associations only of up to order $P-1$. We can then use "indicator functions" of up to $P-1$ sites as the kernel basis of $\Delta^{(P)}$. Denote the k-site indicator function by $I^{e_1,\ldots,e_k}_{s_1,\ldots,s_k}$, where the subscripts $s_1, \ldots, s_k$, each from 1 to ℓ, represent the sites, and the superscripts $e_1, \ldots, e_k$, each from 1 to α, represent the elements. The indicator function has the same shape as ϕ, and a component of $I^{e_1,\ldots,e_k}_{s_1,\ldots,s_k}$ is equal to one if the elements at sites $s_1, \ldots, s_k$ of the corresponding sequence are $e_1, \ldots, e_k$, respectively, and is equal to zero otherwise. See Fig. 4 for some examples. Note that the indicator functions can be expressed as $I^{e_1,\ldots,e_k}_{s_1,\ldots,s_k} = I^{e_1}_{s_1} \odot \cdots \odot I^{e_k}_{s_k}$, where $\odot$ represents element-wise multiplication. The set of indicator functions for $k = 1, 2, \ldots, P-1$ is able to span the kernel of $\Delta^{(P)}$. Taking the "inner product" of $I^{e_1,\ldots,e_k}_{s_1,\ldots,s_k}$ from left with the EOM, the first term vanishes and we have

$$I^{e_1,\ldots,e_k}_{s_1,\ldots,s_k} Q_a = I^{e_1,\ldots,e_k}_{s_1,\ldots,s_k} R \tag{11}$$

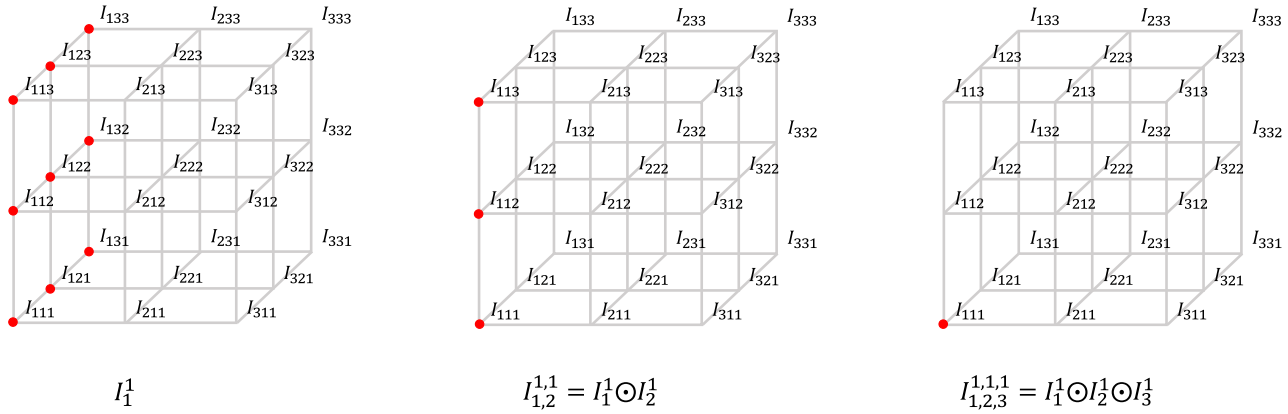

FIG. 4. From left to right, examples of 1-site, 2-site, and 3-site indicator functions. The indicator functions have the same shape and index as the field ϕ in question (in this case, the one in Fig. 2). A node with or without a red dot carries a value of 1 or 0. The 1-site indicator function I_1^1 is equal to 1 in a plane made up of sequences having element 1 at the first site. The 2-site indicator function $I_{1,2}^{1,1}$ is equal to 1 along a line consisting of sequences having element 1 at both site 1 and site 2. Note that this line is the intersection of the plane I_1^1 and the plane I_2^1. Similarly, the 3-site indicator function $I_{1,2,3}^{1,1,1}$ is equal to 1 only at the node that corresponds to the right sequence. This node is the intersection of the three planes I_1^1, I_2^1, and I_3^1.

for $k = 1, 2, \ldots, P-1$. That is, the MAP estimate Q_a has the same marginal frequencies of up to $P-1$ sites as the data R.

Note that in the limit $a \to \infty$, the probability distribution is constrained solely by the marginal frequencies. For $P = 2$, the distribution is constrained by the 1-site marginal frequencies and is thus equivalent to the position weight matrix model. For $P = 3$, the distribution is constrained by the 1-site and 2-site marginal frequencies and is thus equivalent to the Potts model. In other words, these traditional models are limiting cases of our formalism. (We note that, although the position weight matrix model and the Potts model are normally used for categorical data, they can be used for ordinal data as well, as long as the data are discrete.) Moreover, in this limit the distribution is given by $Q_\infty = e^{-\phi_\infty}$, where $\phi_\infty = \sum c^{e_1,\ldots,e_k}_{s_1,\ldots,s_k} I^{e_1,\ldots,e_k}_{s_1,\ldots,s_k}$ with some set of coefficients $c^{e_1,\ldots,e_k}_{s_1,\ldots,s_k}$. These coefficients are the only degrees of freedom the distribution has in this limit, and they are chosen so that the distribution satisfies the constraints in Eq. (11). From a theorem in information theory [39], we know that Q_∞ must therefore be the unique maximum entropy distribution. That is, it is the most uniform, and hence has the maximum entropy, distribution satisfying the moments constraints, Eq. (11). Incidentally, the position weight matrix model and the Potts model are also known as the additive and pairwise maximum entropy (MaxEnt) estimates, respectively.

In summary, we use a field ϕ to parametrize the probability distribution Q that we want to infer from a sample of ordinal biological sequences. We use a grid graph to capture the relationships among the sequences, and then we construct an operator $\Delta^{(P)}$ that can compute the smoothness of the field (or equivalently, the distribution). The optimal distribution is a compromise between the data and a smoothness measure that is controlled by a hyperparameter a. By varying the value of the hyperparameter a from zero to infinity, we can obtain a family of probability distributions from the most rugged data frequency to the most uniform maximum entropy estimate, and each distribution in this family has the same marginal frequencies of up to $P-1$ sites as the data. The remaining question is then how to find the optimal value of the hyperparameter. In Bayesian field theory, there are two ways to answer this question: one is by computing Bayesian evidence and the other is by doing cross validation. Here we choose to do k-fold cross validation. In k-fold cross validation, we first split the data into k even partitions, and then we repeatedly solve for Q_as with $k-1$ partitions and compute the likelihood of the remaining partition with the resulting Q_as. The optimal value of a is then determined by the likelihood averaged over the k partitions, and we denote the optimal distribution by Q^*. This procedure may be time-consuming and computationally demanding, but it is more applicable than computing Bayesian evidence in most practical situations.

III. RESULT

To demonstrate the utility of our density estimation method, here we apply it to the aneuploidy data of cancer patients from the TCGA project [37]. A cell or an organism is said to be euploid if it has an exact multiple of the haploid number of chromosomes, otherwise it is called aneuploid. For instance, a human cell having three copies of each chromosome is euploid, whereas a human cell having three copies of 22 of the chromosomes but four copies for the remaining chromosome is aneuploid. Aneuploidy is closely related to karyotype. A karyotype records the absolute copy number of each chromosome, while aneuploidy describes by how much the subject deviates from being euploid. An euploid cell or organism may still survive because of balanced gene expression. For an aneuploid cell or organism, unbalanced gene expression most often results in reduced cellular growth rates, but some aneuploid states can in fact accelerate growth [40].

In our previous work [14], we analyzed the TCGA aneuploidy data of 10 522 cancer patients across 33 tumor types [34]. The data set contained the aneuploidy of the 22 autosomes, that is, chromosomes other than the sex chromosome. For each patient, we first classified each chromosome as normal, if it was neutral, or abnormal, if it was deleted, amplified,

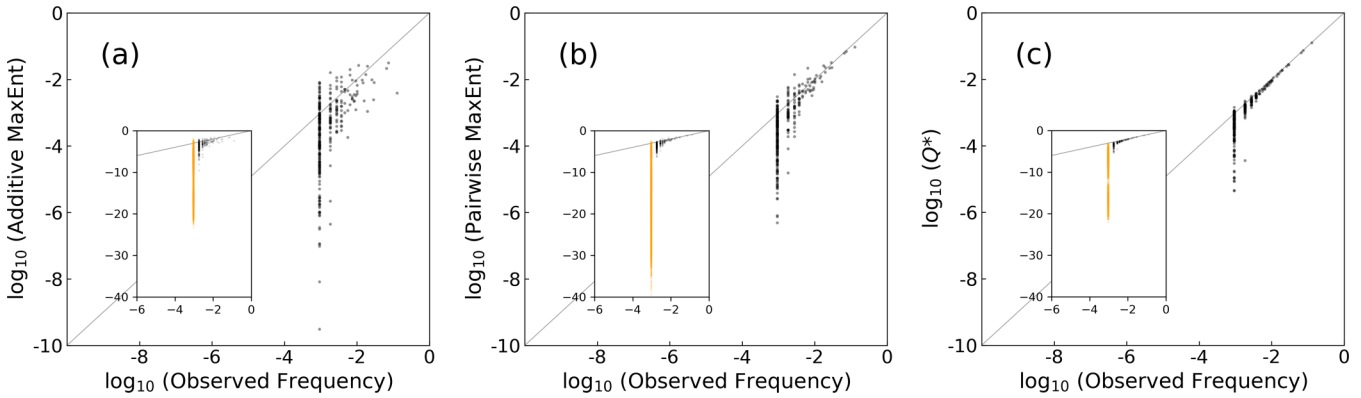

FIG. 5. Probabilities of the 294 unique sequences in the TCGA gliomas data set estimated with (a) additive MaxEnt, (b) pairwise MaxEnt, and (c) ordinal SeqDEFT. The insets show the same thing but for all the 16 384 possible sequences in this problem. In order to plot the sequences that were not observed, we add a pseudocount to the data set. The unobserved sequences are represented by the orange dots in the last strip in each inset.

or something more complex. Thus, the aneuploidy data of each patient were simplified to a string of 22 binaries. Then we used SeqDEFT to estimate the probability distribution with the data set. From a visualization of the resulting probability landscape, we were able to identify some outstanding peaks associated with particular tumor types. One of the peaks corresponded to simultaneous alterations in chromosomes 6, 7, 9, 10, 19, 20, and was found to associate with gliomas. Now we can analyze the case one step further by taking the original four states of aneuploidy into consideration.

We focus on the subset of 1086 patients that have gliomas [41–43]. According to the result of our previous analysis [14], we look only at the subset of chromosomes that consists of chromosomes 1, 6, 7, 9, 10, 19, 20. (Chromosome 1 is included because of the 1p/19q codeletion commonly observed in low-grade gliomas [36].) Each chromosome can take on one of the four states: deleted (D or "−"), neutral (N or " "), amplified (A or "+"), or complex (C or "∗"). The four states are arranged in the order D, N, A, C to represent the amount of each chromosome relative to being euploid. This choice of chromosomes and states results in a sequence space of dimension $4^7 = 16\,384$. By comparison, the data set contains 1086 sequences (CNNNNCN, NNANDNN, etc.), among which only 294 sequences are unique. In other words, the observed sequences occupy only 1.79% of the whole sequence space.

A. Probability distribution

We use our method with $P = 2$, along with fivefold cross validation, to estimate the probability distribution. In order to compare, we also fit an additive MaxEnt model and a pairwise MaxEnt model to the same data set. The choice of the value of the hyperparameter P deserves a few words. In principle, the value of P can be determined entirely from the data. To do so, we compute each family of probability distributions for $P = 1, 2, \ldots, \ell$. The value of P can then be selected from the cross-validated log likelihoods of the data, for example. In practice, however, this method is not very practical. This is because sequence data are usually so sparse that the magnitude, or even the existence, of higher-order associations cannot be established, and thus the calculations with high values of P may be problematic. As a result, the value of P is usually hand-picked by user. We have seen that we should not choose a high value of P. On the other hand, the operator $\Delta^{(P)}$ with $P = 1$ is identical to the graph Laplacian, and thus its function is to make the difference between neighboring sequences as small as possible. But we have seen in biology that a single point mutation can have a dramatic effect on a sequence. Thus, a value of $P > 1$ should be chosen. Here we find that $P = 2$ works best for the current problem. From the theory we know that the resulting distribution will have the same 1-site marginal frequencies as the data.

The results are shown in Fig. 5. From the figure we see that the three models all make reasonably accurate estimates of probabilities for the observed sequences, with the pairwise model providing more precise estimates than the additive model, and our method providing even more precise estimates than the pairwise model. This trend is also reflected in the resulting cross-validated log likelihoods; the additive, pairwise, and ordinal SeqDEFT models give an average cross-validated log likelihood of −1325, −1180, and −1140, respectively. For the unobserved sequences (represented in the last strip in the inset of each subplot), however, it is slightly different. We see that for the unobserved sequences the estimates of the additive model is actually tighter than that of the pairwise model. Intuitively, this is because, although the data set is large enough to establish the frequency at each site, it is not large enough to establish the correlations between each pair of sites. As a result, the pairwise model gives a wider range of predictions for the unobserved sequences. For ordinal SeqDEFT, on the other hand, in regions where there are no data the distribution is mainly determined by the smoothness operator in the prior. Thus, the distribution therein, which is obtained by pushing the posterior action to zero as much as possible, is similar to the one that is restricted within the kernel of the operator, which for $P = 2$ is an additive model.

We have seen that our method is capable of inferring the underlying probability distribution given a sample of ordinal sequences. In some cases, the probability distribution itself is what we want. Oftentimes such a probability distribution is then used in follow-up analyses or applications. An example is the generative modeling, in which a probability distribution

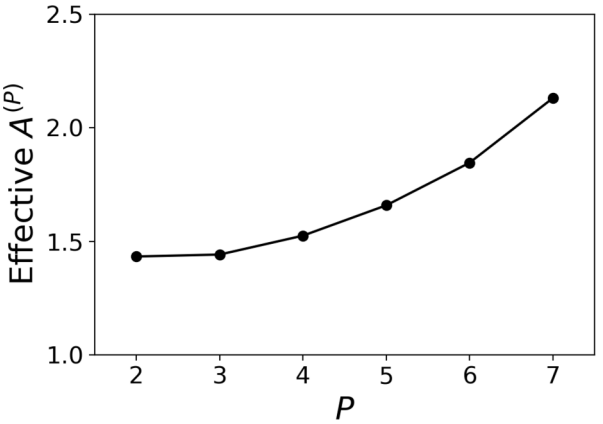

FIG. 6. Association spectrum inferred from the probability distribution Q^*. For each order P, the effective magnitude of the association is defined as the exponential of the root-mean-squared value of $\log \mathcal{A}^{(P)}$ computed by using Eq. (12). The larger the value of $\mathcal{A}^{(P)}$, the stronger the association. An $\mathcal{A}^{(P)}$ of 1.0 means that there is no association at all.

is used to generate samples having similar features of interest. Below we demonstrate two more things one can do with a probability distribution. One is studying the associations of the sequence sites, the other is probing the global geometry of the probability landscape. We will look at the associations first and then move on to the probability landscape.

B. Associations of sequence sites

From the expression of the smoothness operator in Eq. (6) we see that sandwiching the operator $\Delta^{(P)}$ with the field ϕ representing a distribution results in a sum of squared log P-associations over all P-dimensional hypercubes. With this quantity we can easily compute the root-mean-squared value of $\log \mathcal{A}^{(P)}$ as

$$(\log \mathcal{A}^{(P)})_{\text{rms}} = \sqrt{\frac{1}{s}\,\phi^T \Delta^{(P)} \phi}, \tag{12}$$

and use its exponential as the effective magnitude of the Pth-order association predicted by the distribution in question. Given a distribution, we can repeatedly use Eq. (12) with $P = 1, 2, \ldots, \ell$, to estimate the strength of the association among P sequence sites. The resulting "association spectrum" computed with the distribution Q^* we obtained above is shown in Fig. 6. Note that the result with $P = 1$ is more like a measure of the overall uniformity of the distribution, rather than an association, so it is not included in the figure. From Fig. 6 we can see that the sequence sites actually have stronger higher-order associations, although all the associations are only mild. This should not be too surprising since the chromosomes used in the data set were chosen because of the seeming concurrence of their alterations in gliomas.

In principle, each value of effective $\mathcal{A}^{(P)}$ in the association spectrum should be computed with a number of log P-associations, each corresponding to a particular set of P sites and a certain background. Equation (12) provides a direct formula for the root-mean-squared log P-association. But, still, we can compute those log P-associations explicitly and look into them to try to extract more information. When computing a value of, say, log 2-association (i.e., odds ratio), we need to specify two elements at each of the two sites. For example, 1 and 2 at one site, and also 1 and 2 at the other. Together, they form four patterns, 11, 12, 21, 22, at the two sites, which correspond to a face in the grid graph representing the sequence space;see Fig. 3. Notice that, for a given face, there are multiple ways to form the odds ratio, such as $(Q_{22}/Q_{12})/(Q_{21}/Q_{11})$, $(Q_{12}/Q_{11})/(Q_{22}/Q_{21})$, $(Q_{21}/Q_{22})/(Q_{11}/Q_{12})$, $(Q_{11}/Q_{12})/(Q_{21}/Q_{22})$, and so on. But no matter how the odds ratio is formed, we obtain the same magnitude for this log 2-association. This is because there is no intrinsic "direction" in the computation of $\log \mathcal{A}^{(P)}$. For the sake of convenience, however, we may regard the first site as being mutating from 1 to 2, and the second site from 1 to 2 too. In this case, the odds ratio is given by $(Q_{22}/Q_{12})/(Q_{21}/Q_{11})$, and the magnitude of its logarithmic value can be interpreted as the coupling strength between the two mutations. When doing so, association becomes very similar to the epistatic coefficient in biology, which is often used to describe how the effect of a gene mutation is affected by one or more other gene mutations. In our case, it is how the frequency, or the existence, of the sequence is affected by the mutations.

Although this thinking may seem appealing, it may lead us to a wrong conclusion when interpreting the result. For instance, we may tend to think a strong association between the two sites means that when the first site mutates from 1 to 2, the second site mutates from 1 to 2 as well, and thus the pattern appearing at the two sites should be 11, 22, or both, rather than 12 or 21. Though this is indeed one possibility, other possibilities do exist. In fact, one of the other possibilities is a complete opposite of our conclusion, that is, the pattern that appears at the two sites could be 12, 21, or both, rather than 11 or 22. As stated above, this problem comes from the fact that a value of log P-association is given by a ratio of 2^P probabilities, and there are multiple ways to arrange the probabilities so that the log P-association has the same magnitude. Thus, there are many scenarios that can result in a strong association between the P sites. The safest way to find out the pattern corresponding to a strong association is therefore by looking into the probabilities of the sequences involved. In our example, we may find that 22 has a much larger probability than 11, 12, and 21, and thus is the dominant pattern at the two sites. Note that the pattern found in this manner is merely a "local rule," since we are comparing only the four patterns, 11, 12, 21, and 22. If we investigate another face at the same two sites formed by patterns, say, 22, 23, 32, and 33, we may find that a different pattern 33, instead of 22, is the most probable one. Local rules like these might be useful in some situations, but a more useful piece of information would be a complete list of the patterns that do appear at the P sites. Such a picture can be patched up with local rules derived from all the $(\alpha - 1)^P$ hypercubes. A more straightforward, and perhaps simpler, way to get the picture is by looking into the probabilities of the α^P associated sequences directly.

Having computed the associations, the next question is how to present the result in an informative way. For the case of $P = 2$, it is an easy task as pairwise associations can be readily represented by using a heatmap. For an example of splice sites in human genome, see Ref. [14]. For the cases of $P > 2$,

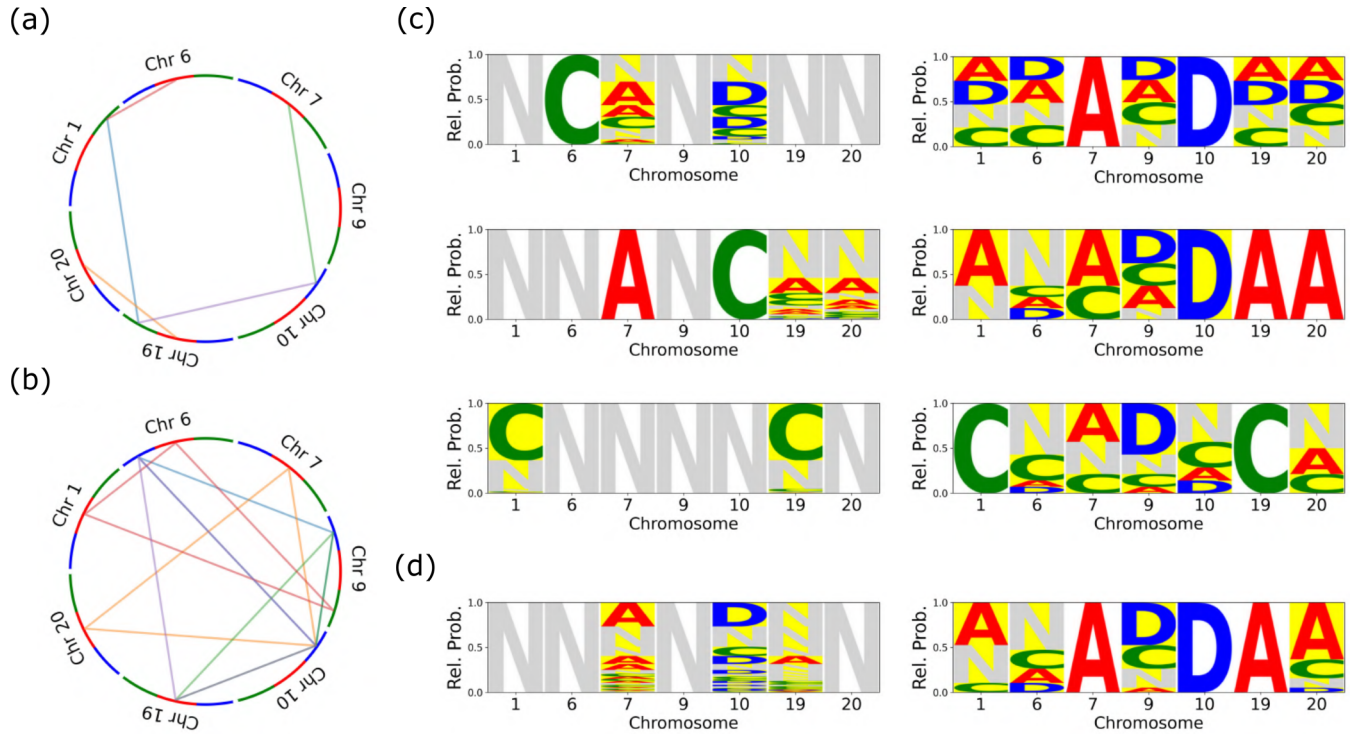

FIG. 7. (a), (b) Schematic representation of the five strongest associations for $P = 2$ and for $P = 3$ inferred from the probability distribution Q^*. In each plot, the circumference of a circle is partitioned into seven arcs, each corresponding to a chromosome. Each arc is also partitioned into three segments, each corresponding to one of the mutations: D $\rightarrow$ N (blue), N $\rightarrow$ A (red), A $\rightarrow$ C (green). Each Pth-order association is represented by a P-sided polygon within the circle. There are five polygons in each plot, colored in blue, orange, green, red, and violet, according to their association strengths in descending order. (c, d) Left: Sequence logos that show the patterns appearing at the P sites of interest (highlighted in yellow) conditional on a certain background. Note that each pattern consists of P letters in the horizontal direction, so a letter may appear multiple times at one site. The height of each pattern is proportional to its relative probability. Right: Sequence logos that show the backgrounds (highlighted in yellow) in which a particular pattern tends to appear at the P sites of interest. The height of each letter is proportional to its relative probability.

however, presenting those higher-order associations becomes a difficult task. It is clear that the solution should depend on what we want to know about the system. If what we want to know is what combinations of mutations have the strongest associations, and in what backgrounds, we can represent the result in the following manner. First, we partition the circumference of a circle into ℓ arcs, each arc corresponding to a sequence site, and then we partition each arc into $\alpha - 1$ segments, each segment corresponding to a mutation. Then any Pth-order association can be represented by a P-sided polygon within the circle connecting the particular mutations involved. This is similar to the "circos plots" commonly used in visualizing genomic data [44]. Meanwhile, if desired, the backgrounds can be represented by a "sequence logo" [45].

The five strongest pairwise and three-way associations inferred from the probability distribution Q^* are shown in Figs. 7(a) and 7(b). From Fig. 7(a), we can see some well-known chromosome abnormalities in gliomas, such as $7^+/10^-$ [36] (conditional on $1/6^*/9/19/20$) and $19^+/20^+$ [46] (conditional on $1/6/7^+/9/10^*$). We also see a strong association between 1^* and 19^*, conditional on 6/7/9/10/20, which is likely a manifestation of the 1p/19q codeletion [36]. As explained above, to see what is really going on at those pairs of sites, we need to look into the probabilities of the associated sequences. Here we examine all the patterns that are allowed at the two sites, not just the patterns formed by the mutations, to get a more complete picture. The results are displayed as sequence logos in the left panel of Fig. 7(c). We can see that the patterns $7^+/10$, $19^+/20^+$, and $1^*/19^*$ are indeed the most prominent pattern at the two sites when both sites mutate. Note that, for chromosomes 1 and 19, the pattern $1^*/19^*$ is more probable than the wild type.

These pairwise associations, $7^+/10^-$, $19^+/20^+$, and $1^*/19^*$, have an effective strength (averaged over all backgrounds) of 2.52, 2.02, and 1.58, respectively, which are greater than 99%, 98%, and 88% of all possible pairwise associations. We plot the backgrounds in which $7^+/10^-$, $19^+/20^+$, or $1^*/19^*$ is the most probable pattern at the two sites as sequence logos in the right panel of Fig. 7(c). The backgrounds for each pattern are found in the following manner. For each background, we first find the probabilities of all the patterns (DD, DN, …, CC) at the two sites. This gives us a vector of 16 components, and we normalize it to 1. Then a "similarity score" can be obtained by taking the inner product of this vector with another vector constructed in a way that the component corresponding to the pattern of interest (e.g., AD) is 1 and 0 otherwise. A background is selected if the similarity score is greater than 0.5. We find 950, 8, and 14 backgrounds for $7^+/10^-$, $19^+/20^+$, and $1^*/19^*$, respectively. From the figure, we can see that the backgrounds for $7^+/10^-$

have no specificity, meaning that each state appears with an equal probability at the background sites. By contrast, the backgrounds for $19^+/20^+$ and $1^*/19^*$ show a mild to medium specificity.

Higher-order associations have been rarely identified in clinical studies of gliomas. A possible third-order association involves $7^+/10^-/19^+$ [47], which, however, does not show up in Fig. 7(b). We compute the probabilities of all the patterns at the three sites with a neutral background. The result is shown in the left panel of Fig. 7(d), from which we can see that $7^+/10^-/19^+$ is indeed the most probable pattern when the three sites all mutate. On the other hand, the effective strength (averaged over all backgrounds) of this abnormality is inferred to be 1.56, which is greater than 81% of all possible third-order associations. Moreover, we find the backgrounds in which $7^+/10^-/19^+$ takes place in the same manner as above. The 19 backgrounds found are plotted in the right panel of Fig. 7(d), which also show a mild specificity.

We have demonstrated that the association analysis presented above can help us find out the patterns that tend to appear at some sites in a certain background. It can also help us find out in what backgrounds a particular pattern is more likely to appear at some sites. This information, in turn, can tell us which mutations tend to take place together. These rules, or "biological grammar," dictate how different sequence sites should cooperate together in order to make a sequence "visible," and thus may be useful in the exploration of underlying physical or biological mechanisms. Here we analyze only the pairwise and three-way associations. Higher-order analysis can be carried out in the same manner.

C. Global geometry of probability landscape

Now we turn to investigate the probability landscape of the distribution Q^*. The distribution Q^* is a very high-dimensional object. Thus, in order to visualize it, we resort to a dimensionality reduction technique based on an evolutionary model [48]. In this method, the distribution Q^* is regarded as being the result of an evolutionary process. This evolutionary process can be represented by a reversible Markov chain, whose rate matrix is given by

$$T_{ij} = \frac{\log Q_j^* - \log Q_i^*}{1 - e^{-(\log Q_j^* - \log Q_i^*)}}, \tag{13}$$

where i and j denote mutationally adjacent sequences, and the diagonal element T_{ii} is chosen so that the row sums of the matrix T are all zero. The derivation of Eq. (13) can be found in Appendix B. Order the eigenvalues of the matrix $-T$ in such a way that $0 = \lambda_1 < \lambda_2 \leqslant \lambda_3 \leqslant \cdots \leqslant \lambda_{\alpha^\ell}$. A d-dimensional visualization of the Markov chain, and hence the distribution Q^*, can be obtained by plotting each sequence i with coordinates $\sqrt{1/\lambda_2}\, r_i^{(2)}, \ldots, \sqrt{1/\lambda_{d+1}}\, r_i^{(d+1)}$, where $r^{(k)}$ is the right eigenvector of $-T$ corresponding to the eigenvalue λ_k. It can be shown that the reciprocal of the square root of the kth eigenvalue, $\sqrt{1/\lambda_k}$, is proportional to the variance explained in the direction of the kth eigenvector. Moreover, the squared distance between two sequences i and j in the visualization approximates the expected "commute time" between them (i.e., the time it takes to go from i to j and then back to i). For more details, see Ref. [48] and the supplement to [14].

Note that we do not actually need to diagonalize the rate matrix to construct the visualizations because we just need a few subdominant eigenvectors. Moreover, because the rate matrix is highly sparse, this can be done efficiently for sequence space containing millions of sequences.

Figure 8 shows a visualization of the probability distribution Q^* using the method described above with the first three rescaled eigenvectors. These rescaled eigenvectors are called "diffusion axes" since they capture the slow modes of the underlying evolutionary process [48]. Each node in the visualization represents a sequence and is colored by its inferred probability, and each edge represents a single point mutation. We can see that the sequences are divided into three distinct groups by Diffusion Axis 1. A closer inspection reveals that higher, medium, and lower values on Diffusion Axis 1 correspond to sequences with chromosome 10 being deleted or neutral, amplified, and complex, respectively. This segregation may seem strange at first since each chromosome is allowed to be in any one of the four states. To understand why this happens, we note that the chromosome abnormality 10^- is very common in gliomas. In fact, in the TCGA data set 10^- is more frequent than the wild type. On the other hand, the abnormality 10^+ is quite rare in the data set. Thus, from an evolutionary point of view, a population starting as a wild type is much more likely to evolve to 10^- than to 10^+. In other words, the commute time between wild type and 10^- is much shorter than the commute time between wild type and 10^+. In the visualization the sequences with chromosome 10 being deleted and the sequences with chromosome 10 being neutral form two individual groups; however, because of their very short commute time, they appear to be overlapping with each other and are far from the group of sequences with 10^+. The chromosome abnormality 10^* is also much more frequent than 10^+, which separates the two groups of sequences by a large distance in the visualization as well. As a result, the sequences are segregated into three distinct groups.

Diffusion Axis 3 is found to play a similar role as Diffusion Axis 1 in the visualization, with the controlling chromosome now being chromosome 6. More specifically, the region with lower values on Diffusion Axis 3 corresponds to a mix of sequences with chromosome 6 being deleted or neutral, whereas the regions with medium and higher values on Diffusion Axis 3 correspond to sequences with chromosome 6 being amplified and complex, respectively. Although the segregation along Diffusion Axis 3 is not as prominent as that along Diffusion Axis 1, it can still be discerned by the various shades of orange along the axis.

Unlike Diffusion Axes 1 and 3, Diffusion Axis 2 does not have a clear connection with the states of a certain chromosome. Nevertheless, the sequences form several stripes that stretch to different extents along Diffusion Axis 2. We find that the farthest nodes of each stripe always possess the pattern $1^*/19^*$ with some background, whereas the middle portion of each stripe consists exclusively of sequences either having $1^*/19^+$ or having $1^+/19^*$, and this trend vanishes as we move further toward the head of the stripe. What distinguishes the stripes? It is found that the sequences forming the leftmost stripe in the first group all have chromosome 9 being neutral

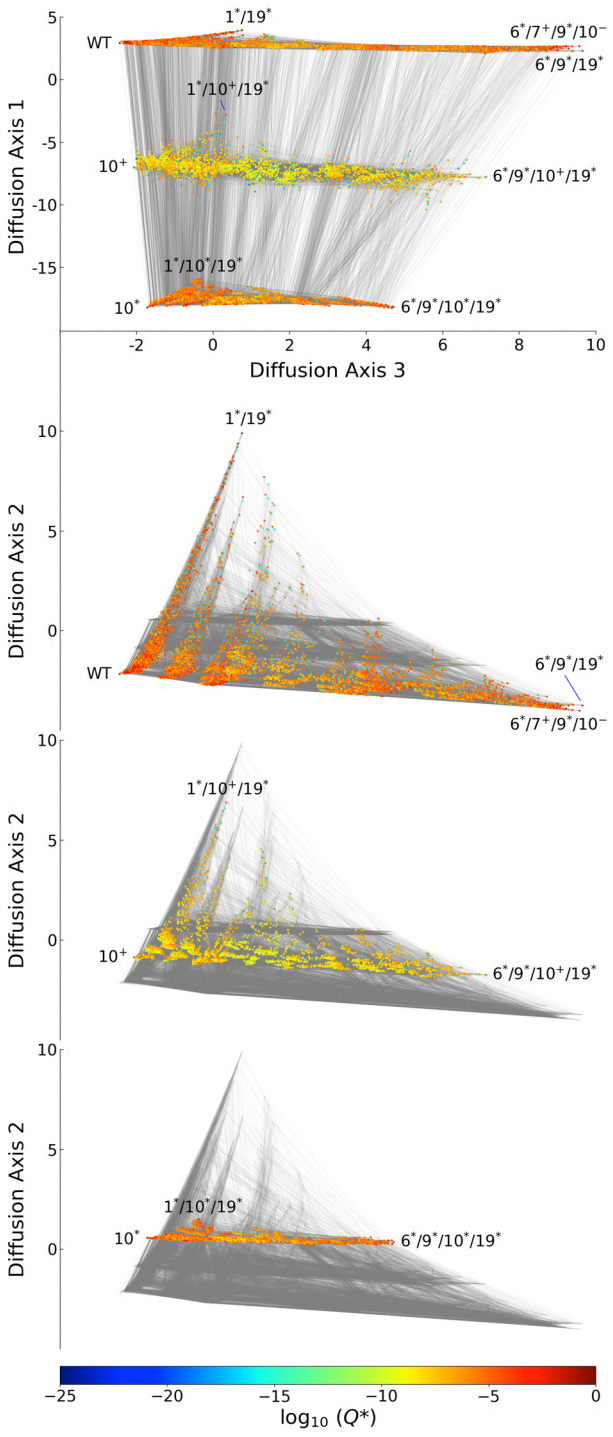

FIG. 8. Visualization of the probability distribution Q^*. Each node represents a sequence and is colored by its inferred probability, and each edge represents a single point mutation. The sequences form three groups, conditional on $10/10^-$, 10^+, or 10^*, respectively, which are plotted separately in the lower panel for the sake of visibility. WT, wild type.

or deleted, and the sequences forming the next stripe therein all have chromosome 9 being amplified. However, distinctions between other stripes are hard to find as the boundaries between them are blurry. In addition, we also find that in each group sequences having 7^+ tend to gather around the bottom (i.e., regions with lower values on Diffusion Axis 2), although some of them may scatter deep into a stripe.

From the global geometry of the probability landscape, Fig. 8, we can imagine an evolutionary process as follows. A population starts as a wild type, with all the chromosomes being neutral. As the population wanders around the sequence space, it has a much higher probability of staying in the original group than jumping to the group with 10^+. Even if the population does jump to the group with 10^+, it will soon jump back to the original group or jump to the group with 10^*. Let us focus on what is likely to happen to the population, assuming that it stays in the original group. Because the group with 10^- overlaps thoroughly with the group with 10 neutral, the population is very likely to pick up the abnormality 10^-. In addition, since the wild type is surrounded by sequences having 7^+ (which tend to gather near the bottom of the group), the population is also very likely to pick up the abnormality. This combined pattern $7^+/10^-$ is a signature of glioblastoma [36]. Sequences having both 7^+ and 10^- abnormalities form the routes from wild type to the bottom-right corner along Diffusion Axis 3. Our population may also take on alternative routes in the direction of Diffusion Axis 2. In doing so, it picks up either 1^* or 19^*, and eventually possesses the pattern $1^*/19^*$, which in our analysis is characteristic of low-grade gliomas [36]. Another possibility is that the population may pick up the abnormality $7^+/10^-$ on its way toward the end of a stripe. This corresponds to the clinical scenario that some low-grade gliomas may progress to glioblastoma [43].

Figure 9 shows how many times a particular chromosome abnormality with 10 or 10^- is observed in the TCGA data set. From the figure we can see that the abnormalities with glioblastoma being the diagnosis scatter primarily around the bottom, whereas the abnormalities with low-grade gliomas being the diagnosis tend to pile up at the upper corner. This picture is consistent with what we have inferred from an imaginary evolutionary process taking place in the probability landscape shown in Fig. 8.

IV. SUMMARY AND DISCUSSION

Within the framework of Bayesian field theory, we developed a novel approach to estimate the probability distribution from which a sample of ordinal biological sequences were drawn. This approach was based on the idea of representing the sequence space by a suitable graph. We found that for ordinal sequences a grid graph, which is a Cartesian product of path graphs, was a suitable choice as it can take into account the ordering of sequence elements. In such a representation, we then constructed an operator which can compute the associations among sequence sites of any order, and these associations were used to form a smoothness measure. This operator played a role like a "regularizer" in machine learning, allowing us to control the smoothness of the resulting distribution by tuning the value of a hyperparameter, and the optimal value of the hyperparameter was determined via k-fold cross validation. We applied this approach to the TCGA aneuploidy data of gliomas, and we found that the optimal distribution

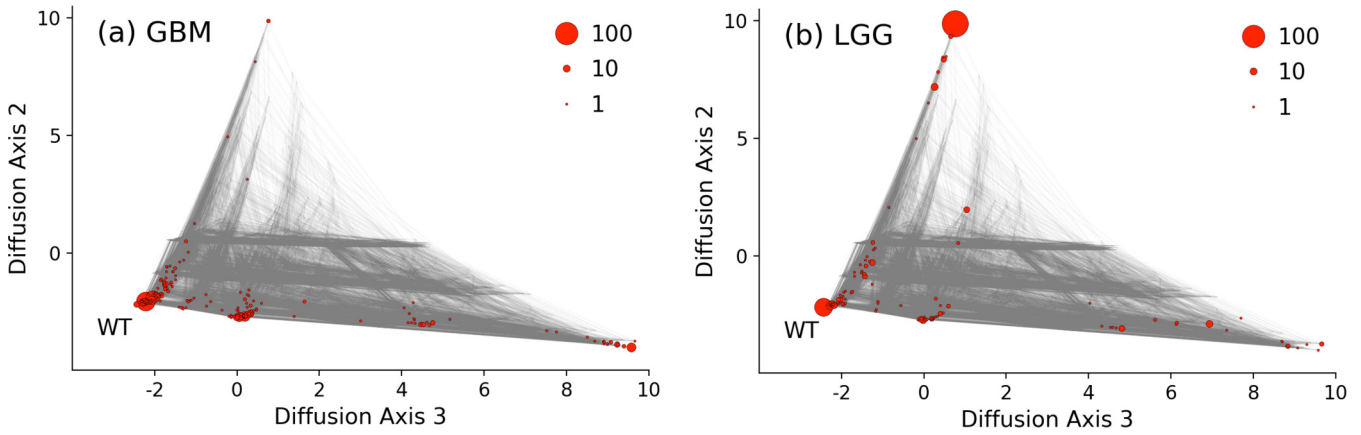

FIG. 9. Number of patients with a particular pattern of the seven chromosomes from the TCGA data set. Note that the data set is based on the old definitions of cancer types. Only the patients belonging to the group with higher values on Diffusion Axis 1 in Fig. 8 are counted. The patients are further separated into two subgroups, one consisting of those patients annotated as having glioblastoma (GBM, left panel) and the other with patients annotating as having low-grade gliomas (LGG, right panel). The size of each dot is proportional to the number of patients having the corresponding pattern. WT, wild type.

inferred with this approach had a better performance than the parametric additive and pairwise MaxEnt models.

We also did two follow-up analyses with the optimal distribution. First, we computed an association spectrum that revealed the association strengths of the sequence sites of all orders. We also looked into the associations of second and third orders to search for events of interest, for example, what combinations of mutations tended to occur simultaneously and in what backgrounds. This analysis provided us a way to figure out the underlying biological grammar for the system. Indeed, we were able to recover several well known chromosome abnormalities in gliomas. Second, we used a dimensionality reduction technique based on an evolutionary model to visualize the probability landscape. We saw that the global geometry of the probability landscape reflected some prominent features of the system. Moreover, we investigated how a population starting as the wild type might evolve to other states and eventually ended up in glioblastoma or low-grade gliomas. Looking at the system from this perspective helped us understand the underlying biological mechanisms.

The method developed in this work is very similar to its categorical counterpart introduced in Ref. [14]. In fact, the derivation of the posterior action $S_a[\phi]$ presented in Appendix A holds for both categorical and ordinal discrete data. The difference shows up only in the implementation of the methods through the prior, which requires setting up relationships among the sequences and constructing an appropriate operator in accord with the sequences' nature. Therefore, SeqDEFT and ordinal SeqDEFT share the same advantages and disadvantages. One disadvantage of the methods lies in their computational complexity. There are many factors contributing to it, and the most significant one is the exponential growth of the size of the arrays with respect to the sequence length. This inevitably restricts the applicability of these methods to small problems. Nevertheless, we have seen that, within the scope of small problems, the flexibility of our methods can lead to much better results than the traditional methods subject to fixed functional forms.

Here we applied the method to aneuploidy data, but the method can also be applied to other types of ordinal data in genetics, such as variations in gene copy number [49] and genotypes consisting of alleles ordered by their phenotypic effect [50]. In addition to naturally discrete data, biologists frequently convert inherently continuous data into discrete, ordered categories to standardize measurements across multiple sources and alleviate measurement noise. One notable example is the Relevé data from ecology [51], which is often used to describe the composition, structure, and distribution of plant species in an ecological community. When compiling Relevé data, a discrete scale (e.g., the Braun-Blanquet scale) is typically used to describe the abundance of a list of plant species at different locations within a community. Our method can thus be applied to Relevé data to model the covariation between species and reveal key ecological interactions.

Finally, we conclude with an issue which we believe is worth further investigation. In our previous work of density estimation on categorical sequence data [14], the sequence space was represented by a Hamming graph, namely, a Cartesian product of complete graphs. We found that there existed a simple and elegant expression of the smoothness operator $\Delta^{(P)}$ in terms of the Laplacian matrix of the Hamming graph and its eigenvalues. Specifically, for categorical sequences of length ℓ and α elements, $\Delta^{(P)}$ can be expressed as

$$\Delta^{(P)} = \frac{1}{P!}\prod_{k=0}^{P-1}(L - k\alpha I), \tag{14}$$

where L is the Laplacian matrix of the graph, I is the identity matrix, and $k\alpha$, for $k = 0, 1, 2, \ldots, \ell$, are the eigenvalues of L. This operator can compute associations for all possible combinations of P different sites, as in Eq. (6). This expression not only suggests a more efficient way to implement the computation of associations but also gives us an insight into the sequence space decomposition. Unfortunately, when developing the theory of this work, we found that this expression no longer worked out, and we had to stick with a raw

definition of $\Delta^{(P)}$ as in Eq. (4). The failure of Eq. (14) for grid graphs can be easily seen by trying with a toy model. But why does it not work out? Does the expression exist only for Hamming graphs? We think it is worthy to investigate if a decomposition similar to Eq. (14), possibly with different entities or even in a different mathematical form, exists for other composite graphs. If not, what properties disable a graph from having such a decomposition? We believe answering these questions will be of interest to graph theory in general and sequence data analysis in particular.

Our implementation of ordinal SeqDEFT is available at Ref. [52] and the TCGA data set of aneuploidy in human cancer is from the Supplemental Material of Ref. [34].

ACKNOWLEDGMENTS

W.C.C. acknowledges support from the National Science and Technology Council of Taiwan, R.O.C., under Grant No. NSTC 111-2112-M-194-008-MY3 and additional funding from National Chung Cheng University. D.M.M. acknowledges support from NIH Grant No. R35GM133613, an Alfred P. Sloan research fellowship, and additional support from the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory.

APPENDIX A

Here we derive the posterior action $S_a[\phi]$ given in Eq. (9). We follow the steps outlined in Ref. [20], which is devoted to density estimation for one-dimensional continuous variables. Our derivation here can be regarded as a discrete version of the derivation therein.

For the sake of simplicity, we rewrite the parametrization of the probability distribution Q by the field ϕ, Eq. (1), as

$$Q_i = \frac{e^{-\phi_i}}{\sum_j e^{-\phi_j}}, \tag{A1}$$

where i and j, both running from 1 to α^ℓ, are index of the sequences. Observe that if ϕ is shifted by some constant, say, c, Q will still be the same. So we can write ϕ as $\phi = \psi + c\mathbf{1}$, where ψ is the part that changes from sequence to sequence, $\mathbf{1}$ is a vector of the right length full of ones, and c is an arbitrary constant. Then Eq. (A1) can be expressed as

$$Q_i = \frac{e^{-\psi_i}}{\sum_j e^{-\psi_j}}. \tag{A2}$$

Because of this "degeneracy," the prior and posterior distributions of Q and ϕ are related by

$$p(Q|a) = \int_{-\infty}^{\infty} dc\; p(\phi|a) \tag{A3}$$

and

$$p(Q|\text{data}, a) = \int_{-\infty}^{\infty} dc\; p(\phi|\text{data}, a), \tag{A4}$$

respectively. Our goal is to find out $p(\phi|\text{data}, a)$, and we achieve this by first computing $p(Q|\text{data}, a)$ and then using Eq. (A4) to figure out $p(\phi|\text{data}, a)$.

The posterior distribution of Q can be written as

$$p(Q|\text{data}, a) = \frac{p(\text{data}, Q|a)}{p(\text{data}|a)}. \tag{A5}$$

In turn, the probability $p(\text{data}, Q|a)$ can be written as

$$p(\text{data}, Q|a) = p(\text{data}|Q)\, p(Q|a). \tag{A6}$$

Below we will compute $p(Q|a)$, $p(\text{data}|Q)$, and $p(\text{data}|a)$, and then put them together to obtain $p(Q|\text{data}, a)$.

1. Derivation of $p(Q|a)$

We adopt the prior distribution on ϕ:

$$p(\phi|a) = \frac{1}{Z_a^0}\, e^{-S_a^0[\phi]} \tag{A7}$$

with

$$Z_a^0 = \int \mathcal{D}\phi\; e^{-S_a^0[\phi]}. \tag{A8}$$

The prior action $S_a^0[\phi]$ is given by

$$S_a^0[\phi] = \frac{a}{2s}\, \phi^T \Delta^{(P)} \phi. \tag{A9}$$

The operator $\Delta^{(P)}$ has a kernel and so we "regularize" it as follows:

$$S_a^0[\phi] \to \frac{1}{2}\, \phi^T \left(\frac{a}{s} \Delta^{(P)} + \epsilon \right) \phi \tag{A10}$$

with $\epsilon \to 0$. Let $\phi = \psi + c\mathbf{1}$. We have

$$S_a^0[\phi] = S_a^0[\psi] + \epsilon\beta c + \tfrac{1}{2}\epsilon\alpha^\ell c^2, \tag{A11}$$

where $\beta = \sum_i \psi_i$. Then the prior distribution of Q, given in Eq. (A3), can be found to be

$$p(Q|a) = \sqrt{\frac{2\pi}{\epsilon\alpha^\ell}}\, e^{\frac{\epsilon\beta^2}{2\alpha^\ell}} \times \frac{1}{Z_a^0}\, e^{-S_a^0[\psi]}. \tag{A12}$$

The coefficients in the front came from the Gaussian integral with respect to c.

2. Derivation of $p(\text{data}|Q)$

The term $p(\text{data}|Q)$ is the data likelihood and is equal to

$$p(\text{data}|Q) = \prod_n Q_n, \tag{A13}$$

where $n = 1, 2, \ldots, N$, is the index of the sequences in the data set. Using Eq. (A2), we have

$$p(\text{data}|Q) = \frac{1}{\gamma^N}\, e^{-\sum_n \psi_n}, \tag{A14}$$

where $\gamma = \sum_j e^{-\psi_j}$. Let us look at the exponential term first. Note that ψ_n can be expressed as $\psi_n = \sum_i \delta_{ni}\psi_i$, where δ_{ni} is the Kronecker delta. Thus, the exponent of the term can be rewritten as

$$\sum_n \psi_n = \sum_n \sum_i \delta_{ni}\psi_i = N \sum_i R_i \psi_i, \tag{A15}$$

where $R_i = N^{-1} \sum_n \delta_{ni}$ is the observed frequency of the ith sequence. On the other hand, the factor in Eq. (A14) can be

expressed as

$$\frac{1}{\gamma^N} = \frac{N^N}{\Gamma(N)} \int_{-\infty}^{\infty} dc\, e^{-N(c+\gamma e^{-c})}. \tag{A16}$$

Using the definition of γ, we obtain

$$\frac{1}{\gamma^N} = \frac{N^N}{\Gamma(N)} \int_{-\infty}^{\infty} dc\, e^{-N(c+\sum_i e^{-\phi_i})}. \tag{A17}$$

Putting Eqs. (A15) and (A17) into Eq. (A14), we have

$$p(\text{data}|Q) = \frac{N^N}{\Gamma(N)} \int_{-\infty}^{\infty} dc\, e^{-N\sum_i (R_i\phi_i + e^{-\phi_i})}. \tag{A18}$$

Note that we multiplied the constant c in the exponent of Eq. (A17) by $\sum_i R_i$, which is equal to 1, before we combined it with Eq. (A15).

3. Derivation of $p(\text{data}|a)$

Now that we have derived $p(Q|a)$, Eq. (A12), and $p(\text{data}|Q)$, Eq. (A18), we can multiply them together to obtain $p(\text{data}, Q|a)$, and the result is

$$p(\text{data}, Q|a) = \frac{N^N}{\Gamma(N)} \sqrt{\frac{2\pi}{\epsilon\alpha^\ell}}\, e^{\frac{\epsilon\beta^2}{2\alpha^\ell}} \frac{1}{Z_a^0} \int_{-\infty}^{\infty} dc\, e^{-S_a[\phi]} \tag{A19}$$

with

$$S_a[\phi] = S_a^0[\phi] + N\sum_i R_i\phi_i + N\sum_i e^{-\phi_i}. \tag{A20}$$

Here we have used the fact that $S_a^0[\psi] = S_a^0[\phi]$ when taking the limit $\epsilon \to 0$. See Eq. (A11).

The probability $p(\text{data}, Q|a)$ can also be written as

$$p(\text{data}, Q|a) = \int_{-\infty}^{\infty} dc\, p(\text{data}, \phi|a). \tag{A21}$$

Comparing this equation with Eq. (A19), we obtain

$$p(\text{data}, \phi|a) = \frac{N^N}{\Gamma(N)} \sqrt{\frac{2\pi}{\epsilon\alpha^\ell}}\, e^{\frac{\epsilon\beta^2}{2\alpha^\ell}} \frac{1}{Z_a^0}\, e^{-S_a[\phi]}. \tag{A22}$$

Therefore, the probability $p(\text{data}|a)$, which is given by

$$p(\text{data}|a) = \int \mathcal{D}\phi\; p(\text{data}, \phi|a), \tag{A23}$$

is equal to

$$p(\text{data}|a) = \frac{N^N}{\Gamma(N)} \sqrt{\frac{2\pi}{\epsilon\alpha^\ell}}\, e^{\frac{\epsilon\beta^2}{2\alpha^\ell}} \frac{Z_a}{Z_a^0} \tag{A24}$$

with

$$Z_a = \int \mathcal{D}\phi\, e^{-S_a[\phi]}. \tag{A25}$$

4. Putting things together

Finally, with Eqs. (A19) and (A24), the posterior distribution $p(Q|\text{data}, a)$, Eq. (A5), can be expressed as

$$p(Q|\text{data}, a) = \frac{1}{Z_a} \int_{-\infty}^{\infty} dc\, e^{-S_a[\phi]}. \tag{A26}$$

Comparing this equation with Eq. (A4), we obtain the posterior distribution of ϕ:

$$p(\phi|\text{data}, a) = \frac{1}{Z_a}\, e^{-S_a[\phi]} \tag{A27}$$

with the posterior action $S_a[\phi]$ given in Eq. (A20).

APPENDIX B

Here we provide an exposition of Eq. (13). The standard Kimura probability of fixation of an individual mutant gene for a haploid Wright-Fisher process is given by [53]

$$\frac{1 - e^{-2s}}{1 - e^{-2Ns}}, \tag{B1}$$

where N is the population size and s is the selective advantage. For small s, the probability of fixation can be approximated as [54,55]

$$\frac{2s}{1 - e^{-2Ns}}. \tag{B2}$$

If each accessible mutation occurs at rate 1, then mutation enters a population of size N at rate N, and hence the rate of introduction of mutations destined for fixation is

$$\frac{2Ns}{1 - e^{-2Ns}}, \tag{B3}$$

which depends only on the compound parameter $2Ns$. Thus, if genotype i has fitness f_i and genotype j has fitness f_j, the rate from i to j is

$$\frac{2Nf_j - 2Nf_i}{1 - e^{-(2Nf_j - 2Nf_i)}}, \tag{B4}$$

and it is easy to verify that the resulting Markov chain satisfies detailed balance with stationary distribution

$$\pi_i \propto e^{2Nf_i}. \tag{B5}$$

Therefore, if we set $2Nf_i = \log Q_i^*$, we obtain the rate matrix given in Eq. (13) and produce a Markov chain with stationary distribution $\pi_i = Q_i^*$.

[1] F. Sanger, S. Nicklen, and A. R. Coulson, DNA sequencing with chain-terminating inhibitors, Proc. Natl. Acad. Sci. USA **74**, 5463 (1977).

[2] J. Shendure, S. Balasubramanian, G. M. Church, W. Gilbert, J. Rogers, J. A. Schloss, and R. H. Waterston, DNA sequencing at 40: Past, present and future, Nature (London) **550**, 345 (2017).

[3] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, Genome-wide mapping of *in vivo* protein-DNA interactions, Science **316**, 1497 (2007).

[4] N. Cloonan *et al.*, Stem cell transcriptome profiling via massive-scale mRNA sequencing, Nat. Methods **5**, 613 (2008).

[5] R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker, Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*, Cell **133**, 523 (2008).

[6] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-seq, Nat. Methods **5**, 621 (2008).

[7] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, The transcriptional landscape of the yeast genome defined by RNA sequencing, Science **320**, 1344 (2008).

[8] B. T. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. J. Penkett, J. Rogers, and Jürg Bähler, Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution, Nature (London) **453**, 1239 (2008).

[9] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, Cambridge, England, 1998).

[10] E. Schneidman, M. J. Berry, II, R. Segev, and W. Bialek, Weak pairwise correlations imply strongly correlated network states in a neural population, Nature (London) **440**, 1007 (2006).

[11] J. Humplik and G. Tkačik, Probabilistic models for neural populations that naturally capture global coupling and criticality, PLoS Comput. Biol. **13**, e1005763 (2017).

[12] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Inverse statistical physics of protein sequences: A key issues review, Rep. Prog. Phys. **81**, 032601 (2018).

[13] A. J. Riesselman, J. B. Ingraham, and D. S. Marks, Deep generative models of genetic variation capture the effects of mutations, Nat. Methods **15**, 816 (2018).

[14] W. C. Chen, J. Zhou, J. M. Sheltzer, J. B. Kinney, and D. M. McCandlish, Field-theoretic density estimation for biological sequence space with applications to 5′ splice site diversity and aneuploidy in cancer, Proc. Natl. Acad. Sci. USA **118**, e2025782118 (2021).

[15] W. Bialek, C. G. Callan, and S. P. Strong, Field theories for learning probability distributions, Phys. Rev. Lett. **77**, 4693 (1996).

[16] I. Nemenman and W. Bialek, Occam factors and model independent Bayesian learning of continuous distributions, Phys. Rev. E **65**, 026137 (2002).

[17] T. A. Enßlin, M. Frommert, and F. S. Kitaura, Information field theory for cosmological perturbation reconstruction and nonlinear signal analysis, Phys. Rev. D **80**, 105005 (2009).

[18] T. A. Enßlin, Information field theory, AIP Conf. Proc. **1553**, 184 (2013).

[19] J. B. Kinney, Estimation of probability densities using scale-free field theories, Phys. Rev. E **90**, 011301(R) (2014).

[20] J. B. Kinney, Unification of field theory and maximum entropy methods for learning probability densities, Phys. Rev. E **92**, 032107 (2015).

[21] W. C. Chen, A. Tareen, and J. B. Kinney, Density estimation on small data sets, Phys. Rev. Lett. **121**, 160605 (2018).

[22] R. Staden, Computer methods to locate signals in nucleic acid sequences, Nucleic Acids Res. **12**, 505 (1984).

[23] G. D. Stormo, Modeling the specificity of protein-DNA interactions, Quant. Biol. **1**, 115 (2013).

[24] A. S. Lapedes, B. Giraud, L. Liu, and G. D. Stormo, in *Statistics in Molecular Biology and Genetics*, edited by F. Seillier-Moiseiwitsch (Institute of Mathematical Statistics, 1999), pp. 236–256.

[25] G. Yeo and C. B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals, J. Comput. Biol. **11**, 377 (2004).

[26] W. Bialek and R. Ranganathan, Rediscovering the power of pairwise interactions, arXiv:0712.4397.

[27] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing, Proc. Natl. Acad. Sci. USA **106**, 67 (2009).

[28] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Jr., Maximum entropy models for antibody diversity, Proc. Natl. Acad. Sci. USA **107**, 5405 (2010).

[29] S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S. I. Lee, and C. J. Langmead, Learning generative models for protein fold families, Proteins **79**, 1061 (2011).

[30] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. USA **108**, E1293 (2011).

[31] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models, Phys. Rev. E **87**, 012707 (2013).

[32] E. van Nimwegen, Inferring contacting residues within and between proteins: What do the probabilities mean? PLoS Comput. Biol. **12**, e1004726 (2016).

[33] R. M. Levy, A. Haldane, and W. F. Flynn, Potts Hamiltonian models of protein co-variation, free energy landscape, and evolutionary fitness, Curr. Opin. Struct. Biol. **43**, 55 (2017).

[34] A. M. Taylor *et al.*, Genomic and functional approaches to understanding cancer aneuploidy, Cancer Cell **33**, 676 (2018).

[35] B. Nogrady, How cancer genomics is transforming diagnosis and treatment, Nature (London) **579**, S10 (2020).

[36] B. T. Whitfield and J. T. Huse, Classification of adult-type diffuse gliomas: Impact of the World Health Organization 2021 update, Brain Pathol. **32**, e13062 (2022).

[37] C. Hutter and J. C. Zenklusen, The Cancer Genome Atlas: Creating lasting value beyond its data, Cell **173**, 283 (2018).

[38] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivaruate Analysis: Theory and Practice* (MIT Press, Cambridge, MA, 1975).

[39] T. M. Cover and J. A. Thomas, in *Elements of Information Theory*, 2nd ed. (John Wiley & Sons, Hoboken, NJ, 2006), pp. 409–425.

[40] A. Vasudevan, K. M. Schukken, E. L. Sausville, V. Girish, O. A. Adebambo, J. M. Sheltzer, Aneuploidy as a promoter and suppressor of malignant growth, Nat. Rev. Cancer **21**, 89 (2021).

[41] The Cancer Genome Atlas Research Network, Comprehensive genomic characterization defines human glioblastoma genes and core pathways, Nature (London) **455**, 1061 (2008).

[42] The Cancer Genome Atlas Research Network, The somatic genomic landscape of glioblastoma, Cell **155**, 462 (2013).
[43] The Cancer Genome Atlas Research Network, Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas, N. Engl. J. Med. **372**, 2481 (2015).
[44] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, Circos: An information aesthetic for comparative genomics, Genome Res. **19**, 1639 (2009).
[45] A. Tareen and J. B. Kinney, Logomaker: Beautiful sequence logos in Python, Bioinformatics **36**, 2272 (2020).
[46] C. Geisenberger *et al.*, Molecular profiling of long-term survivors identifies a subgroup of glioblastoma characterized by chromosome 19/20 co-gain, Acta Neuropathol. **130**, 419 (2015).
[47] A. Cohen, M. Sato, K. Aldape, C. C. Mason, K. Alfaro-Munoz, L. Heathcock, S. T. South, L. M. Abegglen, J. D. Schiffman and H. Colman, DNA copy number analysis of Grade II–III and Grade IV gliomas reveals differences in molecular ontogeny including chromothripsis associated with IDH mutation status, Acta Neuropathol. Commun. **3**, 34 (2015).
[48] D. M. McCandlish, Visualizing fitness landscapes, Evolution **65**, 1544 (2011).
[49] S. H. Almal and H. Padh, Implications of gene copy-number variation in health and diseases, J. Hum. Genet. **57**, 6 (2012).
[50] G. Romeo and V. A. McKusick, Phenotypic diversity, allelic series and modifier genes, Nat. Genet. **7**, 451 (1994).
[51] J. Podani, Braun-Blanquet's legacy and data analysis in vegetation science, J. Veg. Sci. **17**, 113 (2006).
[52] W. C. Chen, J. Zhou, and D. M. McCandlish, OrdinalSeqDEFT (2024), GitHub repository, https://github.com/wcchen-ccu/OrdinalSeqDEFT.
[53] M. Kimura, On the probability of fixation of mutant genes in a population, Genetics **47**, 713 (1962).
[54] R. A. Fisher, *The Genetical Theory of Natural Selection* (Oxford University Press, Oxford, England, 1930).
[55] S. Wright, Evolution in Mendelian populations, Genetics **16**, 97 (1931).