Nonparametric Combinatorial Sequence Models

Fabian L. Wauthier, Michael I. Jordan, and Nebojsa Jojic

[1] University of California, Berkeley, flw@cs.berkeley.edu
[2] University of California, Berkeley, jordan@cs.berkeley.edu
[3] Microsoft Research, Redmond, jojic@microsoft.com

Abstract. This work considers biological sequences that exhibit combinatorial structures in their composition: groups of positions of the aligned sequences are "linked" and covary as one unit across sequences. If multiple such groups exist, complex interactions can emerge between them. Sequences of this kind arise frequently in biology but methodologies for analyzing them are still being developed. This paper presents a nonparametric prior on sequences which allows combinatorial structures to emerge and which induces a posterior distribution over factorized sequence representations. We carry out experiments on three sequence datasets which indicate that combinatorial structures are indeed present and that combinatorial sequence models can more succinctly describe them than simpler mixture models. We conclude with an application to MHC binding prediction which highlights the utility of the posterior distribution induced by the prior. By integrating out the posterior our method compares favorably to leading binding predictors.

Keywords: Sequence models, Chinese restaurant process, Chinese restaurant franchise, MHC binding, mixture models

1 Introduction

Proteins and nucleic acids, polymers whose primary structure can be described by a linear sequence of letters, are found in nature in an astounding diversity. Understanding the diversity of biological sequences has been a major topic in computational biology. Through inheritance, and close functional coupling, the nearby sequence positions in a family of biological sequences are often at a linkage disequilibrium, i.e., the letters at nearby sites tend to covary. However, in their folded form, these molecules also have secondary, tertiary, and quaternary structure, which may reveal geometric proximity, and provide a basis for potential interactions of residues at distant sequence sites and even across different molecules. This creates significant difficulties in modeling diversity of certain families of sequences, where both the nearby and distant sequence positions may exhibit patterns of covariation. This difficulty is exacerbated by the fact that with only a limited number of sequences available for analysis we could arrive at multiple diversity models which are almost equally well supported by data. We model such sequence data starting with a basic componential strategy outlined in Figure 1. We show four aligned subsequences from Influenza HA1

(a) Observations (b) Profiles

Fig. 1. 1(a) Four aligned short subsections of the sequences exhibiting the combinatorial pattern according to the partition highlighted by color. The blue component, $z_{\text{site}} = 1$, comes in two variants, `TGCATC` and `CATGAT`, while the green component, $z_{\text{site}} = 2$, follows either `ACA` or `CTG`. All four combinations of types of these segments are found in the data. Each of those configurations can be combined with two further variants, `GGG` and `AAA`, in the red component, $z_{\text{site}} = 3$. 1(b) Slight perturbations on the basic types are possible as captured by the profiles inferred whose appropriate sections are shown. The profiles and subsequences correspond to appropriate sections of the Influenza HA1 genes analyzed in Section 4. The sequence sites switch among profiles in groups—the entire component follows one of the three profiles. (In general, some components may be less entropic than others and the sequences may then not be mapped to all three different types.) The four sequences in this example can be represented by the pointers z_{prof} for each of the three components which map the components to the appropriate profiles: 213, 113, 223, and 222. Such compression of the variability can increase statistical power of techniques mapping genotypic and phenotypic variation as we demonstrate for the case of MHC binding prediction in Section 4.

genes whose diversity is well explained by first partitioning the sites into three groups and then representing each partition's induced subsequences by one of several prototypes. The site groupings do not need to follow linear patterns, and distant sites may be grouped together. Assuming that the three types in the three groups can be arbitrarily mixed, the model represents 27 different variants, and could thus also be expressed as a mixture with that many components. However, the use of a traditional mixture model would require considerably more data for training, as having obtained only 50–100 sequences it is likely that we did not see all 27 combinations. On the other hand, it is likely that we observed all three types in all three components multiple times, thus facilitating parameter estimation in a componential model. Furthermore, the componential structure itself may be of importance. If for instance, a phenotype of interest is linked only to one variant of one of the components, then the mixture model would capture this variant in nine components needed to represent the relevant type in combination with three types in each of the other two components. Thus a traditional clustering would lead to nine different statistical tests, lowering statistical power by an order of magnitude. In this sense, the combinatorial structure allows for pooling the traditional mixture components based on the finer-grained patterns of covariation. In this paper we outline a probabilistic model that can be used to discover such structure in several gene and protein families while coping with

Family	Forces shaping componential diversity
Immunoglobulin/TCR	Clonal V(D)J recombination
Pathogenic proteins	Recombination, mutation
MHC/KIR	Large and small scale recombination, mutation

Table 1. Sequence families exhibiting combinatorial structures.

the dearth of sequence data and the possible additional correlations among the groups. Such combinatorial diversity is ubiquitous at larger scales such as entire chromosomes. However, some very important biomolecules have relatively short segments that are under significant diversifying selection. In Table 1, we highlight molecules involved in host-pathogen interactions and whose subsequences fit the model discussed above. All these families of molecules have to maintain their biological function, while exhibiting a high degree of variation concentrated in a short subsequence, and the solution to these conflicting requirements has componential structure.

As the first example, we point to the genes encoding immunoglobulin and T cell receptor proteins which are split into multiple gene segments in the germline. These segments are made contiguous by recombination in somatic tissues by the well known V(D)J recombination process [3]. To assemble an antigen receptor gene, one V (variable), one J (joining) and, sometimes, one D (diversity) segment are joined to create an exon that encodes the binding portion of the receptor chain. As there are typically many V, D, and J gene segments, V(D)J recombination creates an immense combinatorial diversity of antibody and TCR binding specificities, responding to the diversity of the immune system's targets.

Pathogen proteins whose subsequences are often targets of immunoglobulin and TCR binding also exhibit combinatorial diversity. For instance, VAR2CSA, a member of the *P. falciparum* erythrocyte membrane 1 protein family and a potential vaccine candidate for pregnancy-associated malaria, contains short segments in which the isolate variation can be well summarized by a small number of very different types. While human-infecting *P. falciparum* isolates exhibit combinatorial diversity resulting from fairly arbitrary mixing of segment types, each type is remarkably conserved across isolates that have them, including isolates of *P. reichenowi* which infects other primates [2]. This indicates a possible role of recombination with other V gene segments in creating combinatorial diversity in the binding domains of these proteins, which have to facilitate adhesion to the placenta while avoiding recognition by the immune system.

The third example we highlight is the major histocompatibility complex (MHC) class I family of molecules which again participate in the interaction between the host immune system and pathogens. In virtually all cells of higher organisms, these molecules present antigenic cellular peptides on the cellular surface for surveillance by cytotoxic T cells. The T cell receptor proteins discussed above may bind to the complex made of the MHC molecule and the antigenic peptide which can lead to the destruction of the infected cell. To properly facilitate the surveillance of the cellular proteome, MHC molecules are again faced

with complex requirements: Across different situations, the MHC molecules will encounter a large number of different targets that may need to be carried to the surface, but at any given time, the cellular presentation should be limited to useful targets. Furthermore, as pathogens adapt to immune pressure quickly, a population of hosts is more resilient if it is diverse in its immune surveillance properties. Nature's solution here is somewhat different than in the case of the TCR and immunoglobulin. The immune system needs to learn to tolerate normal self proteins and the variation in binding properties through clonal recombination in one individual would complicate this tolerance. Instead, in humans, three highly diverse loci encode for MHC class I, leading to diversity of MHC binding specificities across individuals, not within one host. The residues forming the peptide binding groove of the MHC molecules have been found to be under a diversifying selection. The statistical study of MHC alleles has yielded evidence of both large-scale recombination events (involving entire exons) and low-scale recombination events (involving apparent exchange of short DNA segments), but convergent evolution in parts of the MHC from different alleles is also supported by the data [5]. Thus, a variety of mutation and recombination events, whose combinations were selected based on the resulting binding properties of the MHC groove lead to the immense diversity at this locus, the most polymorphic in the human genome.

In these three examples, and many more (Figure 1 illustrates diversity in an influenza protein), the functional requirements have created sequence families that exhibit high levels of diversity with combinatorial structure similar to the one illustrated in Figure 1. Models that capture such structure have immediate applications in low-level tasks such as sequencing, haplotype recovery, as well as in higher level tasks involving the matching of the genetic diversity to phenotypic variation. In the case of the immunoglobulin, this structure is essentially encoded in the human genome, and the different V, D, and J variants can be directly read off there. But, when diversity is maintained on a population level, as is the case with most pathogen proteins and RNA molecules, as well as MHC or KIR (receptor on natural killer cells) among human proteins, then we can only recover the structure by analyzing sequences from a number of individuals. This is complicated by two effects: first, the illustration in Figure 1 is a simplification. The groups of sites are only approximately independent of each other. Some residual weak linkage is expected to exist even in the case of the optimal sequence partition. Secondly, due to the high polymorphism in the families of interest the structure in Figure 1 can only be estimated reliably when sufficient data is available. When data is scarce, multiple different solutions are possible that differ little in the data fit.

In this paper we propose a model that differs from existing models in the way it addresses these two issues. In [1, 2], the partition is assumed to consist of contiguous segments, a constraint that does not hold for many interesting diversity patterns (cf. Figure 1), and a single optimal segmentation (cf. the pattern library/epitome approach of [7, 8]). A combinatorial optimization algorithm for site clustering that does not promote contiguous segments is proposed in [12],

but, as the basic generative model creates blocks with limited diversity [6, 8], the result is again a single optimal segmentation which can be sensitive to the size of the sequence set used to estimate it.

We propose a Bayesian hierarchical site clustering approach with a minimal number of parameters which not only captures weak linkage among components at the first level of the clustering hierarchy, but also naturally adjusts to the size of the dataset. Furthermore, we develop a sampling procedure that produces an estimate of the posterior over possible sequence partitions. In Section 4 we illustrate for three families of proteins—MHC class I, Influenza HA1 and KIR—that the componential model discussed here is a better fit than traditional mixture models, which cluster entire sequences (phylogenetic methods fall in this category). We also show an example where by using the distribution over multiple partitions we improve on the ability to match the genetic diversity with the phenotype variation. In particular, by representing MHC sequences by the latent variables in our model we train simple MHC class I–peptide binding estimators. We show that by integrating over possible MHC sequence representations based on different partitions we obtain better predictions than when we use the latent variables for the MAP estimate of the segmentation structure.

2 Model

Most approaches to capturing diversity in sets of aligned sequences treat each sequence as a whole, applying clustering techniques (e.g., neighbor-joining or maximum likelihood approaches) or building a hierarchical clustering of sequences (e.g., a phylogenetic tree). A special case of such approaches are mixture models which describe aligned sequences as being sampled from a mixture of a small number of "latent profiles," also known as "position-specific scoring matrices," e.g., [10]. As outlined above, a considerable drawback of a whole sequence mixture model is that each observed sequence corresponds in its entirety to one latent profile. Our model is a generalized mixture model that relaxes this constraint and allows different sequence positions to correspond to different profiles. To retain some structure, however, our model introduces a latent partitioning that groups site positions into linked sites that must be sampled from the same profile. Each such "site group" thus induces a different mixture model on its component sites. This allows us to capture combinatorial diversity that is not captured by a flat mixture model—n site groups with k profiles would need n^k mixed profiles if the data was to be represented by a flat mixture. Moreover, as discussed in Section 1, we wish to also couple the mixture models in order to capture additional weaker links among the site groups. Our model achieves this by implicitly coupling the mixing proportions of the different mixtures.

When analyzing data with traditional mixture models, one is faced with the perennial problem of choosing the number of mixture components. Since the model we are proposing can be thought of as a refined mixture model, it is not immune to this issue. While information-theoretic techniques do exist for estimating the structural parameters in mixture models, they are difficult to justify

when the number of components required to represent a large dataset is large [24]. In a number of biological applications [17, 22–24] nonparametric methods based on the Chinese restaurant process (CRP), or the closely related Dirichlet process, have been shown to elegantly circumvent such issues by effectively introducing a prior distribution on the number of latent components. A second advantage is that the induced prior automatically accommodates more latent components as the amount of data grows. This allows us to infer conservative representations with few components when little data is available while being flexible enough to represent complex patterns emerging from larger datasets.

Our model relies on a composition of two nonparametric priors—the Chinese restaurant process (CRP) [16] and the related Chinese restaurant franchise (CRF) [21]. By incorporating these two nonparametric priors we circumvent fixing the number of site groups and the number of profile variants a priori, and instead average over these choices under a posterior distribution.

In this section we present our model by means of a sequential, generative description. In this description we use the index s to index sequences and i to index the sites (sequence positions) within a sequence. Let M denote an $S \times I$ matrix of aligned sequences, so that m_s denotes the s-th sequence and $m_{s,i}$ denotes the i-th symbol in the s-th sequence. Our model relies on four sets of latent random variables: $z_{\text{site}}, z_{\text{clust}}, z_{\text{prof}}$ and θ, sampled in top-down fashion according to a CRF that is conditioned on a partition sampled from a CRP.

2.1 Chinese Restaurant Process Linkage Model

The CRP [16] is a nonparametric prior on partitions of a set of items. In its generative form it describes a sequential process that produces a dataset exhibiting clusters. The language of the CRP likens the sequential process to a (potentially endless) stream of customers entering a restaurant one by one. Upon entering, each patron randomly chooses a table to sit at with probability proportional to the number of customers already seated there, or sits at an empty table. Each table is assigned a parameter that is shared by all customers at that table. For clustering, the datapoints are thought of as patrons, and the clusters as tables, which are parameterized by the tables' parameter.

The first step in our model is to sample a partition of the site indices into groups of linked sites. At this level of the model site indices are not yet associated with any data—we only use the CRP seating process to induce a site partitioning. The partition is sampled from a CRP where sites act as customers and site groups as tables. Representing the allocation of sites to groups (tables) by a set of latent variables $z_{\text{site}}(i), i = 1, \ldots, I$, the process operates as follows: Customers (site indices) enter the restaurant one by one and choose to sit either at an existing table or to open a new table. At each step of the sequential process, let the number of existing site tables be denoted by n_{site}, and the number of site indices at table t by $c_{\text{site}}(t), t = 1, \ldots, n_{\text{site}}$. If we parameterize the CRP by α_{site}, then the seating probabilities for site i given the seating assignment for all previous

sites $1, \ldots, i-1$ are given as

$$p(z_{\text{site}}(i) = t | z_{\text{site}}(1{:}i-1)) \propto \begin{cases} c_{\text{site}}(t) & \text{if } t \leq n_{\text{site}} \\ \alpha_{\text{site}} & \text{if } t = n_{\text{site}} + 1 \ . \end{cases} \tag{1}$$

From this definition we see that just as the number of sites visiting the restaurant can in principle be unbounded, so can the number of tables at which they sit. However, as the number of sites grows, it becomes less likely that new tables will be opened; indeed, the growth rate can be shown to be $O(\alpha_{\text{site}} \log i)$. Note the role of the parameter α_{site} in scaling this growth rate in the prior distribution.

In the following, a site group is treated as an inseparable entity which can be grouped further. In the overall process, it is the preliminary site grouping which captures most of the site linkage in the observed data.

2.2 Chinese Restaurant Franchise Observation Model

The second part of our model represents a combinatorial observation model over aligned sequences in the form of a CRF [21] that is conditioned on the initial partitioning z_{site} by the CRP. The CRF is a generalization of the CRP to allow multiple parallel restaurants to share parameters. Specifically, where in the CRP each table is given a parameter which is shared among its occupants, in the CRF these parameters can also be shared across multiple CRPs. It will turn out that the "parameters" that are being shared in our application are pointers to profiles, rather than the profiles themselves. As such, our model can be thought of as an instance of a dependent nonparametric process, discussed by MacEachern [11], where individual parameters are replaced by stochastic processes. In the CRF we interpret each observed sequence as its own restaurant. But instead of thinking of site positions as customers, as in a standard application of the CRF, we now consider the previously induced site groups to be customers. Each restaurant is visited by all site groups, so that the union of the site groups at each restaurant captures the entire set of sequence indices. The CRF is defined as follows. At each sequence m_s the n_{site} site groups indicated in z_{site} are seated at tables a second time according to the rules of a CRP. The seating arrangement of the site groups is represented by latent variables $z_{\text{clust}}(s,t), t = 1, \ldots, n_{\text{site}}$. Denote by $n_{\text{clust}}(s)$ the number of (second-level) tables formed at sequence s at each step of the process, and let $c_{\text{clust}}(s,u), u = 1, \ldots, n_{\text{clust}}(s)$ denote the number of site groups present at the table u. If we parameterize the sequential seating process at each restaurant by α_{clust}, then conditioned on the seating assignment of the site groups $1, \ldots, t-1$, the seating probabilities for group t are

$$p(z_{\text{clust}}(s,t) = u | z_{\text{clust}}(s,1{:}t-1)) \propto \begin{cases} c_{\text{clust}}(s,u) & \text{if } u \leq n_{\text{clust}}(s) \\ \alpha_{\text{clust}} & \text{if } u = n_{\text{clust}}(s) + 1 \ . \end{cases} \tag{2}$$

In order to produce observed sequences, the CRF model next introduces parameters. Each table u in a sequence restaurant s in the CRF is assigned a latent variable $z_{\text{prof}}(s,u)$, that indicates which of a set of shared parameters θ is used at table u of restaurant s. We will refer to one such shared parameter θ_p as

a "sequence profile." As before, at each step of the sequential algorithm, the variable n_{prof} denotes how many distinct profiles the set of z_{prof} variables points to. The function $c_{\text{prof}}(p), p = 1, \ldots, n_{\text{prof}}$ reports how many of the tables in all processed sequence restaurants picked profile p. In the sequential description of the CRF, the choice of profile made by each table is influenced by the number of other tables that have previously chosen that profile. That is, the process can be thought of as another CRP in which distinct profiles can be thought of as tables. If we use parameter α_{prof} to define this CRP, then the probability that table u in restaurant s chooses profile p, given the profile choices of all tables in restaurants $1, \ldots, s-1$ and tables $1, \ldots, u-1$ in restaurant s, is given by

$$p\left(z_{\text{prof}}(s,u)=p | z_{\text{prof}}(1{:}s-1,\cdot), z_{\text{prof}}(s,1{:}u-1)\right) \propto \begin{cases} c_{\text{prof}}(p) & \text{if } p \leq n_{\text{prof}} \\ \alpha_{\text{prof}} & \text{if } p = n_{\text{prof}}+1. \end{cases} \quad (3)$$

For sequences with an alphabet of size A, each sequence profile $\theta_p, p = 1, \ldots, n_{\text{prof}}$ is comprised of I A-vectors, one for each site index. Each vector $\theta_p(\cdot, i)$ is a probability distribution over the A possible symbols that could be observed at position i. When a new table in one of the restaurants chooses a new profile θ_p which has not yet been chosen before, the profile vectors $\theta_p(\cdot, i), i = 1, \ldots, I$ are sampled from a Dirichlet prior, parameterized by α_{dir}.

Once all latent variables and profiles have been sampled, the observed sequences are generated as follows: given the latent variables $z_{\text{site}}, z_{\text{clust}}, z_{\text{prof}}$ and profiles θ, we generate the symbol at position i in sequence s by sampling from a multinomial with parameter $\theta_p(\cdot, i)$, where $p = z_{\text{prof}}(s, z_{\text{clust}}(s, z_{\text{site}}(i)))$.

The sampling procedure generates data that exhibit the combinatorial structure discussed in Figure 1 and found in a variety of biological sequence families. Of course, our goal is to reverse this process. Starting from the observed sequences we need to reconstruct the latent variables $z_{\text{site}}, z_{\text{clust}}, z_{\text{prof}}$ and the profile sequences, while making explicit our uncertainty over these structures. In the next section we develop an inference algorithm that achieves this by approximating the full posterior over latent structures.

3 Inference

We use a collapsed Gibbs sampler in which the profiles θ are integrated out. The algorithm cycles through resampling the site grouping z_{site}, the secondary grouping of site groups z_{clust} and the assignment of profiles z_{prof}, at each step conditioning on all remaining latent variables. A central property of the CRP and CRF that facilitates this sampling process is *exchangeability*. Exchangeability allows us to treat any customer of a restaurant as if it were the last customer to enter the restaurant. This is consistent with our modeling assumption that sites have unique positions that need to be grouped, but that the ordering of these positions is of little value, since parts may be non-contiguous. The consequence of this exchangeability is that we can now easily sample an updated table seating for any customer in a restaurant. In the following we show the main computations for resampling the site grouping z_{site}. The posteriors for sampling updated variables z_{clust} and z_{prof} can be derived analogously to Teh et al. [21].

3.1 Resampling Site Groupings z_{site}

We denote by $z^{-i}_{\text{site}}, z^{-i}_{\text{clust}}$ and z^{-1}_{prof} the latent variables that remain when site i is removed from the representation. Let n^{-i}_{site} be the number of distinct site tables when site i is removed. Similarly, let $c^{-i}_{\text{site}}(t)$ be the number of site indices seated at table t when site i is removed. Due to the exchangeability of site indices in the top CRP, we may treat site i as if it were the last to enter the restaurant. In order to sample a new site grouping we must compute the probability that a particular site i is seated at a table, given all other relevant information:

$$p\left(z_{\text{site}}(i) = t | m_{\cdot,i}, z^{-i}_{\text{clust}}, z^{-i}_{\text{prof}}\right) \ . \tag{4}$$

Because we treat i as the last customer to enter the restaurant the prior probability of seating site i at table t is given by

$$p\left(z_{\text{site}}(i) = t | z^{-i}_{\text{site}}\right) \propto \begin{cases} c^{-i}_{\text{site}}(t) & \text{if } t \leq n^{-i}_{\text{site}} \\ \alpha_{\text{site}} & \text{if } t = n^{-i}_{\text{site}} + 1 \ . \end{cases} \tag{5}$$

In a collapsed sampler, if t is an existing site table then we compute the likelihood of seating site i at table t by integrating the induced conditional likelihood of sequence symbols at position i against the prior distributions on $\theta_p(\cdot, i), \forall p$. If we define for $z_{\text{site}}(i) = t \leq n^{-1}_{\text{site}}$ (an existing table was chosen) the count that a symbol at position i is of type a and is generated by profile p as

$$c_t(a, p) = \sum_s \mathbf{1}(m_{s,i} = a, z_{\text{prof}}(s, z_{\text{clust}}(s, t)) = p), \tag{6}$$

then for $t \leq n^{-1}_{\text{site}}$ the integrated likelihood of the observed sequence symbols in position i can be computed as

$$p(m_{\cdot,i} | z_{\text{site}}(i) = t, z^{-i}_{\text{clust}}, z^{-i}_{\text{prof}}) = \prod_p \frac{\Gamma(\sum_a \alpha_{\text{dir}}(a))}{\prod_a \Gamma(\alpha_{\text{dir}}(a))} \frac{\prod_a \Gamma(\alpha_{\text{dir}}(a) + c_t(a, p))}{\Gamma(\sum_a \alpha_{\text{dir}}(a) + c_t(a, p))}. \tag{7}$$

It is more complicated to compute the likelihood that site index i is seated at a new table $t = n^{-1}_{\text{site}} + 1$ since the creation of a new site index table triggers a cascade of other choices that need to be made for the z_{clust} and z_{prof} variables. In computing the likelihood of a new site table, the parameters $\theta_p(\cdot, i)$, as well as these new choices need to be integrated out. Rather than computing this complicated integral, we adopt a simpler strategy and approximate the likelihood by sampling a set of new assignments for $z_{\text{clust}}(s, n^{-1}_{\text{site}} + 1), s = 1, \ldots, S$ and if necessary also $z_{\text{prof}}(s, z_{\text{clust}}(s, n^{-1}_{\text{site}} + 1)), s = 1, \ldots, S$ by following the sequential generative model outlined before. Once sample allocations have been generated for the proposal that $t = n^{-1}_{\text{site}} + 1$, we can compute the integrated likelihood of the seating proposal by similar means as in equation (7), giving us the last term

$$p(m_{\cdot,i} | z_{\text{site}}(i) = n_{\text{site}} + 1, z^{-i}_{\text{clust}}, z^{-i}_{\text{prof}}) \ . \tag{8}$$

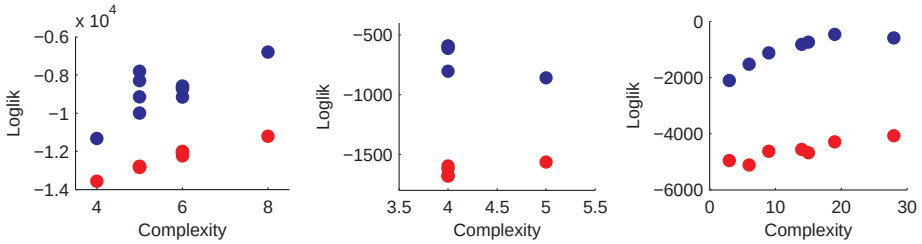

Fig. 2. Comparison of the log likelihood assigned by the combinatorial model (blue scatter) with the log likelihood assigned by a mixture model of comparable complexity (red scatter) as a function of different model complexities. From left to right are shown results for MHC, Flu and KIR sequences. Flu sequences require only relatively few profiles and site groups (cf. Figure 1); thus only two model complexities were explored.

Combining this likelihood with those computed in (7) and the prior in equation (5) allows us to compute the posterior in equation (4) from which we may now sample a new site group allocation for site index i. If an existing site group is chosen, nothing more needs to be done. If a new site group is created we copy the previously sampled allocations into the current state z_{clust} and z_{prof}.

3.2 Resampling z_{clust} and z_{prof}

Once the site partition z_{site} has been resampled, the resampling of z_{clust} and z_{prof} conditioned on z_{site} is performed in similar fashion as in the standard CRF. As before, our implementation integrates out the profile parameters to improve sampling efficiency. The computations can be readily derived from Teh et al. [21].

4 Results

To demonstrate the versatility of our model we applied it to three sequence datasets in which we expect combinatorial patterns to exist. In the following we have focused our analysis on a small number of the most polymorphic sites in each dataset. The first dataset are 526 aligned amino-acid sequences of length 50 for MHC class I proteins from all three alleles A, B, C. The flu dataset comprises aligned 22-long amino-acid sequences for 255 HA1 genes in influenza strains covering the years 1968–2003. The KIR dataset are sequences of unordered (i.e., unphased) pairs of haplotype measurements at 229 SNPs. These SNPs encode variability of a killer cell immunoglobulin-like receptor. If we knew the phase, we could order each pair and turn the data into aligned sequences that could easily be analyzed as outlined before. We have thus extended our model to work with unphased KIR data by introducing extra latent variables z_{phase} that encode the phasing information for each pair. The modified algorithm iterates between sampling phasing variables to turn aligned sequences of pairs into aligned sequences, and then sampling new latent variables $z_{\text{site}}, z_{\text{clust}}$, as well as z_{prof}, as before.

We have carried out experiments to demonstrate that our model successfully isolates combinatorial structures from the data and learns a much more parsimonious sequence model that yields higher log likelihood than a comparable mixture model. For each of the datasets we set up a combinatorial sequence model and computed posterior samples for different settings of the model parameters α_{site}, α_{cluster}, α_{profile}, and α_{dir}. Each combination of parameters induces a different nonparametric prior over aligned sequences. We wish to compare the average data likelihoods assigned by the posterior combinatorial models to the likelihoods obtained from flat mixture models. To facilitate this comparison, we ensure that the mixture models we compare against have similar complexity as the nonparametric model. We estimate the complexity of a given model by measuring how many parameters it would take to represent a set of sequences in a typical posterior sample. If for example a sample with n_{site} site tables and n_{prof} profiles of length I with a symbol alphabet of size A was found, we require a total of $I(n_{\text{site}}-1)+n_{\text{prof}}I(A-1)$ parameters as a shared representation across all sequences. The first $I(n_{\text{site}}-1)$ parameters encode which site position is allocated to which site group while the remaining account for the profile parameters. In comparison, a mixture model that links all site positions asserts that $n'_{\text{site}}=1$ and for n'_{prof} profiles requires $n'_{\text{prof}}I(A-1)$ parameters. Assuming that a single set of such parameters is fixed, to encode a set of sequences we would need to also infer for each sequence the posterior distributions over latent variables (mixture components for the mixture model, or profile pointers z_{prof} in our model). Then any remaining uncertainty as to the identity of the letters in individual positions would also have to be collapsed by encoding these individual letters. Information theory prescribes techniques for making the minimum required code length for encoding all this directly dependent on the uncertainties in the data, with less uncertain pieces encoded with shorter messages, so that the total code length in bits reduces to the $\log_2$ likelihood under the model [18]. By adding the cost of encoding the parameters that are shared by the sequences (profiles, partitioning information), we would obtain a description length of the dataset. The cost of encoding parameters would be proportional to the number of the parameters. Similarly, for comparing model fits, statistical literature recommends the use of the Bayesian information criterion (BIC) [19] or the Akaike information criterion (AIC) which combine the log likelihood of the data with a penalty reflecting the number of free parameters. However, rather than comparing the two models by an MDL, BIC or AIC score for only one model complexity, we present a stronger argument here: it turns out that for a wide range of model complexities, the log likelihood of the data is higher under the combinatorial model than under the mixture model.

To show this, for posterior samples of varying complexity under our model, we compute the smallest number of mixture profiles that would exceed it in complexity, i.e., n'_{prof} so that $I(n_{\text{site}}-1)+n_{\text{prof}}I(A-1)\leq n'_{\text{prof}}I(A-1)$. We then fit five mixture models on the data using n'_{prof} profiles and compute the average log likelihood assigned to the data. In Figure 2 we show for the three datasets the average log likelihood of the combinatorial model across samples

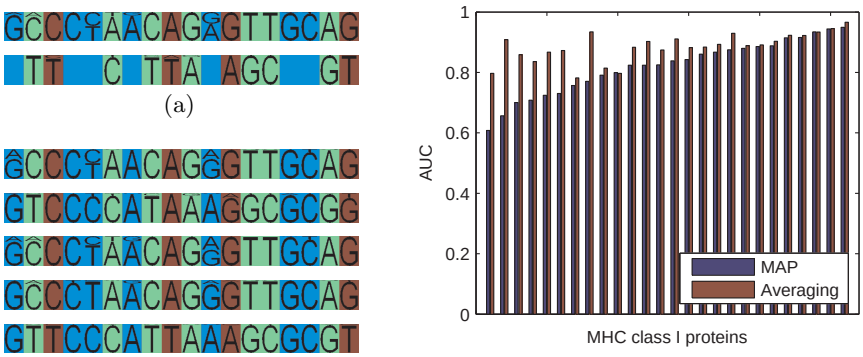

Fig. 3. (a) Factorized representation for the first 18 SNPs inferred by our model on the KIR data. Empty fields in the profiles denote that no further variants were found for a site group. (b) The 5 profiles for the first 18 SNPs learnt by a mixture model on KIR data. (c) AUC scores for the MHC I binding prediction task across 26 MHC proteins. Averaging regression results across posterior samples significantly improves the AUC score over using only the MAP sample to fit a regression.

as a blue scatter and the average log likelihood of the mixture model as a red scatter. For all three datasets, the log likelihood of the combinatorial model exceeds that of the mixture model considerably. Additionally, our model provides a better representation for sequence clustering. The clustering induced by our combinatorial model for Influenza HA1 sequences matches the hemagglutinin inhibition clusters of Smith et al. [20] closer than the clusters obtained by simple mixture modeling, achieving an average adjusted rand index [4] of 0.70 versus 0.55[4]. In Figure 3(a), we visualize profiles as well as site groups for the first 18 SNPs of the KIR dataset for one posterior sample of the combinatorial model. Figure 3(b) shows relevant parts of the 5 profiles that were inferred by a simpler mixture phasing model. As can be seen, our model factorizes the profiles inferred by the simpler mixture model into a parsimonious form that can still explain the mixture variants. The green group has variants `CACGTTA` and `TCTAGCG`, while the red group follows either `CAGG` or `TTAT`. Three of the four possible combinations of these patterns occur in the profiles estimated by the mixture model. As a side effect of the compact representation, our model allows for a more careful use of data for profile parameter inference. Mixture models can capture many combinations, but they achieve this by using a substantially greater number of parameters, while still missing many of the combinations outside the region shown. This leads to significantly lower likelihood in comparison with the componential model of similar parametric complexity, as shown in Figure 2.

[4] To compute these scores we encoded the sampled latent state of each sequence as a binary vector and clustered these into the same number of clusters as the target clustering. The results were averaged over many samples from the posterior.

4.1 MHC Class I Binding Prediction

The latent structure inferred under the model fit to MHC class I sequences above can be used to match these sequences to their binding affinities, and in this way predict epitopes for different MHC molecules. We model the binding affinity (measured in terms of the log IC50 concentration) of an MHC class I protein to an epitope as a linear function that allows sharing across several related protein variants. For any particular protein, our sequence model produces a combinatorial representation in terms of site groups and their associated profiles[5]. For a given set of M MHC proteins, we encode this latent structure in binary vectors $b_s, s = 1, \ldots M$. This structure compresses the links produced by co-evolution of the specific sites in the MHC groove. Assuming that some of this co-evolution is driven by selection for particular binding specificity patterns, the latent structure under our model is expected to be useful in binding prediction tasks. For each protein s, a given set of n_s epitopes examples is encoded as binary vectors $e_{sj}, j = 1 \ldots, n_s$. If we denote the corresponding binding affinities as $y_{sj}, j = 1, \ldots, n_s$, then the linear regression we solve in terms of Θ is written as $y_{sj} = e_{sj}^{\top} \Theta b_s$. The sharing among related proteins is induced by the latent structure b_s. We evaluated two variants of this regression. The first variant uses only the MAP sample from our model posterior to produce a single encoding b_s, while the second fits one regression for each posterior sample (each inducing a different encoding b_s) and then averages the final prediction across samples. The two regression tasks were trained on a total of about 28000 binding affinities over 26 different human MHC molecules. Some MHC molecules were characterized by only a handful of binding measurements, while others were tested against over a thousand different peptides. The results in Figure 3(c) show the AUC score (averaged over five cross-validation runs) obtained from classification into binding and non-binding epitopes. Integration across latent structure significantly boosts the prediction accuracy. Averaged across the 26 MHC variants the averaged predictor yields an AUC score of 0.8846, while the MAP variant achieves a score of only 0.8197. Our result compares favorably with state of the art methods summarized in Peters et al. [14]. The reviewed methods achieve average AUCs of 0.8500 to 0.9146 on a subset of 21 of the 26 proteins for which our averaging method gives a mean of 0.8911. Importantly, the method of Nielsen et al. [13] uses carefully designed nonlinearities and separately known properties of amino-acids to produce improved prediction results. Other leading methods [9, 15] use further feature design or exploit the protein structure to boost prediction results. In contrast, even though we use a simple binary representation of epitopes and MHCs, we produce comparable results by virtue of a refined latent sharing structure which is integrated out.

[5] The parameters used for the combinatorial sequence model were $\alpha_{\text{site}} = 0.1, \alpha_{\text{clust}} = 5, \alpha_{\text{prof}} = 10, \alpha_{\text{dir}} = 0.5$. Posterior samples typically had 3 profiles and 10 site groups over sequences of length 34.

5 Conclusion

This paper presented a nonparametric combinatorial sequence prior that was found to be a good match for a wide range of sequence families. An important feature of the model is that it induces a posterior distribution over latent factorized representations. Our work on MHC binding prediction demonstrates that integrating out this distribution can be an important ingredient in inferences that follow the initial sequence analysis. One way to explain why averaging across predictors should be beneficial in the case of MHCs is to consider the potential for suboptimal parsing of the MHC groove. Although many MHC alleles currently present in human populations are known, we cannot directly access the extinct alleles. Thus, our estimate of the site covariation and the resulting optimal sequence partition must suffer from the limited number of sequences used to fit our model. Picking any one segmentation with a high likelihood over MHC sequences may lead to an oversimplification of the sequence representation. A posterior over the partitions, accompanied with latent variables giving sequence types in different parts, reflects more information about a set of amino acids in each MHC sequence than a latent structure based on one optimal segmentation.

Acknowledgements

We would like to thank Daniel Geraghty for providing access to the KIR dataset.

References

1. Joseph Bockhorst and Nebojsa Jojic. Discovering patterns in biological sequences by optimal segmentation. In *Proceedings of the 23st International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.
2. Joseph Bockhorst, Fangli Lu, Joel H. Janes, Jon Keebler, Benoit Gamain, Philip Awadalla, Xin zhuan Su, Ram Samurdala, Nebojsa Jojic, and Joseph D. Smith. Structural polymorphism and diversifying selection on the pregnancy malaria vaccine candidate VAR2CSA. *Molecular and Biochemical Parasitology*, 155:103–112, 2007.
3. Sebastian D. Fugmann, Alfred I. Lee, Penny E. Shockett, Isabelle J. Villey, and David G. Schatz. The RAG proteins and V (D) J recombination: complexes, ends, and transposition. *Annual Reviews of Immunology*, 18:495–528, 2000.
4. Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.
5. Austin L. Hughes, Marianne K. Hughes, and David I. Watkins. Contrasting roles of interallelic recombination at the HLA-A and HLA-B loci. *Genetics*, 133:669–680, 1993.
6. Nebojsa Jojic and Yaron Caspi. Capturing image structure with probabilistic index maps. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 212–219, 2004.
7. Nebojsa Jojic, Vladimir Jojic, Brendan Frey, Chris Meek, and David Heckerman. Using "epitomes" to model genetic diversity: Rational design of HIV vaccine cocktails. In *Advances in Neural Infomration Processing Systems (NIPS)*, number 18, pages 587–594, 2006.

8. Nebojsa Jojic, Vladimir Jojic, and David Heckerman. Joint discovery of haplotype blocks and complex trait associations from SNP sequences. In *Proceedings of the 20th International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2004.
9. Nebojsa Jojic, Manuel Reyes-Gomez, David Heckerman, Carl Kadie, and Ora Schueler-Furman. Learning MHC I–peptide binding. *Bioinformatics*, 22(14):e227–e235, 2006.
10. Oliver D. King and Frederick P. Roth. A non-parametric model for transcription factor binding sites. *Nucleic Acids Research*, 31(19):e116, 2003.
11. Steven N. MacEachern. Dependent Nonparametric Processes. *Proceedings of the Section on Bayesian Statistical Science*, pages 50–55, 1999.
12. Mukund Narasimhan, Nebojsa Jojic, and Jeff Bilmes. Q-clustering. In *Advances in Neural Infomration Processing Systems (NIPS)*, number 18, pages 979–986, 2006.
13. Morten Nielsen, Claus Lundegaard, Peder Worning, Sanne Lise Lauemoller, Kasper Lamberth, Soren Buus, Soren Brunak, and Ole Lund. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Science*, 12:1007–1017, 2003.
14. Bjoern Peters, Huynh-Hoa Bui, Sune Frankild, Morten Nielson, Claus Lundegaard, Emrah Kostem, Derek Basch, Kasper Lamberth, Mikkel Harndahl, Ward Fleri, Stephen S Wilson, John Sidney, Ole Lund, Soren Buus, and Alessandro Sette. A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput Biol*, 2(6):e65, 2006.
15. Bjorn Peters, Weiwei Tong, John Sidney, Alessandro Sette, and Zhiping Weng. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics*, pages 1765–1772, 2003.
16. Jim Pitman. *Combinatorial stochastic processes.* Springer Lecture Notes in Mathematics. Springer-Verlag, 2002. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, 2002.
17. Zhaohui S. Qin. Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics*, 22(16):1988–1997, 2006.
18. Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry Theory.* World Scientific Publishing Co., Inc., River Edge, NJ, USA, 1989.
19. Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.
20. Derek J. Smith, Alan Lapedes, Jan C. de Jong, Theo M. Bestebroer, Guus F. Rimmelzwaan, Albert D. M. E. Osterhause, and Ron A. M. Fouchier. Mapping the antigenetic and genetic evolution of influenza virus. *Science*, 305:371–376, 2004.
21. Yee W. Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
22. Daniel Ting, Guoli Wang, Maxim Shapovalov, Rajib Mitra, Michael I. Jordan, and Roland L. Dunbrack, Jr. Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLoS Comput Biol*, 6(4):e1000763, 04 2010.
23. Eric P. Xing, Roded Sharan, and Michael I. Jordan. Bayesian haplotype inference via the Dirichlet process. In *Proceedings of the 21st International Conference on Machine Learning*, pages 879–886. ACM Press, 2004.
24. Eric P. Xing, Kyung-Ah Sohn, Michael I. Jordan, and Yee W. Teh. Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture. In *Proceedings of the 23st International Conference on Machine Learning*, pages 1049–1056. ACM Press, 2006.