BAGS OF WORDS MODELS OF EPITOPE SETS: HIV VIRAL LOAD REGRESSION WITH COUNTING GRIDS

ALESSANDRO PERINA, PIETRO LOVATO and NEBOJSA JOJIC*

Microsoft Research, One Way Microsoft, Redmond WA, 98052.
E-Mail: jojic@microsoft.com, Tel: +1 (425) 705-5865

The immune system gathers evidence of the execution of various molecular processes, both foreign and the cells' own, as time- and space-varying sets of epitopes, small linear or conformational segments of the proteins involved in these processes. Epitopes do not have any obvious ordering in this scheme: The immune system simply sees these epitope sets as disordered "bags" of simple signatures based on whose contents the actions need to be decided. The immense landscape of possible bags of epitopes is shaped by the cellular pathways in various cells, as well as the characteristics of the internal sampling process that chooses and brings epitopes to cellular surface. As a consequence, upon the infection by the same pathogen, different individuals' cells present very different epitope sets. Modeling this landscape should thus be a key step in computational immunology. We show that among possible bag-of-words models, the counting grid is most fit for modeling cellular presentation. We describe each patient by a bag-of-peptides they are likely to present on the cellular surface. In regression tests, we found that compared to the state-of-the-art, counting grids explain more than twice as much of the log viral load variance in these patients. This is potentially a significant advancement in the field, given that a large part of the log viral load variance also depends on the infecting HIV strain, and that HIV polymorphisms themselves are known to strongly associate with HLA types, both effects beyond what is modeled here.

Keywords: Gene expression, Modeling host-pathogen interactions, Bag of Peptides

1. Introduction

The mammalian immune system consists of a number of interacting subsystems employing various infection clearing paths, with cellular presentation playing a central role in many of them. Most of the cells present a sample of peptides derived from cellular proteins as a means of advertising their states to the immune system. This facilitates globally coordinated action against viral infection.

The input to the cellular immune surveillance is illustrated in Fig.1. We show a simplified illustration of an infected cell which expresses both self (black) and viral (red) proteins (Fig.1A). Major histocompatibility complex (MHC) type I molecules bind to a small fraction of peptides from these proteins, created by proteasomal cleavage (Fig.1B). Inside these MHC complexes, the peptides are transported to the surface of the cell, where they may be detected by the cytotoxic T cells (CTL), which then may send self-destruct signals to the infected cell, thus stopping further infection (Fig.1C). Peptides that are a target of immune surveillance are often referred to as *epitopes*. As the sampled peptides do not appear in a particular spatial organization on the surface, the immune system effectively sees the infection as a bag of MHC molecules loaded with different viral peptides. Depending on the application, this representation may be further simplified into a *bag of viral peptides* (Fig.1D), under the assumption that the main effect of the MHC molecules is the peptide selection (e.g. choosing conserved vs non-conserved targets[6]).

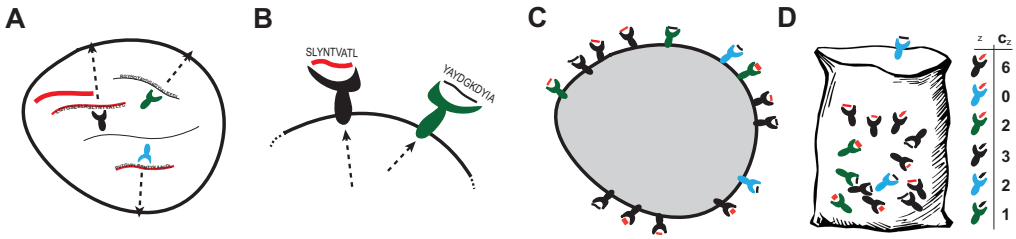

Fig. 1. Modeling immune surveillance input as a bag of words. **A** An Infected cell. **B** MHC binds to a fraction of peptides. **C** Sampled pepdites appear without particular order on the cell surface. **D** A bag of peptides represents the relative counts c_z of the features seen on cellular surface.

This paper has a dual purpose: *i)* it argues for the new application of *bag of words* models,[2,9] which have already been successfully applied in various other areas of machine learning, as a set of tools for capturing correlations in the immune target abundances in humoral and cellular immune surveillance, and *ii)*, it proposes a novel way of modeling bags of words which differs from PCA-like approaches not only in its treatment of observed epitope abundances as counts, but also moves away from the traditional componential structure towards a spatial embedding that captures smooth changes in cellular presentation.
In the experimental section, we restrict to the analysis of the links between the HIV viral load and the patients HLA types, leading to significant improvement with respect to the state of the art. Beyond the particular application tackled here, a good probability model of the epitope co-presentation has several direct applications, from correcting association studies, to detecting patients or populations that are likely to react similarly to an infection, to the rational vaccine design.

Related Work Explaining the differences in viral loads in different HIV patients has received a lot of attention from the HIV community, ever since the early longitudinal studies showed that changes in viral load occur in synchrony with the emergence of new HLA class I epitopes in immune assays.[4]
However, in case of the highly polymorphic HIV, a handful of epitopes usually fail to control the infection, and so researchers turned to population studies in search for optimal immune targets. Early studies failed to detect significant links between patients HLA types and viral load as the straightforward statistical approaches could not handle small dataset sizes (typically around 200 patients or less). But the evidence of HLA pressure on HIV was recognized in strong associations between viral mutations and patients' HLA types.[5] Viral load is highly variable and it may depend on numerous factors, such as gender, age, prior infections and general health of the individual. Thus it seemed likely that only the strongest MHC-driven effects would be visible through the noise. Still, any statistically significant result has been seen as having important consequences to HIV research. Eventually, larger cohorts allowed researchers to detect links between HLA types and viral load. Certain HLA B types, esp. B57 and B5801 were found to strongly associate with low viral load in a cohort of over 700 HIV patients in southern Africa.[7] In these studies, despite the statistically strong associations, the viral load in B57 or B5801 positive and negative patients still had such large variance that each of these HLA types alone could only explain less than 2% of the total log viral load variance in the

Table 1. The percentage of viral load (VL) explained in literature as the square of the Pearson's linear correlation coefficient (See Tab.2)

Ref.	**Major Result**
5	VL considered too noisy. Associations with mutations found
7	1-2% of VL variance explained through individual allele association
6	4% of VL variance explained through by targeting efficiency
10	4.3%-9% of VL variance explained by combinations of epitopes
This Paper	Up to 13.5% *of VL variance explained by embedding into Counting Grids*

population.

Multiple hypothesis testing issues and linkage disequilibrium among HLA loci complicated this research and the employed straightforward statistical approach did not present obvious ways to move from singular features (such as a binary labeling of patients as having B57 or not) to combinations of features that would provide higher explanatory power. However, by analyzing the tendency of the HLA molecules to bind to conserved targets in the HIV, it is possible to create a patient score (dubbed targeting efficiency) that captures binding characteristics of all 6 HLA molecules relative to HIV proteins.[6] At least on one cohort,[5] targeting efficiency explained a little less than 4% of the log viral load variance[a]. On the same cohort, another recent method deals with multiple features and their correlations, the *correlation sifting*,[10] explaining 4.3% of the log viral load variance by patients' HLA types. We show here that the bag of words models[3,9] lead to even better regression to viral load. This is especially the case for the new counting grid model[9] that efficiently captures correlations in cellular presentation by embedding patients in a grid, where the embedding coordinates can be used to explain 13.5% of log viral load variance, more than twice the current state of the art.

To put these numbers into perspective, it is important to make two observations. First, even weak signals, had the tendency to move the entire field,[5,7] as valuable characteristics of the interaction between HIV and the host immune system were revealed, informing both the research on HIV drugs and the research on HIV vaccine. Second, in addition to high variation of the viral load due to factors that relate to age and general health, it is known that the set point viral load depends strongly on the infecting strain,[8] and as HIV was found to mutate in its reactions to HLA presentation, this variation in fitness in the infecting strains may itself be due to the HLA pressure from previous hosts. Thus the increase in explanatory power of HLA types from around 4% of the log viral load to around 13.5% is potentially of great importance. Further analysis in selected combinations of features in the counting grid may lead to further advances in understanding the evolutionary arms race between HLA and the human immune system.

2. Bag of words models

In machine learning research, data samples are often represented as bags of features without a particular order. This choice is typically motivated by the difficulty or computational effi-

[a]Note again that the original analysis based on individual alleles failed to detect significant links with viral load there

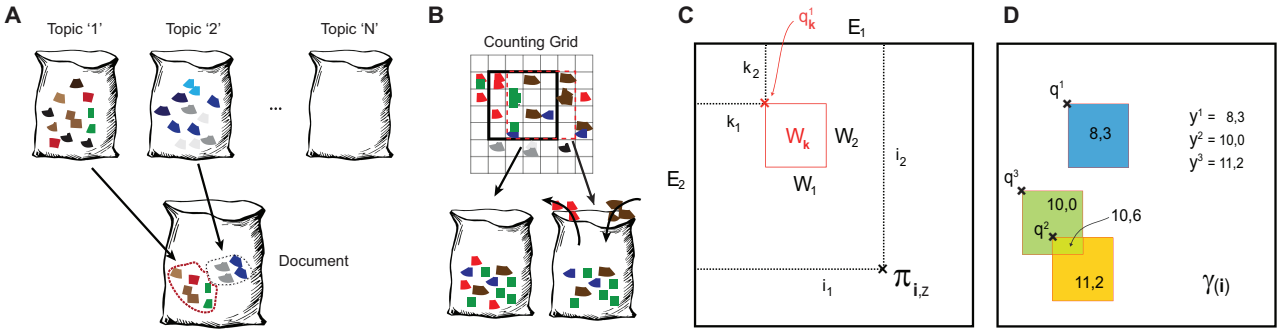

Fig. 2. Capturing dependencies in bags of words.

ciency of modeling the feature structure. Computational biology is abundant with examples of data where the structure is truly unknown, rather than just sacrificed for computational efficiency: for example, a gene expression array has been modeled as a bag of genes with expression levels simply corresponding to counts because most of the time little is known about the cellular pathways that employ these genes.[11–14] Without such knowledge there is no clear gene ordering. But biology is also abundant with situations where the raw data of interest actually has no (known or unknown) structure. In particular, in this paper we develop models of the sets of immune system targets.

Topic models[1,2] were introduced by the text analysis community and have been particularly successful in representing text documents. These simplified models of text assume that a text document has been generated simply by mixing words from a subset of possible topics. In typical applications, the number of possible topics is large, and these topics are inferred from the data by analyzing word co-occurrence patterns, and so the topic scope can vary from very narrow to quite broad, e.g., from near homonyms, to words found in most stories on US politics. An individual document is assumed to use only a fraction of all possible topics, and so the resulting bags of words will exhibit strong co-occurrence patterns: when the president is mentioned, so is the congress, as both appear in the same topic.

These models can be used in other domains by simply replacing words with some other set of features of interest. In bioinformatics, for example, words are replaced by genes and their counts by expression levels[3,12] to model microarray experiments. Visual descriptors are extracted from salient points of brain images and clustered into "visual words" replacing traditional words in bags of words and these representations were then used to classify schizophrenic patients from controls.[18] Peaks in nuclear magnetic resonance (NMR) spectrometry were also clustered and used as words.[19] Finally, protein sequences are sometimes broken into segments or *fragments*, which serve as words for comparing protein structures.[20]

Among topic models one of the best known is the Latent Dirichlet Allocation (LDA).[2] To formally define this model, we will index possible words (features) by z and denote the set of observed word (or feature) counts in the t-th bag of words by $\{c_z^t\}$. The latent (hidden) variables describe the choice of topics indexed by k. The choice of topics follows a distribution $p(k|\theta) = \theta_k$, and each topic has its own distribution over all the words $p(z|k,\beta) = \beta_{z|k}$. The

vector that depecits the topic distribution for one document θ is sampled from a Dirichlet distribution with parameters α. The following probability of generating a particular document is induced by this simple generative process (after picking the topic distribution θ, pick a topic, then pick a word form the topic, then pick a topic and a word from it again and again till all the words in the document are generated):

$$p(\{c_z^t\}|\alpha,\beta) = \int p(\theta|\alpha) \cdot \prod_z \Big(\sum_k (p(z|k,\beta) \cdot p(k|\theta))^{c_z^t} \Big) \; d\theta \tag{1}$$

The model parameters are estimated based on a training set so as to maximize the product of probabilities of all training documents. The topic proportions θ for individual documents can be used as a compact representation of the bag of words that discards the superfluous aspects of the data. For example, the HIV viral load can be regressed directly to these hidden variables in patient cohorts that are too small for the full representation of the viral presentation. Modeling cellular peptide presentation as a mixture of topics can capture some of the presentation patterns discussed above. Upon model fitting, the topics may correspond to individual MHC molecules that are more frequent in the patient cohort, or entire families of MHC types that have similar presentation (sometimes referred to MHC supertypes). In this case, all viral peptides would be indexed by z, and the topic probability distribution would reflects the probabilities of binding of a particular MHC (super)type to these different peptides. Some topics may also capture the HIV clade structure as mutations in each clade alter the MHC binding patterns.

Estimating bags of peptides for individual HIV patients

The concentration of any viral peptide on the cellular surface depends on the source protein's expression level. But different HIV proteins are expressed at different times in the HIV's infection and reproduction cycle. Instead of trying to estimate appropriate weighting factors, we simply considered each of the HIV proteins in isolation in our experiments.

As most epitopes are of length 9, for each analyzed protein we created a vocabulary of all *9-mers* that exist in this protein, indexed by z. Each human host has up to 6 different MHC I molecules (two from each of the three ancient duplicated and highly polymorphic loci A, B, C in the HLA region). In addition, in our experiments we dealt with a cohort in which we had the HLA types for each patient and we had access to an MHC I - peptide complex prediction algorithm that can estimate the *binding energy* $E_b(z,m)$ for each of the peptides z and the different patient's HLA molecules indexed by m.[21] Finally, we also used a *cleavage energy*[22] estimate $E_c(z)$ and turned the total energy into a count (concentration) as follows

$$c_z = e^{-E_c(z) - \min_m \left[E_b(z,m)\right]} \tag{2}$$

In a simplified model, the individual's immune system sees this variation in peptide counts (with many counts close to zero), and thus needs to recognize a virus not as a whole but as a set of disordered viral peptides.

Estimation of surface peptide (relative) counts could use any number of other epitope prediction techniques recently developed in computational biology.[23] Here we used the adaptive

double threading technique,[21] as it provides prediction for arbitrary MHC types simply defined by their protein sequence. NET MHC Pan[24] predictors provides similar functionality.

The counts c_z are not independent. The MHC system, as well as viral mutations, create links among the abundances of different viral peptides in the observed bag. Each MHC molecule has its binding preferences that lead to selection of only one of a hundred to a thousand of peptides. The human leukocyte antigen (HLA) region (human MHC) is the most polymorphic region of the human genome. As a result, two patients infected by the same virus, e.g. HIV, are highly unlikely to have the exact same MHC molecules. Each of their molecules will select specific targets from HIV proteins, and the patients' sets of immune targets will likely overlap only partially. The variation of the HIV epitope sets found in different patients exhibits strong co-occurrence patterns where a high count of one peptide often implies inclusion of several others, as they are all good binders to a particular MHC allele (families of different alleles can also share binding preferences). These links in epitope presentations are further expanded by weak linkage disequilibrium among MHC types as well as viral adaptation, which is itself correlated across sequence sites.
This all means that good models of bags of epitopes that constitute the immune surveillance input need to capture these correlations and this is precisely what the probability models of bags of words where meant to do for text documents.

3. The Counting Grid model

In the counting grid model, individual distributions over words are arranged on a grid (see Fig.2). Each of these distributions is relatively tight, with only a few features having significant probability. To generate a bag of words, instead of mixing topics, it is assumed simply that a window into the grid is opened, and the feature counts in the cells inside the window are combined to create the appropriate words in appropriate abundance. The window floating over the grid captures well variation in certain types of documents where we can see slow evolution of the topics, where certain words are dropped and new ones introduced: think for example to news stories over time, as interest in certain news slowly vanes in favor of new ones. Although traditional topics have been embedded in time or space and made slowly varying in certain directions, these variations do not quite capture the simple constraints present in CG models where a small window shift in the grid simply drops certain words and adds new ones. Furthermore, the counting grids are learned from the data for which the embedding in time or space is *not available*; this is the case for epitope bags. As we will show shortly, counting grids for this data can never the less be produced by iteratively estimating the grid distributions and inferring the mapping of the data to appropriate windows in it, thus resulting in the embedding of the data to a grid.

Formally, the basic counting grid $\pi_{\mathbf{i},z}$ is a set of normalized counts of words / features indexed by z on the D-dimensional discrete grid indexed by $\mathbf{i} = (i_1, \ldots, i_D)$ where each $i_d \in [1 \ldots E_d]$ and $\mathbf{E} = (E_1, \ldots, E_D)$ describes the extent of the counting grid. Since π is a grid of distributions, $\sum_z \pi_{\mathbf{i},z} = 1$ everywhere on the grid. A given bag of words/features, represented by counts $\{c_z\}$ is assumed to follow a count distribution found somewhere in the counting grid. In particular, using windows of dimensions $\mathbf{W} = [W_1, \ldots, W_D]$, each bag can be generated by

first averaging all counts in the hypercube window $W_{\mathbf{k}} = [\mathbf{k} \dots \mathbf{k}+\mathbf{W}]$ starting at D-dimensional grid location $\mathbf{k}$ and extending in each direction d by W_d grid positions to form the histogram $h_{\mathbf{k},z} = \frac{1}{\prod_d W_d} \sum_{\mathbf{i} \in W_{\mathbf{k}}} \pi_{\mathbf{i},z}$, and then generating a set of features in the bag. In other words, the position of the window $\mathbf{k}$ in the grid is a latent variable given which the probability of the bag of features $\{c_z\}$ is

$$p(\{c_z\}|\mathbf{k}) = \prod_z (h_{\mathbf{k},z})^{c_z} = \frac{1}{\prod_d W_d} \prod_z (\sum_{\mathbf{i} \in W_{\mathbf{k}}} \pi_{\mathbf{i},z})^{c_z} \tag{3}$$

Relaxing the terminology, we will refer to $\mathbf{E}$ and $\mathbf{W}$ respectively as the counting grid and the window size. We will also often refer to the ratio of the window volumes, κ, as a capacity of the model in terms of an *equivalent number of topics*, as this is how many non-overlapping windows can be fit onto the grid. Fine variation achievable by moving the windows in between any two close by but non-overlapping windows is useful if we expect such smooth thematic shifts to occur in the data, and we illustrate in our experiments that indeed they do. Finally, with $W_{\mathbf{k}}$ we indicate the particular window placed at location $\mathbf{k}$ (see Fig.2C). To learn a counting grid we need to maximize the likelihood of the data:

$$\log P = \sum_t \log \Big(\sum_{\mathbf{k}} \cdot \prod_z (h_{\mathbf{k},z}^{c_z^t}) \Big) \tag{4}$$

The sum over the latent variables $\mathbf{k}$ makes it difficult to perform assignment to the latent variables while also estimating the model parameters. The problem is solved by employing an iterative variational EM procedure. The E step aligns each bag of features $\{c_z^t\}$ to grid windows, to match the bag's histograms. In this way we compute the posterior distribution $q_{\mathbf{k}}^t$ over all windows $\mathbf{k}$ so that a better match between $\{c_z^t\}$ and $h_{\mathbf{k},z}$ across all features z yields a higher value for the match. In other words, $q_{\mathbf{k}}^t$ is probabilistic mapping of the $t-th$ bag to the grid widnows $\mathbf{k}$. This mapping is usually peaky, i.e., each bag tends to map to a few nearby locations in the grid. In the M-step we re-estimate the counting grid so that these same histogram matches are even better. To avoid severe local minima it is important to consider the counting grid as a torus, and perform all windowing operation accordingly. For details on the learning algorithm and on its efficiency see the original CG paper.[9]

Regression of continuous values

Once a CG is learned, we show here how one may embed continuous values y^t on the grid (e.g., HIV viral load). This is achieved using the posterior probabilities $q_{\mathbf{k}}^t$ for each bag already inferred and embedding the corresponding viral load inside the entire mapped window(s), and then averaging all overlapping windows (Fig.2D), which is similar to how M step re-estimates the distributions π:

$$\gamma(\mathbf{i}) = \frac{\sum_t \sum_{\mathbf{k}|\mathbf{i} \in W_{\mathbf{k}}} q_{\mathbf{k}}^t \cdot y^t}{\sum_t \sum_{\mathbf{k}|\mathbf{i} \in W_{\mathbf{k}}} q_{\mathbf{k}}^t} \tag{5}$$

The function γ can then be used for regression, in what is essentially a nearest-neighbor strategy: when a new data point is embedded based on its bag of words, the target is simply read out from γ, which is dominated by the training points which were mapped in the same region.

In Fig.4A we show a couple of γs, estimated from the dataset we used in the experiments. The window **W** is shown with a dotted line in the figure.

4. Experiments

In this section we first discuss what aspects of the epitope bags the counting grids may capture. Then we show that counting grids outperform not only traditional bag of words models, which have previously not been applied to this task, but also the state of the art in biomedical and computational biology literature[5–7,10] on analysis of the links between the HIV viral load and the patients HLA types (see Tab.1).

Types of correlations in epitope bags that can be captured with counting grids There are reasons why a counting grid model may be a more appropriate model of variation in epitope bags and perhaps more generally in many computational biology applications. These reasons have to do with the manner in which biological entities interact and adapt to each other leading to patterns of slow evolution characterized by genetic drift, local co-adaptation, as well as punctuated equilibrium. In case of cellular presentation, for example, millions of years of evolution created certain typical variants of MHC as well as minor variation on each of these major types. These variations are at least in part due to the interaction with viruses,[6] and similarly the genetic variation in viruses reflect some of this evolutionary arms race, too. Thus, the HIV clade constraints, as well as MHC binding characteristics may be so interwoven that a rigid view of cellular presentation as a mix of a small number of topics may be inappropriate. In the counting grid, the major variants of cellular presentation can be modeled as far away windows, while minor variations would be captured by slight window shifts in certain regions of the grid. To illustrate this we analyzed the cellular presentation of HIV patients from the Western Australia cohort.[5] We represented each patient's cellular presentation by a set of 492 counts over that many 9-long peptides from the Gag protein, previously found to be targeted by the immune system. The counts were calculated based on the patients MHC class I types (or HLA types, as they are called in humans) and the HLA-peptide binding estimation procedure discussed in Sec.2. This provides us with bags of peptides (*BoP*, counts over the 492 words) that represent GAG in different patients. We used the same process for two more proteins, POL and VPR, resulting in counts matrices of respectively 88×135 and 939×118 words×samples. We analyzed only the clade B infected patients.

Cellular presentation of viral peptides and viral load As the immune pressure depends on cellular presentation, the variation in cellular presentation across patients is expected to reflect on the variation in viral load, at least to some extent.[5,6] Viral load is expected to depend on the cellular presentation for various reasons. If the targeted peptides are conserved, this indicates inability of the virus to escape immune pressure. Even binding to some relatively variable peptides may lead to good outcomes for the patient (low viral load), as long as the CTLs can crossreact effectively across the peptide variants. In addition, there is a possibility that additional qualities of the peptides render some immune responses more effective than the others, or that certain immune responses trigger different viral behaviors. In bio-medical

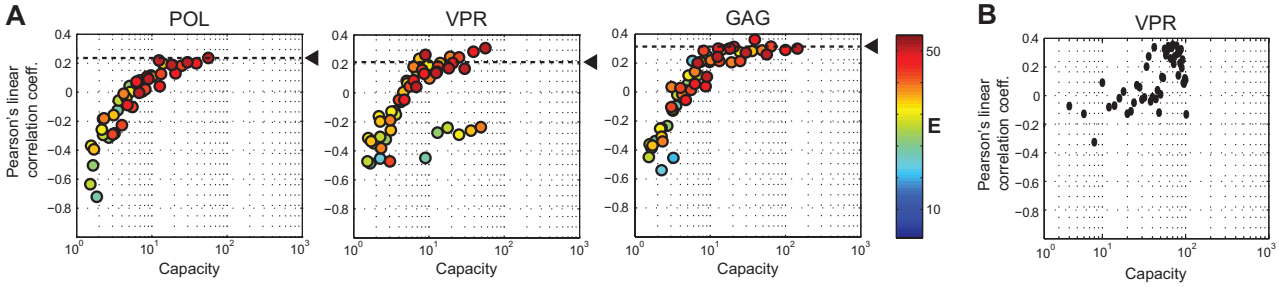

Fig. 3. HIV viral load regression. The variation of the correlation factor ρ for CG and LDA models of different complexities. Color code is used represent the square CG size $\mathbf{E}$ as a single capacity can be obtained with different $\mathbf{E}/\mathbf{W}$ combinations.

literature, analysis of this type of data targeted individual peptides and the discovery of those peptides that have significant association with viral load. However, these results do not explain nearly as much of viral load variance as what follows.

As general procedure, we first trained a CG using the bags-of-peptides c_z but *without* using the regression targets y^t (log viral load). Then, in a leave-out-out fashion, we held out a sample $\hat{t}$ and estimated the regression function γ (see Eq. 5, with $t \neq \hat{t}$) using all the other epitope bag/viral load pairs, and finally, read out γ in the appropriate (probabilistic) location $q^{\hat{t}}_{\mathbf{k}}$ to obtain the viral load prediction for $\hat{t}$ sample as $y^{\hat{t}}_{CG} = \sum_{\mathbf{k}} q^{\hat{t}}_{\mathbf{k}} \cdot \gamma(\mathbf{k})$. Once we computed the estimated regression target for all the samples, we computed ρ, the pairwise correlation coefficient between the true and the estimated viral load, comparing CGs with LDA,[3] and a technique based on phlogenetic trees[15] meant to established how much can the viral laod be predicted simply from the patient's dominant HIV sequence, as different strains may vary in fitness. We considered counting grids of various complexities $\mathbf{E} = [\mathbf{12}, \mathbf{15}, \mathbf{18}, \mathbf{21}, \mathbf{25}, \mathbf{30}, \mathbf{35}, \mathbf{40}, \mathbf{50}]$ and $\mathbf{W} = [\mathbf{2}, \mathbf{3}, \mathbf{4}, \dots]$. We tested only the combinations with capacity κ between 1.5 and $T/2$, where T is the number of samples available.

Rogers' LDA adaptation,[3] LPD originally designed for modeling microarray data was evaluated in a similar fashion. We learned as single model (without using the target) and we predicted the viral load for the left out sample using linear regression based on the topic proportions θ.

To compare with a sequence-based regression, we used the maximum likelihood approach[15] to estiamte a phylogentic tree for all patients' HIV sequences. Few parameters have to be tuned when computing such trees: In our experiments, we pick as a rate substitution matrix the WAG model,[16] and we allowed for rate variations across sites, setting 4 discrete gamma categories.[17] To predict the viral load $\hat{y}$ for a test sequence x using the estimated tree, we detected the training sequences that lie near by in the tree and averaged their viral loads acdcordign to their distance. If t indexes the training sequences x_t and their associated viral load value y_t

$$\hat{y} = \sum_t e^{-C \cdot \mathrm{dist}(x, x_t)} \cdot y_t \tag{6}$$

The parameter C has been found with crossvalidation on the training set. Fig.3, summarizes the performance of CG and LDA across a range of capacities κ for CGs and the number of topics K for LDA. LDA and CGs reach similar results of POL and VPR, while CGs have a clear advantage on GAG. It is important to note that for the Counting Grids, the correlation

factor varies much more regularly with the capacity κ, since this indicates that the complexity can be chosen on the training set through crossvalidation, which then allow us to properly calculate the percent of viral load explainable by the model. For each protein, we performed leave-one-out crossevaluation on the training set, to pick the best model complexity (**E**/**W** for Counting Grids, or the number of topics K for LDA) and we compared the results with the tree regression discussed above.

In leave-one-out experiments, the training set was each time used as a full set for another set of leave-one-out experiments on training data alone, plotting the graphs as above, and picking the best complexity. Then for the test sample we predicted the viral load using this best complexity. It is important to note that in this scheme *i)* the viral load of different patients can in principle be predicted using different complexities, and *ii)* the test sample does not contaminate the prediction model in any way. Results are shown it Tab.2. For Latent Dirichlet Allocation, this process failed and we could not obtain statistically significant results because of severe overtraining issues.

Finally, we also combined CG rpedictions with the idea of regressing the reconstruction error $E_z^t = \tilde{c}_z^t - R_z^t$ on residual viral load $y_{RED}^t = y^t - y_{CG}^t$,[10] where y_{CG}^t is the viral load prediction using the counting grid, and $\tilde{c}_z^t$ the normalized feature count. We used a regularized linear regression with L_1 norm using as before leave-one-out crossevaluation to choose the best model complexity. We computed the correlation factor ρ, setting final viral load prediction to be equal to the sum of y_{CG}^t and the prediction of y_{RED}^t. The idea here is that the deviation from the norm may be detecting viral adaptation and can predict furhter the modualtion of viral fitness. As can be seen in Tab.2, column $CGs\rightarrow^{10}$, this improved the performance in all the cases.

Interestingly, the model complexities chosen by each round of leave-one-out, though they could in principle be different for each patient, did not in fact differ that much. Regardless of the protein considered, for more than 89% of the data points the same complexity was typically chosen, as reported in the last column of Tab. 2.

Table 2. Pearson's linear correlation (after crossevaluation where applicable). Crossevaluation for LDA was found not statistically significant (NS) for **GAG** and **POL**. The last column reports the most common CG's complexity chosen in the rounds of leave-one-out crossevaluation.

	CGs	CGs→10	Trees	LDA	Ridge Regr.	Complexity Chosen
Protein	ρ	ρ	ρ	ρ	ρ	
GAG	0.3301	0.3674	0.3519	*NS*	0.1835	[**30**,**5**] - 89%
VPR	0.2011	0.2546	0.1061	0.1202	*NS*	[**50**,**8**] - 94%
POL	0.2338	0.2443	0.1812	*NS*	*NS*	[**40**,**11**] - 97%

The medical literature has other results obtained by analyzing GAG protein as shown in Tab.1, but the results reported here outperform all these methods, too.

We have one final note on the embedding function γ. The bags of peptides are mapped to the counting grid iteratively as the grid is estimated as to best model the bags, but the regression target, the viral load, was not used during the learning of CGs or LDA models. However, the inferred mapping after each iteration can be used to visualize how the embedded

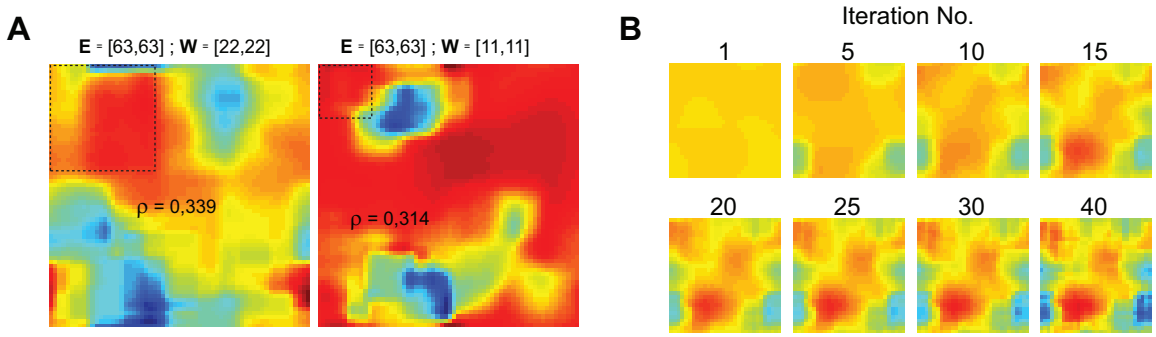

Fig. 4. **A** HIV viral load embedding in the 2D. The window is shown with a dotted line in the figure. **B** Evolution of the viral load across the iterations.

viral load γ evolves. This is illustrated in Fig.4B for a model of complexity $\mathbf{E} = [\mathbf{30} \times \mathbf{30}]$, $\mathbf{W} = [\mathbf{8} \times \mathbf{8}]$. The emergence of areas of high (red) and low (blue) viral load indicates that as the structure in the cellular presentation is discovered, it does indeed reflect the variation in viral load.

5. Conclusions

We propose the use of bag of words models to capture cellular presentation, and more generally the view that the immune system has of the invading pathogens. Furthermore, we demonstrate that the newest of these models, the counting grid, seems to be especially well suited to this task, providing stronger predictions than what can be found in bio-medical literature.
It remains to be understood exactly why CGs exhibit such a strong advantage over topic models (LDA). One intuitive explanation is that the slow smooth variations in count data that can be captured in counting grids better represent the dependencies that were produced by millions of years of coevolution between the HLA system and various invading pathogens.[6] This process involved numerous mixing of both the immune types and the viral strains, and may have produced the sort of thematic shifts in cellular representation that CGs are designed to represent. A more speculative possibility is that the immune system, through some unknown mechanism, collates the reports from circulating CTLs into an immune memory of a similar structure, though this summarization would obviously be performed over different invading pathogens in one patient, while our CGs depect one virus in a population of patients. Our experiments showed that cellular presentation of the Gag protein explains more than 13.5% of the log viral load. Although viral load varies dramatically across patients for a variety of reasons, e.g. gender, previous exposures to related viruses, etc., detection of statistically significant links between cellular presentation and viral load is expected to have important consequences to vaccine research.[7]

References

1. S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman, Latent Semnatical Indexing *Journal of the American Society for Information Science* **41**, 391 (1990)
2. D. Blei, A. Ng and M. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning Research* **3**, pp.993 (2003)
3. S. Rogers, M. Girolami, C. Campbell, R. Breitling, The latent process decomposition of cdna

microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2**, Vol. 2, 143–156 (2005)
4. A. McMichael *et al.*, Cellular immune responses to HIV. *Nature* **410**, 980-987 (2001)
5. C. Moore *et al.*, Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level. *Science* **296**, 436 (2002)
6. T. Hertz *et al.*, Mapping the Landscape of Host-Pathogen Coevolution: HLA Class I Binding and Its Relationship with Evolutionary Conservation in Human and Viral Proteins. *Journal of Virology* **85**, 1310 (2010)
7. P. Kiepiela *et al.*, Dominant in uence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* **432**, 769 (2004)
8. S. Alizon *et al.*, Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. *PLoS Pathogens* **6:9** (2010)
9. N. Jojic, and A. Perina, Multidimensional Coutning Grids: inferring words order from disordered bags of words. *Proceedings of Uncertainty in Artificial Intelligence* (2011)
10. J. Huang and N. Jojic, Variable selection by correlation sifting. *International Conference on Research in Computational Molecular Biology* (2011)
11. A. Perina, P. Lovato, V. Murino, and M. Bicego, Biologically-aware latent dirichlet allocation (balda) for the classification of expression microarray. *Proceedings of the IAPR international conference on Pattern recognition in bioinformatics*, (2010)
12. M. Bicego, P. Lovato, A. Perina, M. Fasoli, M. Delledonne, M. Pezzotti, A. Polverari, V. Murino Investigating topic models' capabilities in expression microarray data classification. *IEEE/ACM Trans. on Computational Biology and Bioinformatics*, **9**, No, 6, pp. 1831-1836 (2012)
13. F. Terrence, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics Journal*, **16**, No. 10, pp. 906-914, (2000)
14. I. Guyon, J. Weston, S. Barnhill, V. Vapnik Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, **46**, No. 1-3, pp. 389-422 (2002)
15. K. Tamura *et al.*, MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods, *Mol. Biol. Evol.*, 28, (2011)
16. S. Whelan and N. Goldman, A generical empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach, *Mol. Biol. Evol.*, 18, (2001)
17. Z. Yang Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods *J. Mol. Evol.*, 39, (1994)
18. U. Castellani, A. Perina, V. Murino, M. Bellani, G. Rambaldelli, M. Tansella, P. Brambilla Brain morphometry by probabilistic latent semantic analysis *International Conference on Medical Image Computing and Computer Assisted Intervention* (2010)
19. G. Brelstaff, M. Bicego, N. Culeddu, M. Chessa, Bag of Peaks: interpretation of NMR spectrometry *Bioinformatics Journal*, **25**, Vol. 5, pp. 258-274 (2009)
20. I. Budowski-Tala, Y. Novb, R. Kolodnya, FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately *Proc. Natl. Acad. Sci.* **107**, pp. 3481-3486 (2010)
21. N. Jojic, M. Reyes-Gomez, D. Heckerman, C.M. Kadie, O. Schueler-Furman Learning MHC I - peptide binding *Bioinformatics* **22**, Vol.14 pp.227-235 2006
22. H. Nielsen, J. Engelbrecht, S. Brunak and G. Von Heijne. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, 1-6, 1997.
23. G. Lan Zhang, V. Brusic et al. Machine Learning Competition in Immunology - Prediction of HLA class I molecules *J Immunol Methods*. 2011 Nov 30;374(1-2):1-4.
24. I. Hoof et al. NetMHCpan, a method for MHC class I binding prediction beyond humans *Immunogenetics*. 2009 January; 61(1): 1-13.