HAMdetector: Combining information to detect HLA-associated mutations with a Bayesian regression model

Daniel Habermann[1], Daniel Hoffmann[1]

Abstract

Motivation
The human leucocyte antigen system (HLA) is of paramount importance to combat viral infections by presenting peptides on the cell surface via MHC I. In this way, CD8+ cytotoxic T-Lymphocytes exert a strong selection pressure towards virus variants that escape that immune recognition pathway, e.g. through point mutations that decrease binding of the respective peptide to MHC I.

Reliably identifying HLA-associated mutations is important for understanding viral evolution, but experimental methods like binding assays are prohibitively expensive for large-scale use and fail to recognize other mechanisms of immune escape like proteasomal processing.
One step in finding these mutations is through the statistical analysis of sequence data. However, existing methods are based on nullhypothesis significance testing and do not make use of all the available information and therefore have unsatisfactory real-world performance.

Results
Here, we present a Bayesian regression model that is easily extensible to include information from different sources (e.g. epitope prediction software) and makes use of recent advances in Bayesian inference, e.g. by using a sparsifying prior. We show that including this kind of information improves predictive performance considerably over state-of-the-art methods.

Availability and Implementation
The source code of this software is available at http://github.com under a permissive MIT license.

Supplementary information
Supplementary data are available at *Bioinformatics* online.

Keywords
human leucocyte antigen system, HLA, multiple sequence alignment, escape mutations, viral escape, Bayesian inference, sparsity, horseshoe, epitope prediction

[1] *Bioinformatics & Computational Biophysics, Faculty of Biology, University of Duisburg-Essen, 45117 Essen, Germany*

Contents

1. Introduction

1.1 The HLA system

One way how the human immune system is able to recognize intracellular viral infections is through the human leucocyte antigen system: In cells with active protein biosynthesis, proteins are continuously synthesized and also degraded by a process called proteasomal degradation, which cleaves proteins into linear peptides of varying length. A small subset of these peptides is presented on the cell surface via receptors called MHC class I. The genomic region encoding for MHC I is known to be highly polymorphic, with more than 20000 different HLA alleles described today. The resulting gene products differ in their binding properties, which means that cells from different individuals present a highly diverse set of peptides on their surface. Cytotoxic T cells are selected during maturation to only weakly bind to peptide/MHC I complexes when the peptide originated from proteins of the usual proteome, but might be able to strongly bind to complexes of MHC I with peptides which are generated from of a viral protein. Upon activation, T cells induce cytolytic activity and recruit other immune cells.

1.2 HLA escape

In this way, the HLA system exerts a strong selection pressure towards virus variants that escape T cell recognition, for example through a point mutation that results in reduced binding of an immunogenic peptide to MHC I or through a set of mutations that alters the viral protein in such a way that it is cleaved into different peptides that are not recognized by the host's T cell repertoire.

The evolutionary events are complex and occur not only on the level of individuals, where a virus adapts to specific features of the host, but also on the population level, because HLA alleles differ in their frequency across geographic regions, as they are inherited according to standard Mendelian rules.

Upon transmission, HLA escape mutations typically quickly revert to their wild type as they usually reduce viral replicative capacity (if a mutation would increase viral replicative capacity regardless of the presence of a given HLA allele, it would probably already be the wild type). Kawashima et al. describe an escape mutation that is selected by HLA allele B*51, does not strongly affect viral replicative capacity and therefore slowly enriches over time in Japan, where B*53 commonly occurs. How quickly a given escape mutation is selected upon transmission in a host depends on the magnitude of the reduction in viral replicative capacity, on the strength of selection pressure and also on the genetic background, e.g. some escape mutations require compensatory mutations which partly attenuate the negative impact on viral replicative capacity.

Studying HLA escape therefore provides an unique opportunity to gain insight into viral evolution, on the host level but also on the population level. Unfortunately, identifying HLA escape mutations is difficult in practice.

1.3 Identifying HLA-escape mutations

There are several experimental methods available to study HLA escape: Recombinant MHC-I molecules can be used in binding assays: Upon complex formation with a peptide, a change in conformation can be detected with conformation-specific antibodies. This method is relatively fast, but only measures binding affinity of a peptide to MHC-I and does not account for antigen processing or immunodominance, which describes the observation that a peptide may be presented via MHC-I on the cell surface but does not induce an immune response. An experimental setup that resembles the conditions in-vitro more closely but is also more time-consuming is to measure CD8+ T cell responses instead. This is usually done by stimulating PBMCs with prototype and variant peptides and measuring the secretion of IFN-gamma by intracellular cytokine staining and fluorescence-assisted cell sorting. To analyse CD8+ T cell responses against endogenously processed antigens it is necessary to generate cell-lines stably expressing the antigen in question and adding antigen-specific CD8+ T cells. This method scales poorly as it requires transfection of cell lines and antigen-specific expansion of CD8+ T cells.

1.4 Computational methods

Because the currently available experimental methods do not scale well enough to analyze whole viral genomes, a useful strategy might be to use annotated sequence data to identfiy candidate HLA escape mutations that can then be verified experimentally.

As the selection pressure exerted by cytotoxic T cells depends on successful recognition of viral peptides on the cell surface, escape mutations are often HLA-allele specific and can therefore be detected as HLA-allele dependent footprints in sequence alignments of viral proteins: At certain alignment positions, a replacement might be more frequently observed in sequences from hosts with a specific HLA allele than in sequences from hosts without that HLA allele. By quantifying this difference for all replacement and HLA allele pairs it is possible to idenfify replacements that are enriched in sequences coming from hosts with certain HLA alleles and are therefore likely to be HLA escape mutations.

One way of quantifying this enrichment is Fisherś exact test. For a given replacement R_i at alignment position i and HLA allele H, a 2-by-2 contingency table is constructed containing the absolute counts of the number of sequences in the four possible categories: (R_i, H), (R_i, $!H$), ($!R_i$, H) and ($!R_i$, H), where $!R_i$ denotes any replacement except R_i and $!H$ denotes any HLA allele except H. Fisherś exact test is a conventional nullhypothesis significance test (NHST) that generates p-values. In this case, the nullhypothesis is that HLA allele H and replacement R_i are independent, and the p-value is the probability of observing a deviation from independence that is at least as extreme as in the data at hand, under the assumption that the nullhypothesis is true.

Fisher's exact test has the advantage of being easy to apply, but has also several disadvantages that are outlined in Carlson et al. The most striking one is that viral sequences share a common phylogenetic history and therfore, treating sequences as independent and identically distributed samples may under- or overestimate effect sizes and in the context of hypothesis testing leads to increased false positive and false negative rates. Another issue of applying Fisher's exact test is that HLA class I loci are located in close proximity on chromosome 6 and are therefore in linkage disquilibrium, which means that HLA alleles are not inherited completely independent of each other, so that inheritance of one HLA allele correlates with inheritance of another HLA allele. When using a statistical method that tests each HLA allele individually without considering the whole set of alleles present, spurious associations might occur : If HLA allele H_1 is associated with an amino acid replacement R, but H_1 is in linkage disequilibrium with another allele H_2, this also means that we observe an association between replacement R and H_2, even in absence of any underlying escape mechanism.

Correlations can not only occur between HLA alleles, but also between replacements. This kind of codon covariation occurs for example in compensatory mutations that attenuate the negative impact of immune escape mutations. For instance, a compensatory mutation might lead to a conformational change in such a way that the mutated protein resembles the original wild type more closely.

Carlson et. al. developed a method called Phylogenetic Dependency Network that accounts for phylogenetic bias, HLA linkage disequilibrium and codon covariation. and is based on nullhypothesis significance tests.

1.5 Issues with using p-values as a screening tool

In addition to the aforementioned biological reasons why Fisher's exact test is not suited for the analysis of sequence data, there are also more fundamental statistical issues that universally occur when using p-values as a screening tool.

In the presence of small effect sizes and high variance between measurements, as it is typically the case when working with biological data, statistically significant results can often be misleading and are likely to be in the wrong direction (a so-called type S error) or greatly overestimate an effect (a so-called type M error). This problem has recently been widely appreciated in the literature in the context of the current „replication crisis“, which describes the circumstance that scientific claims with seemingly strong statistical evidence fail to replicate.

Another issue that occurs when using p-values as a screening tool is the problem of multiple comparisons. When applying a statistical test, the probability of obtaining a statistically significant result increases with each additional test, even in absence of any real effect. When using p-values as a filter it is therefore likely to obtain significant effects that are in fact not real. To circumvent this problem, a common strategy is to control the false discovery rate, which is the expected proportion of false positives. These adjustment procedures have the problem that, when performing many of such comparisons, none but the very largest effects remain. Instead of performing many hypothesis tests and trying to adjust for them we instead prefer to fit a single, multilevel model that contains all comparisons of interest. When using multilevel models, the problem of multiple comparisons can disappear entirely and also yield more valid estimates.

2. System and methods

When possible, we choose to fit Bayesian models. By using prior information, adding problem-specific structure and partial pooling, the accuracy of estimates can often be noticably improved. Prior information does not necessarily mean to use external data, even a rough idea about the expected magnitude of estimates is often surprisingly effective. Additionally, Bayesian statistics provides an accessible way to test models: By comparing data generated under the model's assumptions to the actually observed data, it is possible to identify important aspects of the dataset that the models fails to capture and subsequently improve the model until it is consistent with the observed data.

In the context of identifying HLA-associated mutations we propose to improve existing methods by the following additions, which can be broadly divided into additional information and model structure:

Additional information

Binding affinity prediction

HLA-associated mutations are expected to lie more frequently in regions of known epitopes. There are vast experimental binding affinity data available of different peptides and MHC I molecule pairs and there are well-established computational methods that use these data to extrapolate HLA binding for untested peptides. We show that by including the outcome of these computational tools as input for a probabilistic model, the prediction of HLA-associated mutations can be improved.

Antigen processing prediction

Similarly, there are also HLA-allele independent effects like antigen processing that influence presentation on MHC I. Second generational tools like MHCFlurry 2.0 use both binding affinity and antigen processing data to improve epitope prediction. For our tool HAMdetector, we use the output of MHCFlurry to benefit from this binding affinity and antigen processing data for predicting HLA-associated mutations.

Model structure

Sparsity-inducing priors

Recent advantages in Bayesian inference include so-called sparsity-promoting priors. Sparsity-promoting priors convey the apriori expectation that most coefficients in a regression model are close to 0. This assumption also applies to HLA-associated mutations, because the number of epitopes that are restricted by a given HLA allele is typically small compared to the number of all possible epitopes. If no epitope spanning a given alignment position is presented on the MHC I molecule, any association between a replacement and the respective HLA allele is likely to be due to random variation alone (or HLA linkage disequilibrium).

It has been shown that using a sparsity-promoting prior when non-zero coefficients are sparse can drastically improve predictive performance because the model is better able to differentiate between signal and noise.

We show that a simple logistic regression model with a sparsity-promoting prior on the regression coefficients for the HLA alleles alone already performs roughly on-par with the much more elaborate Phylogenetic Dependency Network approach from Carlson et. al. when using the fraction of identified escapes that lie in known HLA epitopes as a benchmark.

Partial pooling of 4-digit HLA alleles

It has recently been appreciated that binding specifities can vary drastically across HLA alleles from the same allele group, e.g. between HLA-B*51:01 and HLA-B*51:03. For predicting HLA associated mutations, there has therefore been a shift to use 4-digit resolution data whenever possible. This is not without downsides however, because overall, binding specifities are more similar for alleles in the same group and therefore, treating them as completely separate might unnecessarily fragment the available data.

In Bayesian statistics, it is not required to chose one extreme or the other: Instead of either treating all HLA alleles from the same allele group as identical or as completely separate, partial pooling of the estimates allows something inbetween: By partial pooling, estimates do share some information across each other, but they are also allowed to vary if necessary.

In the context of our model this means that observing a relevant association between a replacement and an HLA allele also makes the model more inclined to estimate relevant associations between that replacement and all other alleles from that allele group. The degree of this partial pooling can be estimated from the data, so if we do observe strong similarity of HLA alleles in a given allele group, the estimates are more influenced by each other than if they do not.

3. Algorithm

4. Implementation

5. Discussion

6. References