How genetic sequence data can guide vaccine design

Muhammad Saqib Sohail [1], Ahmed Abdul Quadeer [1], and Matthew R. Mckay [1]

[1]Affiliation not available

October 30, 2023

Abstract

Vaccines have saved more lives than any other medical intervention throughout human history by preventing the spread of infectious diseases. However, despite several decades of research, there is no effective vaccine against fast evolving viruses such as the human immunodeficiency virus (HIV) and the hepatitis C virus (HCV). A confounding factor in the development of a HIV or HCV vaccine is that these viruses have a unique ability to make a lot of mutations in their genetic code. This enables them to escape the human immune system while retaining their ability to propagate infection. For developing a vaccine against such viruses, scientists are developing novel strategies which seek to target specific parts of the virus that are most vulnerable (i.e., where it is difficult for the virus to survive mutations) in order to induce a focused and potentially effective immune response. To determine the existence and location of such parts of HIV and HCV, initial studies have leveraged recently-available sequence data for these viruses, and looked for those positions in the genome for which the frequency of mutation was lowest. Unfortunately, vaccines based on such first-order statistics have not enjoyed much success, and there is increasing evidence suggesting that interactions between mutations is also important and must be considered when designing an effective vaccine against HIV and HCV. It is almost impossible to determine effects of interactions between all mutations experimentally as it requires performing billions of experiments. In this article, we explain how by leveraging virus sequence data, mutational interactions can be estimated using statistical techniques and incorporated in designing novel and potentially effective vaccine strategies against such fast-evolving viruses.

How genetic sequence data can guide vaccine design

Muhammad Saqib Sohail[1], **Ahmed A. Quadeer**[1], **and Matthew R. McKay**[1,2]

[1] Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China.
[2] Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China.

Keywords: sequence data; statistical techniques; data analytics; statistical correlations; random matrix theory; vaccine; vaccine design; virus; HIV; hepatitis C virus; immune system; T cells

Vaccines have saved more lives than any other medical intervention throughout human history by preventing the spread of infectious diseases. Experts credit vaccines with preventing an estimated 6 million annual deaths worldwide. Perhaps the greatest vaccine success story is that of the eradication of the smallpox disease. This disease is estimated to have killed about 300 million people between the years 1900 and 1977, while disfiguring and blinding millions more. For perspective, these are more lives lost than the combined death toll of both world wars.

Standard vaccine design strategies aim to create an inactivated or weakened virus, administered orally or intramuscularly, for eliciting immune responses that mount a lethal attack against the virus. This strategy has been successfully used to develop vaccines against many disease-causing viruses like polio, measles, mumps, rubella, chickenpox, and rotavirus, among others. However, despite several decades of research, there is no effective vaccine against fast evolving viruses such as the human immunodeficiency virus (HIV) and the hepatitis C virus (HCV). Vaccines can also be important for controlling the spread of novel viruses like the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) virus, which has caused the ongoing outbreak of the infectious coronavirus disease 2019 (COVID-19). Analysis of the genetic data of fast-evolving viruses, as well as emerging ones, can reveal novel insights that can guide the experimental efforts towards designing potent vaccines. Many efforts in this direction are currently under way for SARS-CoV-2.

In this article we focus on how data analytics can help guide vaccine design against fast evolving viruses by revealing new vulnerabilities in these viruses that are difficult, if not impossible, to find using traditional methodology. Fast evolving viruses like HIV and HCV infect millions of individuals worldwide and there is an urgent need of a vaccine against them. A confounding factor in the development of a HIV or HCV vaccine is that these viruses have a unique ability to make a lot of mutations in their genetic code. This enables them to escape the human immune system while retaining their ability to propagate infection. For developing a vaccine against such fast evolving viruses, scientists have been rethinking the vaccine design process and developing novel strategies. For example, instead of the standard procedure of using the complete inactivated or weakened virus, “subunit” vaccines are being explored that seek to target specific parts of the virus that are most vulnerable (i.e., where it is difficult for the virus to survive mutations) in order to induce a focused and potentially effective immune response. Such vaccines have been successful in preventing hepatitis B and human papillomavirus infections. To determine the existence and location of potentially vulnerable parts of HIV and HCV, initial studies have leveraged recently-available sequence data for these viruses, and looked for those positions in the genome for which the frequency of mutation was lowest. Unfortunately, vaccines based on such first-order statistics have not enjoyed much success, and there is increasing evidence suggesting that interactions between mutations is also important and must be considered when designing an effective vaccine against HIV and HCV. It is almost impossible to determine effects of interactions between all mutations experimentally as it requires performing billions of experiments. In this article, we explain how by leveraging virus sequence data, mutational interactions can be estimated using statistical techniques and incorporated in designing novel and potentially effective vaccine strategies against such fast-evolving viruses.

Virus, proteins, amino acid sequences

FIG1 Viruses are made up of proteins which are sequences of amino acids. The protein structure is from the publicly available structure of the HCV NS5B protein PDB ID: 1C2P; https://www.rcsb.org.

Before jumping into designing vaccines against viruses using sequence data, we first need to understand what a virus is, what information the sequence data carries, and how a virus leads to infection. Viruses are infectious agents comprising multiple proteins as shown in Figure 1. They contain the genetic material (viral genome) which is used as a template to make viral proteins during replication. Proteins are three-dimensional biomolecules made up of organic compounds called amino acids. A protein sequence is just the one-dimensional representation of a protein comprising a sequence of amino acids.

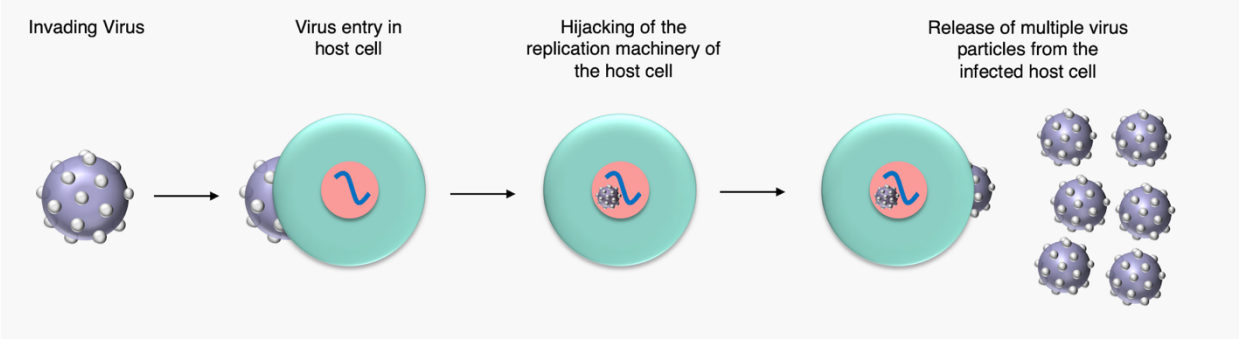

FIG2 Pathway taken by an invading virus to cause infection.

Viruses lack the capability to replicate on their own and thus require cells of other organisms to survive. Figure 2 shows the simplified series of steps a virus generally takes to replicate within a host organism and which eventually leads to infection. Viruses enter a host cell, hijack the cell's replication machinery, and force it to make multiple copies of the virus. The infected cell keeps on producing virus particles until all the resources of that cell are exhausted. The generated virus particles then go on to infect other cells. This cycle continues until the host organism dies due to the infection or the host immune system either clears the virus, or in some rare cases (e.g., in HIV and herpesvirus) keeps it under control but without clearing it.

How the human immune system combats invading viruses

In order to understand how data may aid the design of vaccines, it is important to first comprehend how our body naturally fights infections. The immune system is our body's defence system against invading pathogens (a disease causing microorganism, e.g., virus or bacteria). It consists of a vast network of cells and biological subsystems that can be broadly divided into two inter-linked systems: the innate and the adaptive immune systems. The innate immune system provides a generic rapid response and can clear common pathogens. Some viruses, however, have evolved strategies to evade the innate immune system

and go on to infect host cells. This is when the adaptive immune system spurs into action, and guided by the innate immune system, mounts a highly specific response against the invading virus to clear the infection. Importantly, the adaptive immune system has a unique ability to remember the virus it has encountered in the past, enabling it to mount an immediate and robust response upon reinfection with the same virus.

In this article, we focus on one of the key players of the adaptive immune system: T cells. This emphasis on T cells is naturally motivated by numerous studies that have reported the protective role of T cell responses against HIV and HCV. T cells are a type of white blood cells that have special receptors on their surface called T cell receptors (TCRs). Such receptors can recognize specific short fragment of proteins called “peptides”. They can distinguish between foreign peptides originating from virus proteins and self-peptides originating from proteins found in the body naturally. While an adult human has billions of T cells, they are very diverse with almost none of them having similar TCRs.

Figure 3 demonstrates the typical sequence of events involved in T-cell-based clearance of a viral infection. All nucleated human cells have certain molecules on their surface called the major histocompatibility complex molecules (MHCs). These molecules serve as billboards that display peptides extracted from the proteins being synthesised inside the cell. When a cell is infected by a virus, its cellular machinery is hijacked for producing copies of viral proteins and the MHCs of that cell are loaded with viral peptides. If the TCRs of a nearby T cell recognize a displayed viral peptide, a series of chemical reactions take place that result in the activation of the T cell and subsequent proliferation to produce a large number of T cells with the same specificity (i.e., the same TCRs). This army of activated T cells then moves to the infected area and kills all similar infected cells, thereby also killing the viral particles within these cells. Once the infection is cleared, most of the proliferated T cells die off, while a few remain in the body as long-lived memory cells for mounting a fast and robust response if the same virus is encountered in the future.

This memory of the adaptive immune system is the basis of vaccines. By inducing a controlled quantity of a weakened virus (conventional vaccines) or some specific parts of a virus (subunit vaccines) into the body, the goal of a vaccine is to elicit an immune response for recognizing this particular virus. Thus, if the body gets infected by the same virus in the future, the immune memory generated by the vaccine would enable the memory T cells to mount a strong response and consequently fend off the infection.

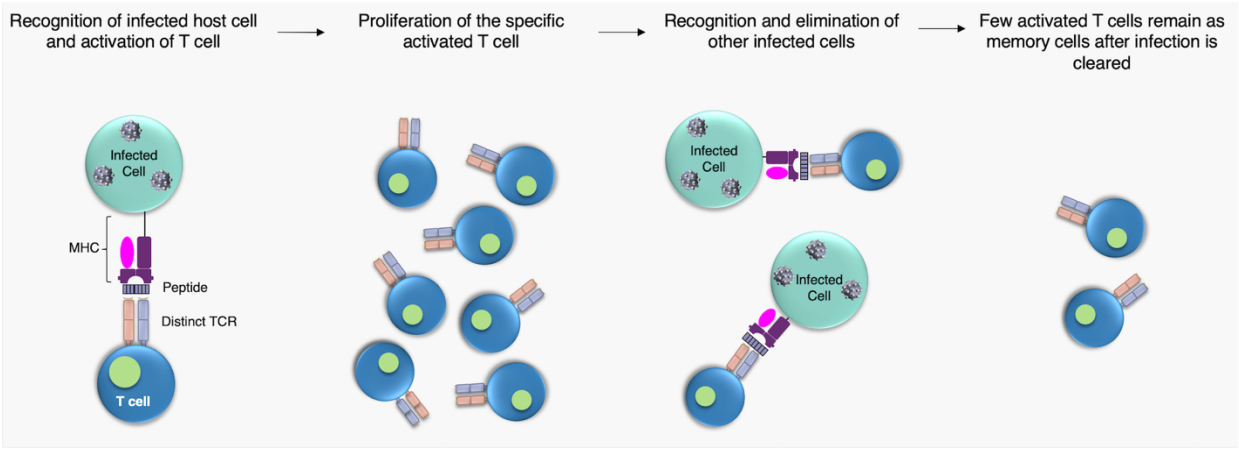

FIG3 Pathway taken by the human immune system to clear infection using T cells.

The big challenge to vaccine design: Genetic mutations

An ideal T-cell based subunit vaccine should comprise of viral peptides that elicit an immune response capable of clearing the virus. However, mutations occurring in the genetic code of a virus during replication

present a challenge. Mutations are random copying errors during replication and may result in an amino acid change in the genetic code of the virus. Due to the high specificity of T cells, a mutant virus may be able to abrogate recognition of vaccine-induced T cells and evade them if it carries a mutation in the part of its genetic code corresponding to the peptide employed in the vaccine. Such an escape mutation may either be harmful, beneficial, or harmless for the virus, depending on whether it strengthens, weakens, or does not affect its fitness (i.e., the ability of the virus to survive, replicate, and pass its genetic material to the next generation), respectively, as shown in Figure 4. From a vaccine design perspective, it is important to identify peptides in which mutations are harmful to the virus so that the chance of escaping the vaccine-induced T cell response is as low as possible. However, to identify the effect of each individual mutation on viral fitness experimentally is laborious and difficult as it involves running billions of experiments. This is precisely where data analytics can step in and potentially guide the vaccine design without conducting a large number of experiments.

Increasing amounts of viral protein sequence data have become available in recent years thanks to the rapid advancement in high throughput sequencing techniques. These sequences are obtained from blood samples collected from infected patients, and they can be leveraged to identify the effect of mutations by examining their statistical patterns in the data. For example, recent vaccine designs try to identify parts of the genetic code of a virus which are “conserved”, i.e., where mutations are not generally observed. The idea is that a higher level of conservation is an indication that mutations in these parts are generally harmful (and hence are not seen). This way, the vaccine elicits T cell response against the critical parts of the viral genetic code (conserved part) where a mutation would be potentially harmful for the virus.

As mentioned in the introduction, the vaccine strategy based on conservation of positions in the genetic code has not been effective against fast evolving viruses such as HIV and HCV. These viruses pose unique challenges for vaccine design as they have been reported to evade the immune system by exploiting pairwise interactions between different positions in the genetic code of the virus. That is, these viruses can escape by acquiring an individually harmful mutation in the parts of the viral genetic code targeted by the immune system but with its harmful effect defused by a compensatory interaction elsewhere in the genetic code. Moreover, for HIV and HCV, a large number of virus particles containing multiple different mutations can arise very rapidly due to the high mutation rates and short replication cycles of these viruses. This also makes the pairwise effects of mutations important as it opens up new compensatory pathways for the virus to escape the immune system as shown in Figure 4. Thus, considering only conservation of individual positions may not be sufficient for designing an effective vaccine against such fast evolving viruses. Rather, it seems important to account for the pairwise effects between mutations, in order to minimize the pathways that such viruses can employ for evading the immune system. Determining all pairwise effects is unfeasible to determine experimentally since, due to the large sizes of viral proteins, it would require running billions of experiments. Here, the role of data analytics becomes crucial for coming up with novel strategies for designing vaccines against scourges like HIV and HCV.

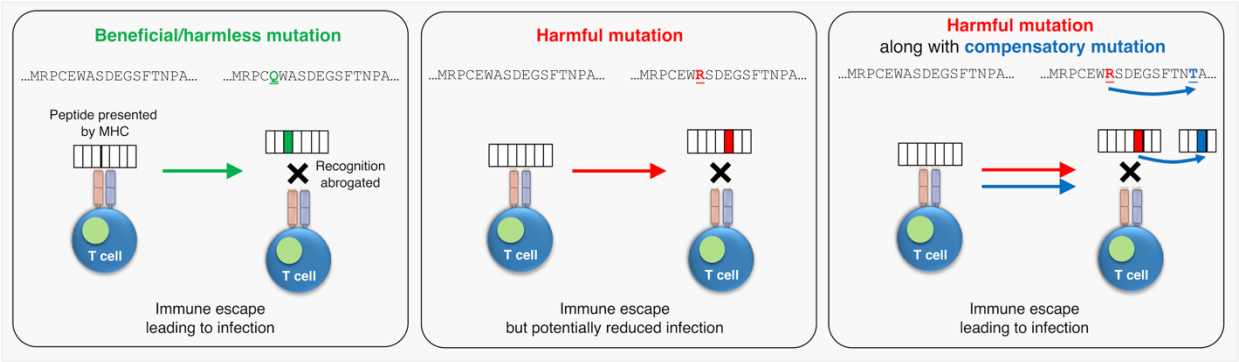

FIG4 Immune evasion strategy taken by fast evolving viruses.

Possible solution using data analytics – Move beyond conservation!

For combating fast evolving viruses, going beyond the conservation-based strategy employed in recent vaccine designs is necessary. In this regard, researchers at MIT (Dahirel et al.) pioneered to study the second order statistics, i.e., correlations between mutations at pairs of amino acid positions, of sequence data to come up with robust vaccine designs. To understand the basic idea of how mutational correlations can be useful, we present a simplistic example in Figure 5 where two (unmutated) viral peptides are considered for eliciting T cells. For the purpose of illustration, let us assume that a mutation is observed at only one amino acid position in each peptide. In this case, there can be two possibilities with respect to the correlation observed between mutations at these positions in the sequence data:

- If the mutations at these positions are positively correlated, it implies that these are observed together in the sequence data more frequently than if the mutations were to occur independently. Thus, one of these mutations may be compensating the individual harmful effect of the other. If a vaccine elicits T cells to target such peptides, the virus can seemingly mutate both these positions and escape both T cells while still retaining its fitness. Thus, peptides with positively correlated mutations would not appear to be good targets from a vaccine design perspective.
- If the mutations at these positions are negatively correlated, it implies that these are observed together in the sequence data less frequently than if the mutations were to occur independently. This would suggest that the combined effect of these mutations may have a harmful effect on the virus (and hence are not seen together). If a vaccine elicits T cells to target such peptides, the virus would either (i) mutate one of these positions and resist mutation at the other, resulting in recognition by one of the two T cells and getting killed; or (ii) mutate both positions, resulting in most likely an unfit virus incapable of causing infection. Thus, targeting peptides with negatively correlated mutations would appear to be a good vaccine design strategy.

The above example demonstrates that statistical correlations can be useful for designing potentially effective vaccines. However, inferring these from available viral sequence data presents a new statistical challenge. The sequence data of viral proteins is high-dimensional, i.e., these sequences are hundreds of amino acids long, and despite having thousands of sequence samples available, it is very difficult to reliably estimate pairwise correlations from such data. This is because for such high-dimensional data, the number of correlations to estimate is often comparable to or even larger than the number of available samples, thereby corrupting the estimates with a large amount of statistical noise. Classical statistical methods fail to provide accurate estimates of correlations for such high-dimensional data. For example, while the classical sample correlation matrix is a good estimator of the true correlation matrix when the number of variables is small and the associated number of samples is very large, it fails to estimate the true correlation matrix in such high-dimensional scenarios even in the simplest case when all involved variables are independent.

Such scenarios are not restricted to the viral sequence data but are in fact quite common these days in many fields including wireless communications, signal processing, and finance. Recently, advanced statistical techniques rooted in large-dimensional random matrix theory (RMT) have been attracting increasing attention for addressing the problem. These RMT-based techniques specifically aim to provide reliable estimates of correlations in cases when the number of variables are comparable to or even larger than the number of samples. While there has been much progress, estimating correlations from high-dimensional data is currently an active area of research in statistics, with still much room for improvement in the accuracy of the available techniques.

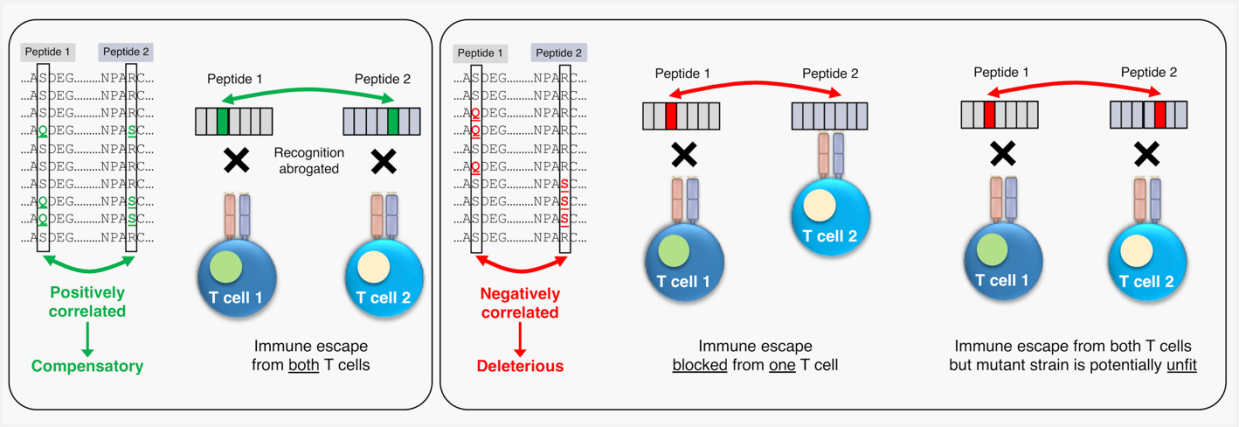

FIG5 Exploiting correlations to guide rational vaccine design against fast evolving viruses.

Robust vaccine design strategies using mutational correlations

In Figure 5, we described a simple example of a single mutation in two viral peptides for showing the potentially useful role mutational correlations can play in designing potentially effective vaccines. In reality, there are numerous peptides in a virus, with each comprising multiple positions where a mutation can occur. For inducing an effective immune response by a vaccine that is expected to ward off the virus, one would like to elicit T cells against a combination of peptides that is enriched in amino acid positions where a large proportion of combinations of mutations is harmful for the virus. Thus, we would like to select peptides such that the proportion of both the fully conserved positions, where no mutations are observed in the sequence data, and the negatively correlated inter-peptide mutation pairs is maximized, as shown in Figure 6. Moreover, we would also like to minimize the possibility of immune escape pathways available to the virus; this would imply that we must also minimize the proportion of positively correlated inter-peptide mutation pairs in the selected peptides. Another important factor that must be taken care of in selecting peptides for a T-cell-based vaccine is the population coverage. The idea here is to select peptides that can be presented by the MHC molecules of a large proportion of the population. Hence, by using statistical information computed from sequence data—conservation and mutational correlations, one can formulate a mathematical optimization problem to determine an optimal vaccine design (i.e., an optimal set of peptides to elicit T cells against).

The above strategy is used by Dahirel et al. and Ahmed et al. to propose robust T cell vaccine designs against HIV. A similar approach is used by Quadeer et al. to recommend robust HCV vaccine designs. All these studies use RMT-based techniques to infer mutational correlations from the available viral sequence data. Notably, these studies have shown that parts of viral proteins enriched in negative correlations are seemingly important for the virus to maintain its fitness. Moreover, such parts have also been shown to carry immunological importance, i.e., these are targeted by T cells of patients that either keep the virus under control (in case of HIV) or spontaneously clear the virus (in case of HCV). Thus, these results support the importance of targeting negatively correlated positions in vaccine design, as described above.

While these recent results are promising, there is a large scope for research in developing and further improving the statistical methods for inferring correlations from high-dimensional sequence data, which thereby can improve the design of vaccines. The ultimate effectiveness of such data-inspired robust vaccine designs hinges on their experimental and clinical outcomes. Testing this will require extensive collaborative efforts between clinicians, experimental biologists, and data scientists.

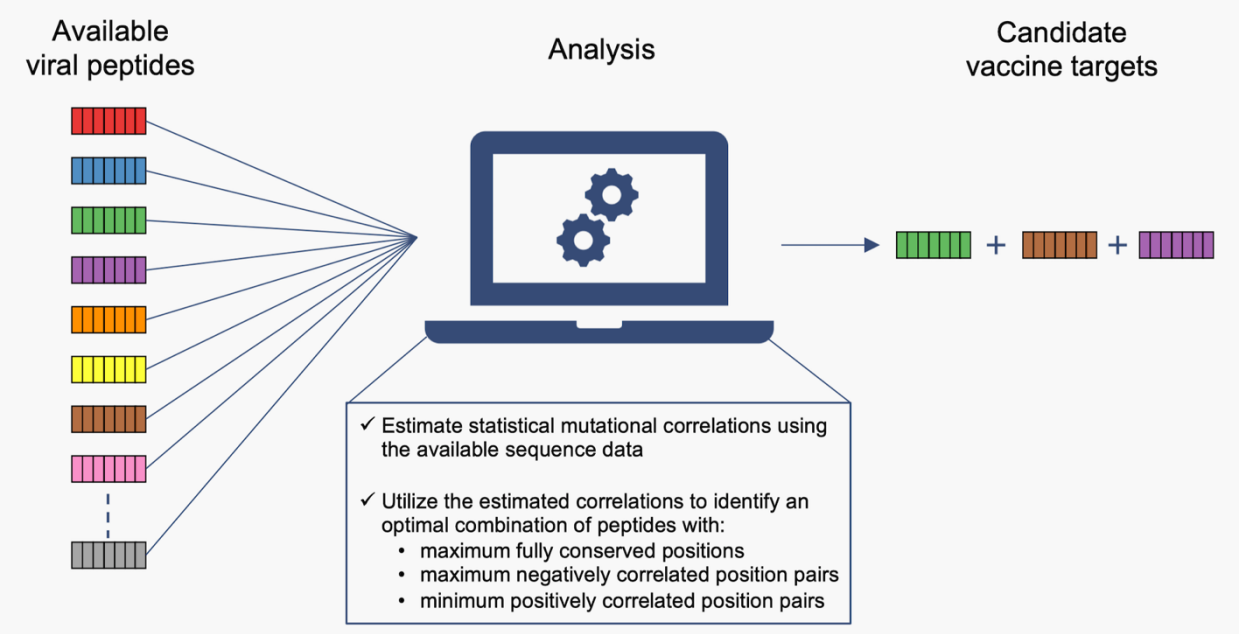

FIG6 Vaccine design based on statistical estimation of mutational correlations from the sequence data

Read more about it

- L. Sompayrac, *How the immune system works*, 4th edition, Hoboken, NJ: Wiley-Blackwell, 2012.
- V. Dahirel *et al.*, "Coordinate linkage of HIV evolution reveals regions of immunological vulnerability," *Proc. Natl. Acad. Sci.*, vol. 108, no. 28, pp. 11530–11535, Jul. 2011.
- S. F. Ahmed, A. A. Quadeer, D. Morales-Jimenez, and M. R. McKay, "Sub-dominnt principal components inform new vaccine targets for HIV Gag," *Bioinformatics*, vol. 35, no. 20, pp. 3884-3889, Oct. 2019.
- A. Quadeer, R. H. Y. Louie, K. Shekhar, A. K. Chakraborty, I.-M. Hsing, and M. R. McKay, "Statistical linkage analysis of substitutions in patient-derived sequences of genotype 1a hepatitis C virus non-structural protein 3 exposes targets for immunogen design," *J. Virol.*, vol. 88, no. 13, pp. 7628–7644, Jul. 2014.

Acknowledgment

We acknowledge financial support from the General Research Fund of the Hong Kong Research Grants Council, under grant numbers 16202918 and 16204519.

About the authors

Muhammad Saqib Sohail (mssohail@connect.ust.hk) is a PhD candidate at the Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China. His research focuses on developing computational solutions for problems in genetics and biology.

Ahmed A. Quadeer (aaquadeer@connect.ust.hk) is a postdoctoral fellow at the Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China. His research interests are focused on the use of statistical physics, signal processing, and machine learning tools for solving problems in biology and immunology.

Matthew R. McKay (m.mckay@ust.hk) is a Professor at the Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China. His research interests include communications and signal processing, random matrix theory, high-dimensional statistics, and computational biology/immunology.