Special Issue Paper

Statistics in Medicine

Received 21 August 2016, Accepted 18 November 2016 Published online 29 January 2017 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7203

Graph-based optimization of epitope coverage for vaccine antigen design

James Theiler[a,b] **and Bette Korber**[a,b*†]

Epigraph **is a recently developed algorithm that enables the computationally efficient design of single or multi-antigen vaccines to maximize the potential epitope coverage for a diverse pathogen population. Potential epitopes are defined as short contiguous stretches of proteins, comparable in length to T-cell epitopes. This optimal coverage problem can be formulated in terms of a directed graph, with candidate antigens represented as paths that traverse this graph. Epigraph protein sequences can also be used as the basis for designing peptides for experimental evaluation of immune responses in natural infections to highly variable proteins. The epigraph tool suite also enables rapid characterization of populations of diverse sequences from an immunological perspective. Fundamental distance measures are based on immunologically relevant shared potential epitope frequencies, rather than simple Hamming or phylogenetic distances. Here, we provide a mathematical description of the epigraph algorithm, include a comparison of different heuristics that can be used when graphs are not acyclic, and we describe an additional tool we have added to the web-based epigraph tool suite that provides frequency summaries of all distinct potential epitopes in a population. We also show examples of the graphical output and summary tables that can be generated using the epigraph tool suite and explain their content and applications. Published 2017. This article is a U.S. Government work and is in the public domain in the USA. Statistics in Medicine published by John Wiley & Sons Ltd.**

Keywords: vaccine; epitope; antigen; algorithm; directed acyclic graph; de Bruijn graph

ep·i·graph: an apposite quotation at the beginning of a book, chapter, or research article.

1. Introduction

The human immunodeficiency virus (HIV) causes a chronic infection that without treatment ultimately leads to AIDS and death, although the virus can be held in check by daily life-long antiretroviral therapy. HIV is a retrovirus with a high mutation rate. Immune responses in natural infection drive diversification *in vivo* by selecting for immune escape variants [1–4]. Thus, distinct and immunologically relevant mutations accumulate over time in every infected individual, creating a complex HIV quasispecies in each infected person; this translates to highly diverse viruses at the population level. This diversity limits the cross-reactivity of single antigen vaccines, such as a natural protein or a consensus sequence [5, 6]. Thus, we, along with our team at Los Alamos National Laboratory, developed a multiple antigen *mosaic* strategy about 10 years ago, expressly aimed at contending with this diversity [7]. The mosaic vaccine employs a genetic algorithm to generate antigens that collectively maximize their coverage of diverse epitopes in an HIV target population.

Epigraph is a recently introduced algorithm [8] that uses a graph-based approach to maximize the same potential epitope coverage that the mosaic algorithm maximized, but with better computational efficiency and, under some conditions, with provably optimal solutions. Building on [8], which emphasized applications of the epigraph code, here, we provide a more detailed mathematical description of the epigraph algorithm, a comparison of various heuristics for removing cycles from the directed graph, and an overview of what is available in the online tool that is hosted on the HIV Database website (https://www.hiv.lanl.gov/content/sequence/EPIGRAPH/epigraph.html). This includes the ability to

[a]*Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A.*
[b]*New Mexico Consortium, Los Alamos, NM 87545, U.S.A.*
[*]*Correspondence to: Bette Korber, Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A.*
[†]*E-mail: btk@lanl.gov*

'characterize potential T-cell epitopes (PTEs)', and we show what that characterization looks like for US B-clade Gag protein.

The foundation of both mosaic and epigraph design is that antigen optimization is based on coverage of PTEs, which are strings of k contiguous amino acids (or k-mers), as first suggested in [9]. Because the optimal length of most cytotoxic T-cell epitopes is nine amino acids, we usually set $k = 9$ [7, 10]. Epigraph's speed allowed us to explore additional vaccine design issues that were previously intractable with mosaics. These include studying the impact of excluding of rare epitopes, optimizing on coverage of imperfectly matched epitopes, and incorporation into more complex iterative clustering algorithms that enabled a tailored therapeutic vaccine design [8].

Vaccines designed using the HIV mosaic code have performed very well when tested experimentally; we expect epigraphs to behave comparably. To date, all mosaics that have been tested produce proteins that are well folded (i.e., they bind to many neutralizing antibodies that recognize discontinuous antibody epitopes – epitopes that consist of nonadjacent amino acids that are brought close together in natural folded proteins). Furthermore, they are highly immunogenic, eliciting both T-cell and B-cell antibody responses [6, 11, 12]. T-cell responses to mosaic vaccines effectively target HIV-infected cells [13], and, as intended, they are far more cross-reactive than those induced by natural proteins [6, 11, 14–16]. They can be applied to whole proteins, or just to regions that are relatively conserved; even the most conserved regions of HIV are variable when considering epitope length fragments [10, 17, 18]. Mosaic designs have also been explored for influenza [19], hepatitis C [20], Ebola [8, 21], and Chlamydia [22].

Of note, the diversity coverage of antibody epitopes as well as T-cell epitopes should theoretically be augmented by using mosaic/epigraph designed polyvalent vaccines rather than using combinations of natural strains. Antibody epitopes are generally discontinuous, but contain short linear stretches of neighboring amino acids that come together in the folded protein. By algorithmically favoring inclusion of commonly found combinations of amino acid that are in close proximity in sequence space, as a consequence of maximizing k-mer coverage in mosaic/epigraph vaccines, and by tending to minimize the inclusion of rare amino acids, even discontinuous antibody epitopes will tend to enriched for forms of the epitopes that are common among natural strains. As epigraphs are complementary sets, common variants will be represented, and through exposure to common variants during affinity maturation, antibodies may be selected that have greater breadth. In contrast, amino acids that are very rare in a given position and also very rare local combinations of amino acids pepper virtually every natural HIV sequence; thus, when natural proteins are used for vaccination, they may yield type-specific vaccine responses if rare amino acids or combinations of amino acids are embedded in a immunogenic epitopes. For example, Env is the key antibody vaccine target for HIV-1. When we generated an epigraph for the HIV Env protein global alignment, containing 4256 sequences each isolated from a different individual, we found there were a remarkable 650,000 distinct 9-mers. Over 500,000 were only found once in the whole population of 4256 sequences. Each Env is less than 900 amino acids long, and this averaged to 120 unique 9-mers per natural strain [8].

2. Formulation

A multiset $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}$ of N sequences is taken to characterize the variability of a virus over a population of interest. Each sequence is a string of alphabetic characters, corresponding to the 20 amino acids. Some special characters are allowed, including a gap character (-) in the case of aligned sequences, and characters indicating a premature stop codons (*), an uncertain amino acid due to an ambiguous base call within the corresponding codon (X), and a frame-shifted codon (#).

For this exposition, we define an *epitope* as any short subsequence of k amino acids. Strictly speaking, this is an abuse of terminology, because true epitopes correspond to very specific subsequences that are known to be immunologically relevant. For viral proteins of interest here, however, such as HIV-1 Gag, T-cell epitopes are ubiquitous across the entire Gag sequence [23]. More formally (and in previous treatments [8]), we refer to these k-mers as 'PTEs'. Here, however, we will simply use 'epitope' to mean k-mer.

For each epitope $\mathbf{e}$, we assign an integer frequency $f(\mathbf{e})$ corresponding to the number of sequences in $\mathcal{S}$ in which the epitope appears. If the epitope appears more than once in a given sequence (e.g., due to a repeat), it is still only counted once for that sequence.

The simple version of the problem is to design an artificial sequence $\mathbf{q}$ that optimally 'covers' the epitopes in $\mathcal{S}$. If we write $\mathcal{E}_{\mathbf{q}}$ as the set of epitopes that appear in the sequence $\mathbf{q}$, then our measure of coverage is the sum $\sum_{\mathbf{e}\in\mathcal{E}_{\mathbf{q}}} f(\mathbf{e})$ of the frequencies for all the epitopes that appear in $\mathbf{q}$. It is convenient to normalize this quantity by the sum of the frequencies of all the epitopes: $\sum_{\mathbf{e}\in\mathcal{E}_{\mathbf{q}}} f(\mathbf{e}) / \sum_{e} f(\mathbf{e})$.

In the 'cocktail' version of the problem, the aim is to generate $M > 1$ artificial sequences $\mathcal{Q} = \{\mathbf{q}_1, \ldots, \mathbf{q}_M\}$ that optimally cover the sequences in $\mathcal{S}$. We write $\mathcal{E}_Q$ as the set of epitopes that appear in at least one of the sequences in $\mathcal{Q}$, and we seek to optimize the sum $\sum_{\mathbf{e}\in\mathcal{E}_Q} f(\mathbf{e})$ of the frequencies for all the epitopes that appear in $\mathcal{Q}$.

2.1. *Graph-based formulation*

A *graph* is composed of two sets: a set of nodes and a corresponding set of edges, where each edge connects a pair of nodes. In a *directed graph*, the edges have a direction, and the connection goes *from* one node *to* the other. The graphs we will construct are directed and very much like the de Bruijn graphs [24] that are used in other kinds of sequence analysis, particularly in sequence assembly [25, 26].

We can construct a single directed graph in which each k-mer epitope $\mathbf{e}$ is associated with a node. A pair of epitopes $\mathbf{e}_a, \mathbf{e}_b$ is connected by a directed edge if the last $k-1$ characters of $\mathbf{e}_a$ match the first $k-1$ characters of $\mathbf{e}_b$. For example, the nodes $\mathbf{e}_a =$ `VTSSNMNNA` and $\mathbf{e}_b =$ `TSSNMNNAD` are connected by a directed edge because they share the $k-1$ characters `TSSNMNNA` and therefore could be adjacent (overlapping) epitopes in a sequence ...`VTSSNMNNAD`.... A *path* through the graph is a sequence of nodes, associated with epitopes $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_L$, such that there is an edge from $\mathbf{e}_\ell$ to $\mathbf{e}_{\ell+1}$ for $\ell = 1, \ldots, L-1$. Such a path corresponds to a sequence of $L + k - 1$ characters.

For computational convenience, we add two extra nodes, a BEGIN and an END node. The BEGIN node connects to all the nodes that lack predecessors (corresponding to epitopes that are the first k characters in a sequence). Nodes that lack successors (because they are the last k characters in a sequence) are connected to the END. All paths of interest, then, go from the BEGIN node to the END node.

In this graph-based formulation, the goal is to identify a path P through the graph that maximizes $\sum_{\mathbf{e}\in P} f(\mathbf{e})$ where the sum is over all the epitopes that are in the path (but if a given node appears more than once in a path, its epitope frequency $f(\mathbf{e})$ is only included once in the sum).

2.2. *Solving the graph-formulated $M = 1$ problem in the ideal case*

If the directed graph of epitopes is acyclic (which means that there are no paths that include the same node more than once) and if only $M = 1$ artificial sequence is needed, then we can find the optimal path through this graph, which corresponds to the optimal vaccine sequence given our criteria of maximizing epitope coverage.

The algorithm for finding the optimal path uses dynamic programming, a strategy that has been widely used for other sequence analysis problems, including sequence alignments [27, 28]. The algorithm involves a forward loop followed by a backward loop.

The forward loop defines the function $F(\mathbf{e})$ to be the largest sum achievable for any path that terminates with the epitope $\mathbf{e}$. This function can be computed for each node in the graph in a stepwise manner. Let $\mathcal{P}(\mathbf{e})$ be the set of predecessors of node $\mathbf{e}$: that is, the set of nodes $\mathbf{e}'$ for which there exists a directed edge that connects from $\mathbf{e}'$ to $\mathbf{e}$. Then we have

$$F(\mathbf{e}) = f(\mathbf{e}) + \max_{\mathbf{e}'\in\mathcal{P}(\mathbf{e})} F(\mathbf{e}') \tag{1}$$

If the set of predecessors $\mathcal{P}(\mathbf{e})$ is empty, then we define $F(\mathbf{e}) = f(\mathbf{e})$.

If the graph of epitopes is a directed acyclic graph, then there exists a 'topological ordering' of the epitopes, $\mathbf{e}_1, \mathbf{e}_2, \ldots$, with the property that if $(\mathbf{e}_i, \mathbf{e}_j)$ is a directed edge, then $i < j$. By proceeding in this topological order, we can straightforwardly evaluate Equation (1) for all the nodes.

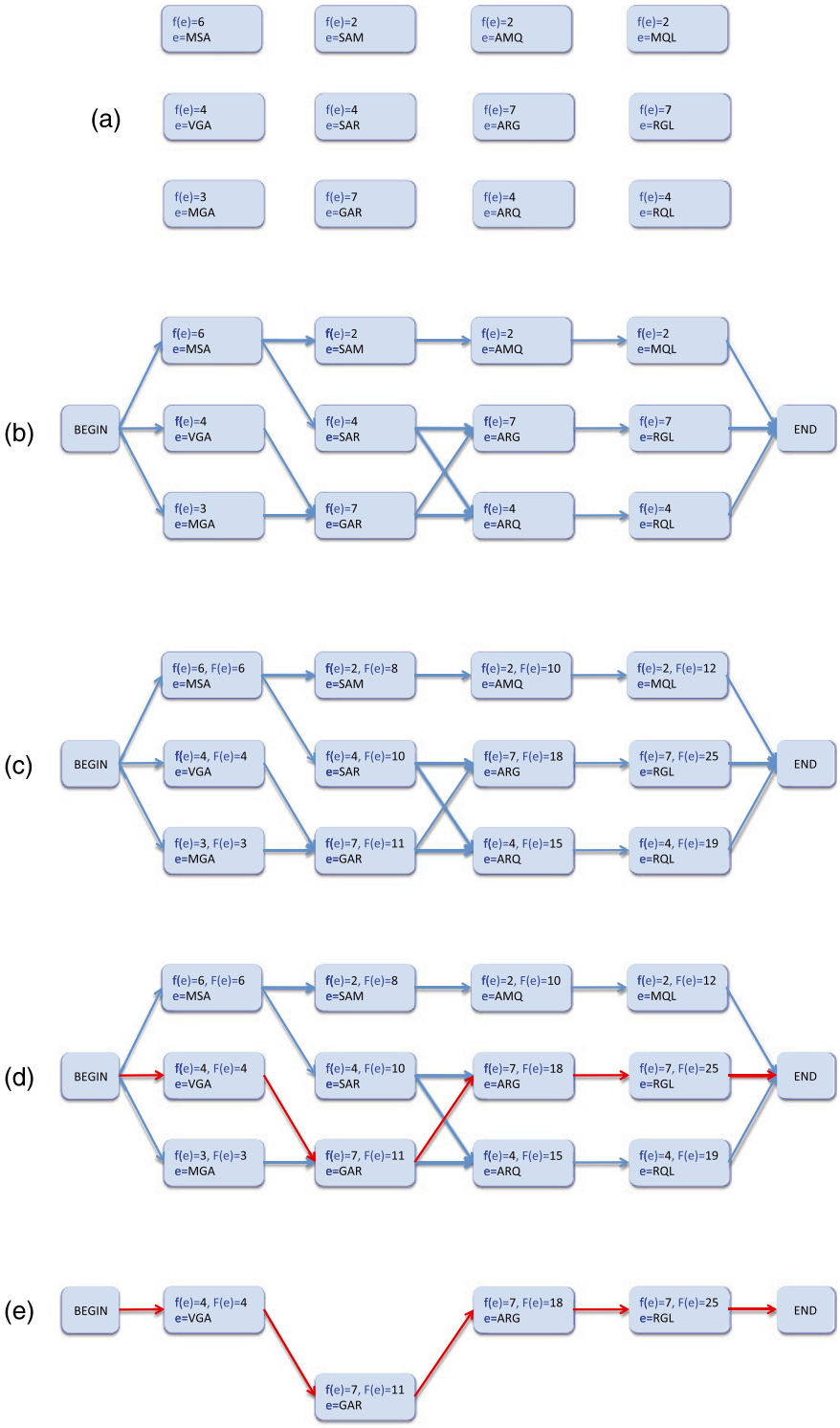
Having evaluated $F(\mathbf{e})$ for all the nodes $\mathbf{e}$, we choose a node with maximum value: $\mathbf{e}^*_0 = \text{argmax}_{\mathbf{e}} F(\mathbf{e})$. This will be the final epitope in our optimal string. Now, we just work backwards:

$$\mathbf{e}^*_{p+1} = \text{argmax}_{\mathbf{e}\in\mathcal{P}(\mathbf{e}^*_p)} F(\mathbf{e}) \tag{2}$$

If the set $\mathcal{P}(\mathbf{e}^*_p)$ is empty, then we are finished, and the sequence of epitopes $\mathbf{e}^*_p, \mathbf{e}^*_{p-1}, \ldots, \mathbf{e}^*_0$ corresponds to a sequence $\mathbf{q}$ of $p + k$ characters that optimizes the epitope coverage.

We remark that the argmax operator may not have a unique value; if it does not, then there will be multiple solutions, all of which are optimal in the sense of coverage.

Furthermore, this optimality is achieved with computational effort that scales only linearly with the size (as measured in edges) of the network. Figure 1 illustrates the creation of a graph and the optimal path through the graph, for a toy example involving $k = 3$-mers, spanning 13 'sequences' of six amino acids each.



Published 2017. This article is a U.S. Government work and is in the public domain in the USA. Statistics in Medicine published by John Wiley & Sons Ltd.
Statist. Med. **2018,** 37 181–194

Note that the constraint $\min_{\mathbf{e}\in\mathcal{E}_Q} f(\mathbf{e}) \geqslant f_o$ is very easy to impose; one simply eliminates all nodes in the graph (and all edges that attach to those nodes) for which $f(\mathbf{e}) < f_o$. The epigraph algorithm then seeks a path through this smaller graph.

2.3. *Multi-antigen ($M > 1$) vaccines*

Here, we seek a set of antigens $\mathcal{Q} = \{\mathbf{q}_1, \ldots, \mathbf{q}_M\}$ that collectively cover the sequences in $\mathcal{S}$. With $\mathcal{E}_Q = \cup_{m=1}^{M} \mathcal{E}_{q_m}$, the set of epitopes that appear in at least one of the sequences in $\mathcal{Q}$, the goal is to maximize $\sum_{\mathbf{e}\in\mathcal{E}_Q} f(\mathbf{e})$.

A variety of strategies were described in [8], but the main idea is to *sequentially* add new antigens q_m in a way that optimizes a *complementary* coverage function. As in the $M = 1$ case, this coverage is a sum of frequencies $f(\mathbf{e})$, but the sum is only over epitopes that have not yet appeared in any of the other antigens. Algorithmically, one proceeds just as for the $M = 1$ case but the sum is over $f^*(\mathbf{e})$ that is equal to $f(\mathbf{e})$ for epitopes that have not yet appeared in the vaccine and is equal to zero for epitopes that have already appeared in the vaccine.

3. Decycling

Unfortunately, the network that is created from a sequence list $\mathcal{S}$ is not guaranteed to be acyclic. In practice, particularly for larger values of k (and larger values of f_o), the network is often 'very nearly' acyclic and can be made acyclic with only a few perturbations to the network. The optimal solution to this perturbed network is then taken as a nearly optimal solution to the original network.

Removing the least number of edges to produce an acyclic graph is equivalent to an NP-hard problem, the 'minimum feedback arc set' problem [29,30]. Thus, we cannot expect a universally efficient algorithm for optimally decycling a graph, and that is what motivates us to consider a variety of heuristic approaches for making the directed graph acyclic with a minimum amount of perturbation.

The first step in eliminating cycles is to identify them. To do this, we decompose the graph into 'strongly connected components' [31]; for an acyclic graph, each node is its own strongly connected component. Within a single strongly connected component, a path can be found from every node to every other node. Cycles can be identified in the following way: if $\mathbf{e}_a$ and $\mathbf{e}_b$ are two nodes in the same component, then the

Figure 1. Simple example with $k = 3$ and a population sample composed of these 13 sequences: 2×MSAMQL, 4×MSARGL, 4×VGARQL, and 3×MGARGL. In this simple illustration, we take $k = 3$, but in actual practice, we choose k in the range 8–12, with $k = 9$ generally preferred. (a) Nodes are associated with *k-mers*, each of which is a potential T-cell epitope. In this simple illustration, there are twelve nodes, corresponding to the twelve distinct 3-mers that appear in the 13 sequences listed in the legend. For each 3-mer **e**, the expression $f(\mathbf{e})$ corresponds to the number of sequences in which e appears. (b) The Graph is produced by connecting pairs of nodes with directed edges. If the last k - 1 characters of one node match the first k - 1 characters of another node, then a directed edge connects the one node to the other. In this case the last 2 characters of the first node match the first 2 characters in the following node. Thus, MSA connects to SAM and to SAR, but not to GAR. Also, we add a BEGIN and an END node to the graph as a bookkeeping convenience. (c) The frequency $f(\mathbf{e})$ corresponds to the number of sequences in the sample population in which the k-mer associated with the node e appears. For each path P that ends on epitope **e**, we can compute the sum of frequencies of the nodes in the path. The maximum of this sum, over all such paths, is given by $F(\mathbf{e})$. For a directed acyclic graph, such as the one in this illustration, we can compute $F(\mathbf{e})$ very efficiently using the expression in Eq. (1). In particular, to compute $F(\mathbf{e})$, find the largest F(e′) among the nodes e′ that are predecessors to **e** and add that to $f(\mathbf{e})$. For instance, ARQ has two predecessors: e′ =SAR and e′ =GAR. F(e′) = 11 for GAR, and adding that to $f(\mathbf{e})$ = 4 for e =ARQ gives $F(\mathbf{e})$ = 15. (d) The optimal path, which maximizes the sum of $f(\mathbf{e})$ over all the nodes in the path, is obtained by starting at the END node and working backward. At each step, the predecessor with the largest value of $F(\mathbf{e})$ is chosen. In this illustration, of the three predecessors to END, the node RGL has the $F(\mathbf{e})$ = 25, which is larger than for the other two predecessors. We work backward to ARG, and then choose between SAR and GAR; we select GAR since its $F(\mathbf{e})$ = 11 is higher than SARs value. We continue moving backward until we reach the BEGIN node. (e) The Epigraph solution (here given by VGARGL) is associated with the optimal path [VGA,GAR,ARG,RGL]. Note that this solution is not among the sequences that were in the population sample, and that it is not the consensus sequence of those sequences. It is, however, the single sequence that maximizes the coverage of the 12 distinct 3-mers, given their frequencies.

Published 2017. This article is a U.S. Government work and is in the public domain in the USA. Statistics in Medicine published by John Wiley & Sons Ltd.
Statist. Med. **2018,** 37 181–194

Table I. Decycling heuristics: a heuristic function provides a value for each edge in the directed graph; edges with the lowest values are removed until the graph becomes acyclic.

Heuristic	Description
sum	A very simple and (empirically) effective heuristic is to take the value to be the sum $f(\mathbf{e}_a) + f(\mathbf{e}_b)$.
max	Another simple heuristic is to take the maximum of the two values associated with the two nodes that define the given edge: that is, max $\big(f(\mathbf{e}_a), f(\mathbf{e}_b)\big)$.
iso	We observe that if $\mathbf{e}_a$ is the *sole* predecessor of $\mathbf{e}_b$, then cutting edge $(\mathbf{e}_a, \mathbf{e}_b)$ will isolate node $\mathbf{e}_b$; similarly, if $\mathbf{e}_b$ is the sole successor to $\mathbf{e}_a$, then cutting the edge will isolate $\mathbf{e}_a$. The cost associated with the iso statistic takes the sum of the $f(\mathbf{e})$ values of the isolated nodes.
sum+iso	The cost associated with edge $(\mathbf{e}_a, \mathbf{e}_b)$ is the sum of the sum and iso costs. This is the heuristic we used in [8].
max+iso	Sum of max and iso heuristics.
posn	Every node can be assigned a position $x(\mathbf{e})$ corresponding to the length of the shortest path that connects the BEGIN node to $\mathbf{e}$. We use $-x(\mathbf{e}_a)$ as the heuristic value, thus preferring to cut edges whose first node is farthest from BEGIN.
del-posn	The value given by $x(\mathbf{e}_b) - x(\mathbf{e}_a)$ tells us the extent to which the directed edge points 'forward' – large negative values correspond to edges that point backward, and are good candidates for cutting.
random	Assign random values to edges.

directed path from $\mathbf{e}_a$ to $\mathbf{e}_b$ can be merged with the directed path from $\mathbf{e}_b$ to $\mathbf{e}_a$ to form a cycle (although possibly not a 'simple' cycle) that includes both $\mathbf{e}_a$ and $\mathbf{e}_b$.

Our approach for eliminating cycles is to keep all the nodes from the original network, but to successively cut edges until an acyclic network is obtained. Each time a cycle is located in the graph, we choose one of the edges in the cycle to remove from the graph. This choice is heuristic, and we tried a number of alternatives. But in general, because cutting edges may have the effect of isolating nodes, we seek cuts that isolate low-value nodes. For each edge $(\mathbf{e}_a, \mathbf{e}_b)$, we can define a value, based on $f(\mathbf{e}_a)$ and $f(\mathbf{e}_b)$; then we choose the edge with the smallest value and remove it from the graph. We considered eight specific heuristic functions, listed in Table I.

Because the choice of which cycle to draw from the graph is based on random choices, we do seven trials for each of the eight heuristics and consider both the average and the maximum coverage values for those seven trials. The results are shown in Figures 2–5. These experiments, to some extent, vindicate our choice of heuristic (sum+iso) used in [8], which is seen as a stable choice that gives results that are, over a range of different proteins and clades, consistently competitive with the more expensive mosaic algorithm. But this heuristic does not always lead to the best coverage. In Pol B, we see the posn heuristic produces significantly higher coverage than any other heuristic, even though posn for other proteins often gives significantly poorer coverage.

As a general rule, we expected the heuristics that removed the fewest edges (i.e., that produced the least 'damage' to the full graphs) to give the highest coverage, and although that trend roughly holds (particularly for the Gag and Env proteins), there are some striking exceptions. For both Nefs B and M, we see that the del-posn heuristic produces not only the worst or second worst damage but also the highest coverage. Further study is called for (both in comparing heuristics and in developing new ones), but a reasonable strategy at this point would be to apply the whole slate of available heuristics to any given population sample of interest and to simply take the one with the best coverage.

Because different heuristics yielded different outcomes depending on the input dataset, we have how added the ability for the user to specify which heuristic to employ in the design phase of epigraph.

4. The epigraph web interface

Users can execute the epigraph algorithm on their own data using a web interface [32] maintained by the Los Alamos HIV database (http://www.hiv.lanl.gov/).

4.1. Human immunodeficiency virus type 1 Gag as a test case

User input for all of the web-based tools in the epigraph tool suite is a set of sequences considered to be a reasonable (or best available) sampling of the pathogen of interest's protein diversity in infected population that is being targeted for a vaccine, or for immune response studies. Epigraph, like all of the tools provided through the HIV database project, has a readily available input sample set to enable users to quickly explore the tool, even if they do not have a data set to hand.

For the epigraph tool, we chose a highly relevant data set as the sample input, incorporating the Gag p24 proteins sampled within the last decade in the USA, one sequence derived from each of 189 different individuals. We pulled this set from the HIV database pre-made alignments [33], and analogous HIV sets can readily be obtained representing any global region, country, or HIV subtype.

The immunology informing Gag p24 as the choice for our example file is worth considering, because Gag is an excellent protein for inclusion in T-cell response vaccines [34] and the example highlights the merits of an epigraph approach. Gag p24 is one of the most conserved proteins in HIV, but conservation is relative, and even p24 is highly variable at the epitope level. The number of vaccine-elicited Gag cytotoxic T-lymphocyte (CTL) responses has been shown to directly correlate with viral control in rhesus macaque SIV challenge models [35], and Gag is very immunogenic. It is spanned by an dense overlay of experimentally well-defined known human T-cell epitopes [10]. To see the extent of known CTL epitopes, see also the HIV database maps of known epitopes collected from the literature [23]. This list of experimentally defined HIV epitopes is certainly an underestimate, as most experimental studies rely on the use of consensus peptides (e.g., see [36, 37]), and when specific peptides are synthesized that exactly match an infecting virus in an individual whose immune response is being studied, CTL responses to novel epitopes are readily discovered [38]. The high density of T-cell epitopes is further supported by population-based T-cell recognition studies [36, 37].

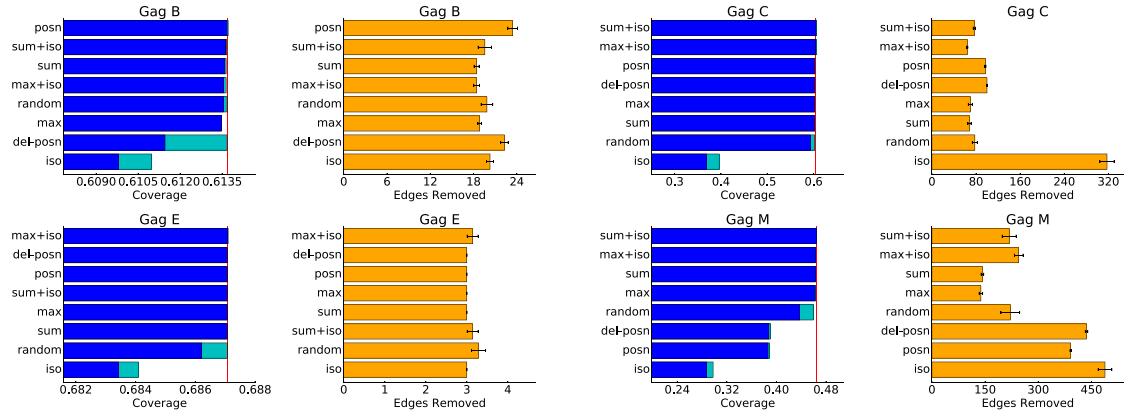


Figure 2. Coverage score (left, blue) and removed edges count (right, orange) for four Gag protein clades B, C, E, and M. Each plot compares results for eight different heuristic decycling algorithms; these heuristics are described in Table I. Dark blue bars correspond to average coverage, over seven trials. The lighter cyan bars show the maximum coverage, over those seven trials. The bars are arranged so that the top bar has the highest average coverage and the bottom bar has the lowest average coverage. The vertical red line corresponds to the value provided by the mosaic algorithm, as reported in [8]. Light orange bars indicate the number of edges removed from the graph, using the different heuristics, in order to obtain an acyclic graph.

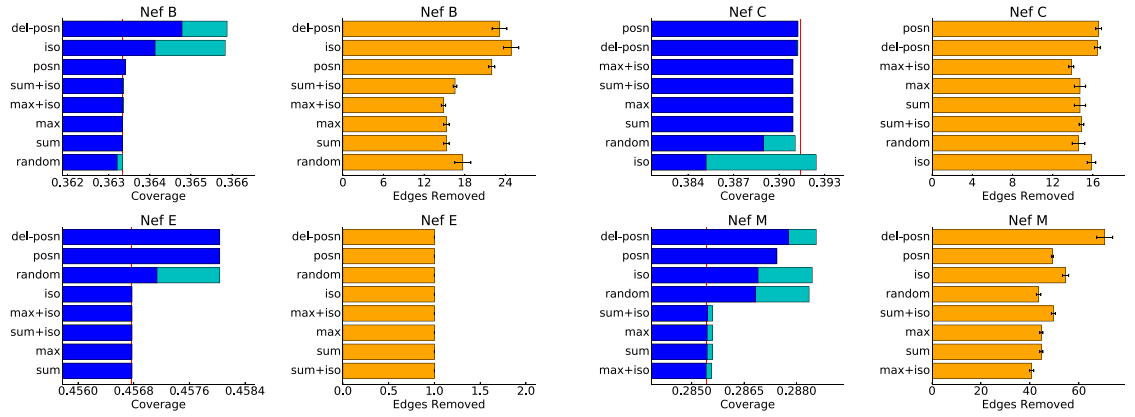


Figure 3. Similar to Figure 2, but for the Nef protein.

Published 2017. This article is a U.S. Government work and is in the public domain in the USA. Statistics in Medicine published by John Wiley & Sons Ltd.

Statist. Med. ***2018,*** 37 181–194

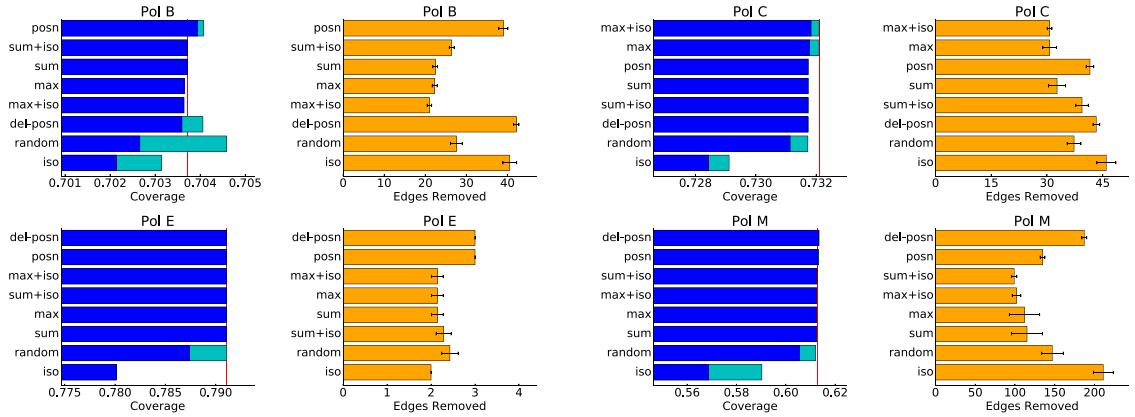


Figure 4. Similar to Figure 2, but for the Pol protein.

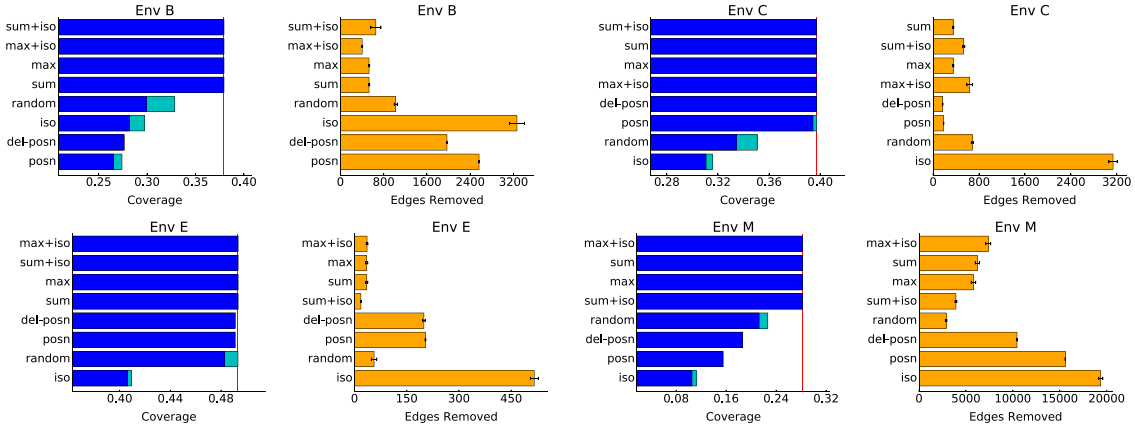


Figure 5. Similar to Figure 2, but for the Env protein.

A high density of known epitopes is not exclusive to HIV Gag and other HIV proteins. For example, the influenza A hemagglutinin protein, another intensively studied viral protein, has over 800 T-cell epitopes listed in the Immune Epitope Database (http://www.iedb.org/), suggesting a comparably dense tiling of epitopes to Gag. Such epitope density is part of the justification for the assumption that any 9-mer fragment is a potential T-cell epitope in epigraph/mosaic design. This is particularly important when considering the array of epitope target possibilities across a diverse vaccinated population with a complex spectrum of human leukocyte antigens (HLAs). HLAs are the human proteins that present T-cell epitopes: They are highly polymorphic, and they play a key role in determining which particular epitopes are recognized in a given individual. While not all 9-mers will be precisely defined optimal epitopes, still they will partially overlap with valid epitopes. The epigraph solution is optimized across all possibly relevant potential T-cell epitopes, without having to define those regions *a priori* when not enough data is available to define and differentiate natural epitope regions.

In a natural infection, a given individual will recognize a relatively small number of T-cell epitopes [38], and this is also true for vaccines using standard delivery methods [39]. But vaccination with a cytomegalovirus (CMV) vector changes this paradigm. Rhesus macaques vaccinated with SIV immunogens delivered in rhesus CMV vectors can clear over half of the infections that are established upon challenge [40, 41]. The T-cell response to these vaccines is extraordinary. A remarkable number of CTL epitopes in Gag are recognized [40, 41], and they are presented in unusual and not readily predictable ways, using the non-classical major histocompatibility complex E molecule for presentation [42]. Epigraphs were originally developed with the design of HIV inserts for preclinical development of CMV vectors in mind [8]. Given the very high density of non-canonical and unpredictable epitope responses generated in the context of the CMV vector, it is reasonable to treat all 9-mers as potential epitopes.

Published 2017. This article is a U.S. Government work and is in the public domain in the USA. Statistics in Medicine published by John Wiley & Sons Ltd.
Statist. Med. ***2018,*** 37 181–194

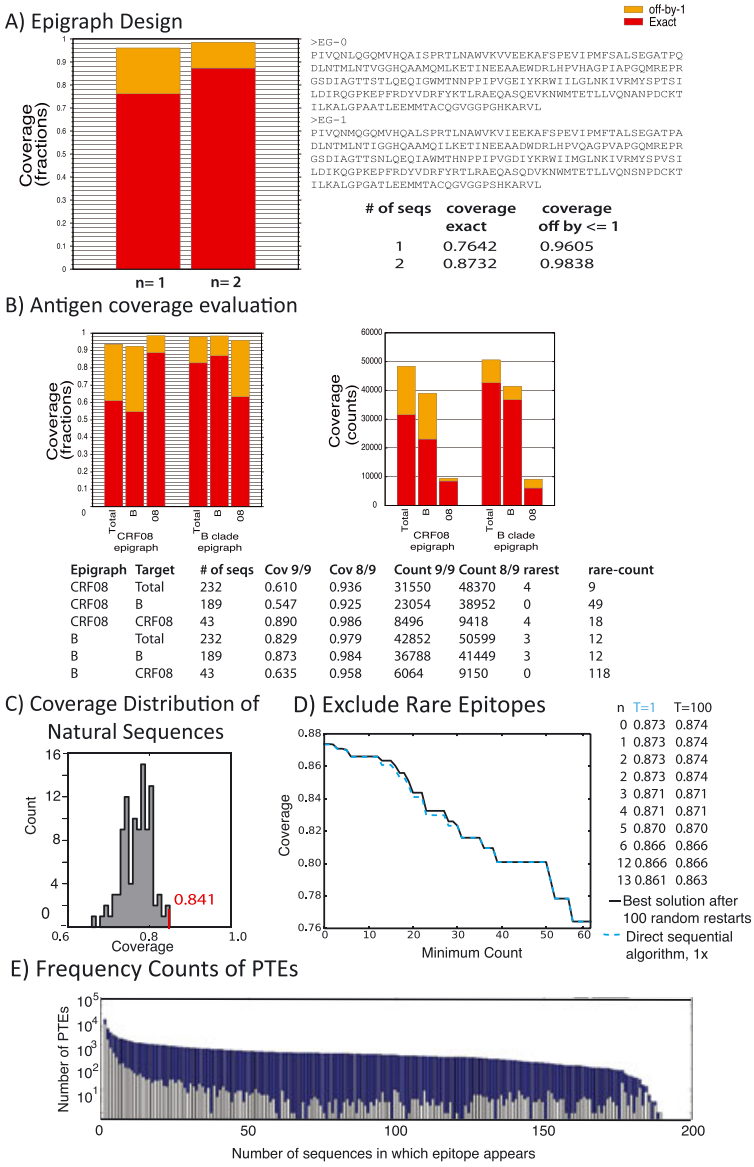


Figure 6. Summary output of the epigraph tool suite using 189 B-clade Gag p24 protein viruses from the USA as an example set. (A) Design output, including the sequences for two-antigen vaccine optimized for 9-mer epitope coverage, a table of coverage information, and a graphic comparing the coverage of the first epigraph (EG-0) to the additional coverage obtained by the addition of the complementary epigraph (EG-1). The red bar measures perfectly matched epitopes, while the orange bar adds the off-by-one matches. (B) Antigen coverage evaluation, showing the fraction of epitope matches between epigraph vaccines and two populations of viral sequences: US B clade and the Chinese CRF08 (labeled 08) recombinant lineage. Shown are fraction of covered epitopes (left) and the raw counts (right). The counts for the CRF08 are much reduced because only 48 CRF08 Gag sequences were available, while 189 B-clade sequences were included. (C) Coverage distribution is shown for 100 randomly selected pairs of natural p24 B-clade sequences. The best natural sequence coverage is 84%, marked in red. (D) exclude rare epitope output. This code sequentially explores the impact on coverage of excluding rare 9-mers, starting with excluding the rarest (those only found once) and stepping up the minimal count by 1 until a graph can no longer be completed (for this data set, that limit occurs at a count of 62). The blue line is a simple epigraph run, the black line is slightly improved and reflects the best scores given 100 random restarts. A complete table is given in the epigraph output, but here, we only track the values up to a minimum count of 13. The maximum coverage if all paths through the graph are available is 0.873. If the antigens are restricted to include only 9-mers that are found 12 or more times, the coverage is reduced by less than a percent, to 0.866. (E) A histogram graphic summarizes the number distinct 9-mers that are found in a given number of sequences, for the full Gag protein in the B US data set. Note that almost 10,000 completely unique 9-mers (i.e., only found in one sequence) are observed in this conserved protein in this relative conserved dataset. The few highly conserved peptides are summarized in Table II.

Published 2017. This article is a U.S. Government work and is in the public domain in the USA. Statistics in Medicine published by John Wiley & Sons Ltd.

Statist. Med. ***2018,*** 37 181–194

Table II. The most conserved Gag peptides in the US B-clade set were identified using the 'characterized PTEs' tool in the epigraph tool suite.

Start	Stop	Frequency	Peptide	Class I HLA presentation
35	43	183	VWASRELER	
			SPR[TL]LNAWVK[VW]	
148	156	181	SPRTLNAWV	B*8101, B*0702
149	157	187	PRTLNAWVK	
150	158	187	RTLNAWVKV	A*0201
164	172	181	FSPEVIPMF	B57, B58, B63
177	185	180	EGATPQDLN	
			MLNTVGGHQAAMQMLK	
187	195	179	MLNTVGGHQ	
188	196	179	LNTVGGHQA	human, unknown HLA
189	197	179	NTVGGHQAA	
190	198	179	TVGGHQAAM	
191	199	182	VGGHQAAMQ	
192	200	185	GGHQAAMQM	
193	201	185	GHQAAMQML	B*1510, B*3901, B38
194	202	184	HQAAMQMLK	B52, A11
			REPRGSDIAGTTS	
229	237	179	REPRGSDIA	
230	238	179	EPRGSDIAG	
231	239	184	PRGSDIAGT	
232	240	184	RGSDIAGTT	
233	241	184	GSDIAGTTS	
			GLNK[I-]VRMYSP	
269	277	181	GLNKIVRMY	B*1501, B62
270	278	182	LNKIVRMYS	
271	279	182	NKIVRMYSP	
			QGPKEPFRDYVDRFYK	
287	295	182	QGPKEPFRD	
288	296	182	GPKEPFRDY	B7
289	297	182	PKEPFRDYV	human, unknown HLA
290	298	182	KEPFRDYVD	
291	299	182	EPFRDYVDR	A*0201
			EPFRDYVDRFF	B81
292	300	182	PFRDYVDRF	human, unknown HLA
293	301	189	FRDYVDRFY	
			FRDYVDRFYK	B*1801, B27
294	302	183	RDYVDRFYK	
			VKNWMTETLL	
313	321	181	VKNWMTETL	B*4801
314	322	181	KNWMTETLL	
			WMTETLLVQN	
316	324	184	WMTETLLVQ	
317	325	185	MTETLLVQN	
			LEEMMTACQGVGGP	
343	351	182	LEEMMTACQ	human, unknown HLA
344	352	182	EEMMTACQG	human, unknown HLA
345	353	182	EMMTACQGV	A*0201
346	354	184	MMTACQGVG	human, unknown HLA
347	355	184	MTACQGVGG	human, unknown HLA
348	356	184	TACQGVGGP	

This generated the data in the frequency and peptide columns. The HIV database sequence locator tool [43] generated positions for each of the peptides, relative to the HIXB2 reference strain. The HIV database ELF tool [44] identified the known database epitopes within the set.

PTEs, potential T-cell epitopes; ELF, Epitope Location Finder; HIV, human immunodeficiency virus; HLA, human leukocyte antigen.

4.2. *Epigraph input and output*

The epigraph tool is organized into multiple tabs, and essential elements of the output of each page are shown in Figure 6. The first tool is simply the epigraph design page. A set of diverse protein sequences are provided by the user, and epigraph produces sequences that optimize epitope coverage are produced. A graphic showing the coverage values relative to the input population is created (Figure 6A). To compare epigraphs to other vaccines, or to explore their cross-reactive potential in other populations, one can use the antigen evaluation tool (Figure 6B), where the frequency of matched k-mers between the vaccine and sequence set can be evaluated. Both the epitope coverage and the number of the rarest epitopes in the vaccine are summarized. In Figure 6B, we compare two two-antigen epigraph vaccines, one based on the US B-clade sequence described earlier, and based on the CRF08 recombinant form that is common in China. One can see the dramatic drop off in epitope coverage when one evaluates the US B-clade epigraphs against CRF08 sequences and vice versa. Figure 6C illustrates the coverage if two natural strains are used instead of an epigraph pair. One hundred sets of two natural strains were selected randomly from the US B-clade set, and coverage of the population of US B-clade viruses was evaluated. If one deliberately selects the natural strains with the best coverage, the potential for cross-reactive vaccine responses can be greatly enhanced over typical choices that might be made for reasons of history or convenience. Epigraphs provide better coverage than the best two natural strains, even in the context (Figure 6A vs. C) of Gag p24, one of the most conserved proteins in HIV. Epigraph has a tool that enables a user to weigh the benefit of excluding rare epitopes against the cost of excluding these epitopes on total coverage (Figure 6D).

Figure 6E illustrates a newly added feature of the epigraph tool suite, introduced for this paper: frequency counts of potential epitopes. Tables are produced that enable users to resolve the frequency of each unique k-mer in their input data (essentially this summarizes the data that is encapsulated in the nodes of the graph), either overall, or by position in an alignment. In the case of the B-clade US HIV Gag sequences, this ranges from a very large number of unique 9-mers that are only sampled once in the population, to a very small number of highly conserved 9-mers that are present in nearly all of the sequences in the dataset. A graphic is produced to illustrate this distribution (Figure 6E). This data can be piped through additional HIV database tools to enhance its usefulness. We provide an example in Table II. Here, for the purpose of illustration, we take all 9-mers that are present in $\geqslant$95% of the US B-clade input data, identifying the most highly conserved epitopes in this data set. This peptide set is then piped through the HIV database sequence locator tool [43], to identify the positions of each conserved 9-mer in Gag. The data is sorted by position, and overlapping 9-mers are combined to define larger highly conserved regions. Each conserved region is then piped through the HIV database Epitope Location Finder tool [44] to identify known epitopes within the region, and their HLA presenting proteins. Epitope Location Finder also enables epitope predictions (data not shown). All of this data is brought to produce Table II. Such a table could provide a useful baseline for interpreting protective immune responses, or for informing peptide vaccine designs.

5. Discussion

Epigraph produces vaccine immunogen cocktail designs that include intact proteins, so they can be delivered using the same strategies as natural proteins. Because of this, epitopes will be naturally processed and presented, and large stretches of proteins with many potential epitopes and many HLA presentation motifs are included in epigraph vaccines. In addition to the design of full-length proteins, there is also an interest in focusing vaccine-stimulated T-cell responses exclusively on conserved regions, to shift immunodominance to epitopes under fitness constraints [45–47]. A variant on this theme is to consider coevolutionary patterns and fitness landscapes and to restrict vaccine design to inclusion of epitopes that require compensatory mutations to escape [48]. Either way, responses to epitopes with high fitness costs may be beneficial in either a preventative or therapeutic setting, but delivery is challenging. HIV peptide vaccines comprised of concatenated epitopes have been essentially immunologically silent when tested in human trials [49, 50]. For this reason, when designing epigraphs and mosaics that conserved epitopes presumed to have a high viral fitness cost associated with escape, we have focused on the use of conserved regions rather than conserved epitopes, so that local protein context is preserved to enable more natural epitope processing. To date, our conserved region HIV mosaic designs are immunogenic in animals studies [17, 18]. Of note, and some concern, responses to conserved region epitopes were stronger when they were embedded in a complete protein than when they were isolated in conserved region fragments [17]. We have recently used epigraph to design a conserved region pan-filovirus vaccine [8], and

immunogenicity testing is currently underway. As an alternative to conserved regions, recent exploration of the use dendritic cells for peptide presentation to T cells may offer a valuable alternative strategy for presenting short peptides to the immune system [51]. Epigraphs could also be helpful for the design of peptide variants in such a system.

Cytomegalovirus vectors have sparked great recent interest T-cell-based vaccines. SIV proteins presented in CMV vectors generate prolific T-cell responses, which enable control and clearance of pathogenic SIV upon infection in over 50% of vaccinated monkeys; thus, CMV-vectored HIV vaccines may have both therapeutic and prophylactic potential [40, 41]. CMV responses in rhesus macaques have advanced the field into new areas of T cell-mediated immunity that we are just beginning to understand, and provided new impetus for exploring T-cell-mediated vaccine approaches [40–42]. Pairing epigraphs inserts with CMV vectors may be able to enhance the cross-reactive potential of the responses, for both preventative and therapeutic vaccine applications [8].

In summary, epigraphs employ a graph-based approach to design antigens for vaccine cocktails against highly variable pathogens. Epigraphs provide significantly improved runtimes over the first generation mosaic vaccine design strategies (seconds to minutes, typically, for epigraph versus hours to days for mosaics). This enhanced computational efficiency allows users to explore aspects of vaccine design that were previously intractable. Epigraph requires that the epitope graph be acyclic, and we have here explored the outcome of using different heuristics for removing cycles from graphs. While some trends were noted, we have found that the optimal heuristic choice varies from one input data set to another. We have also provided more detailed explanations of the basic input and output of the epigraph tool suite.

Acknowledgements

We thank Klaus Früh and Louis Picker, whose innovative suggestions for tailored therapeutic CMV vaccine antigen design initially spurred the development of the epigraph concept. We thank Hyejin Yoon for web implementation; Aric Hagberg, Misha Chertkov, and Diane Oyen for useful discussions about graph-theoretic algorithms; and Amanda Ziemann for valuable comments on the manuscript.

This work was funded by the National Institutes of Health NIH R01-AI100343, and by the Bill and Melinda Gates Foundation Global Health Proposal OPP1108533.

References

1. Fischer W, Ganusov VV, Giorgi EE, Hraber PT, Keele BF, Leitner T, Han CS, Gleasner CD, Green L, Lo CC, Nag A, Wallstrom TC, Wang S, McMichael AJ, Haynes BF, Hahn BH, Perelson AS, Borrow P, Shaw GM, Bhattacharya T, Korber BT. Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS ONE* 2010; **5**(8):e12 303.
2. Liu MK, Hawkins N, Ritchie AJ, Ganusov VV, Whale V, Brackenridge S, Li H, Pavlicek JW, Cai F, Rose-Abrahams M, Treurnicht F, Hraber P, Riou C, Gray C, Ferrari G, Tanner R, Ping LH, Anderson JA, Swanstrom R, Cohen M, Karim SS, Haynes B, Borrow P, Perelson AS, Shaw GM, Hahn BH, Williamson C, Korber BT, Gao F, Self S, McMichael A, Goonetilleke N. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *Journal of Clinical Investigation* 2013; **123**(1):380–393.
3. Bar KJ, Tsao CY, Iyer SS, Decker JM, Yang Y, Bonsignori M, Chen X, Hwang KK, Montefiori DC, Liao HX, Hraber P, Fischer W, Li H, Wang S, Sterrett S, Keele BF, Ganusov VV, Perelson AS, Korber BT, Georgiev I, McLellan JS, Pavlicek JW, Gao F, Haynes BF, Hahn BH, Kwong PD, Shaw GM. Early low-titer neutralizing antibodies impede HIV-1 replication and select for virus escape. *PLoS Pathogens* 2012; **8**(5):e1002 721.
4. Liao HX, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, Fire AZ, Roskin KM, Schramm CA, Zhang Z, Zhu J, Shapiro L, NISC Comparative Sequencing Program, Mullikin JC, Gnanakaran S, Hraber P, Wiehe K, Kelsoe G, Yang G, Xia SM, Montefiori DC, Parks R, Lloyd KE, Scearce RM, Soderberg KA, Cohen M, Kamanga G, Louder MK, Tran LM, Chen Y, Cai F, Chen S, Moquin S, Du X, Joyce MG, Srivatsan S, Zhang B, Zheng A, Shaw GM, Hahn BH, Kepler TB, Korber BT, Kwong PD, Mascola JR, Haynes BF. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* 2013; **496**(7446):469–476.
5. Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, Novitsky V, Haynes B, Hahn BH, Bhattacharya T, Korber B. Diversity considerations in HIV-1 vaccine selection. *Science* 2002; **296**(5577):2354–60.
6. Hulot SL, Korber B, Giorgi EE, Vandergrift N, Saunders KO, Balachandran H, Mach LV, Lifton MA, Pantaleo G, Tartaglia J, Phogat S, Jacobs B, Kibler K, Perdiguero B, Gomez CE, Esteban M, Rosati M, Felber BK, Pavlakis GN, Parks R, Lloyd K, Sutherland L, Scearce R, Letvin NL, Seaman MS, Alam SM, Montefiori D, Liao HX, Haynes BF, Santra S. Comparison of immunogenicity in rhesus macaques of transmitted-founder, HIV-1 group M consensus, and trivalent mosaic envelope vaccines formulated as a DNA prime, NYVAC, and envelope protein boost. *Journal of Virology* 2015; **89**(12):6462–6480.
7. Fischer W, Perkins S, Theiler J, Bhattacharya T, Yusim K, Funkhouser R, Kuiken C, Haynes B, Letvin NL, Walker BD, Hahn BH, Korber BT. Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nature Medicine* 2007; **13**:100–106.
8. Theiler J, Yoon H, Yusim K, Picker LJ, Frueh K, Korber B. Epigraph: a vaccine design tool applied to an HIV therapeutic vaccine and a pan-filovirus vaccine. *Scientific Reports* 2016; **6**:33 987.

9. Li F, Malhotra U, Gilbert PB, Hawkins NR, Duerr AC, McElrath JM, Corey L, Self SG. Peptide selection for human immunodeficiency virus type 1 CTL-based vaccine evaluation. *Vaccine* 2006; **24**(47-48):6893–6904.
10. Korber B, Letvin NL, Haynes BF. T-cell vaccine strategies for human immunodeficiency virus, the virus with a thousand faces. *Journal of Virology* 2009; **83**(17):8300–14. DOI:10.1128/JVI.00114-09.
11. Barouch D, O'Brien K, Simmons N, King S, Abbink P, Maxfield L, Sun Y, La Porte A, Riggs A, Lynch D, Clark S, Backus K, Perry J, Seaman M, Carville A, Mansfield K, Szinger J, Fischer W, Muldoon M, Korber B. Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys. *Nature Medicine* 2010; **16**:319–323.
12. Barouch D, Stephenson K, Borducchi E, Smith K, Stanley K, McNally A, Liu J, Abbink P, Maxfield L, Seaman M, Dugast A, Alter G, Ferguson M, Li W, Earl P, Moss B, Giorgi E, Szinger J, Eller L, Billings E, Rao M, Tovanabutra S, Sanders-Buell E, Weijtens M, Pau M, Schuitemaker H, Robb M, Kim J, Korber B, Michael N. Protective efficacy of a global HIV-1 mosaic vaccine against heterologous SHIV challenges in rhesus monkeys. *Cell* 2013; **155**:531–539.
13. Ndhlovu ZM, Piechocka-Trocha A, Vine S, McMullen A, Koofhethile K C, Goulder PJ, Ndung'u T, Barouch DH, Walker BD. Mosaic HIV-1 Gag antigens can be processed and presented to human HIV-specific CD8+ T cells. *Journal of Immunology* 2011; **186**(12):6914–6924.
14. Santra S, Liao H, Zhang R, Muldoon M, Watson S, Fischer W, Theiler J, Szinger J, Balachandran H, Buzby A, Quinn D, Parks RJ, Tsao C, Carville A, Mansfield K, Pavlakis G, BK BF, BF BH, Korber B, Letvin N. Mosaic vaccines elicit CD8+ T lymphocyte responses that confer enhanced immune coverage of diverse HIV strains in monkeys. *Nature Medicine* 2010; **16**(3):324–8.
15. Santra S, Muldoon M, Watson S, Buzby A, Balachandran H, Carlson KR, Mach L, Kong WP, McKee K, Yang ZY. Breadth of cellular and humoral immune responses elicited in rhesus monkeys by multi-valent mosaic and consensus immunogens. *Virology* 2012; **428**(2):121–127.
16. Abdul-Jawad S, Ondondo B, Gardner A, van Hateren A, Elliott T, Korber B, Hanke T. Increased valency of conserved-mosaic vaccines enhances the breadth and depth of epitope recognition. *Molecular Therapy* 2016; **24**(2):375–384.
17. Stephenson KE, SanMiguel A, Simmons NL, Smith K, Lewis MG, Szinger JJ, Korber B, Barouch DH. Full-length HIV-1 immunogens induce greater magnitude and comparable breadth of T lymphocyte responses to conserved HIV-1 regions compared with conserved-region-only HIV-1 immunogens in rhesus monkeys. *Journal of Virology* 2012; **86**(21): 11 434–11 440.
18. Ondondo B, Murakoshi H, Clutton G, Abdul-Jawad S, Wee EG, Gatanaga H, Oka S, McMichael AJ, Takiguchi M, Korber B, Hanke T. Novel conserved-region T-cell mosaic vaccine with high global HIV-1 coverage is recognized by protective responses in untreated infection. *Molecular Therapy* 2016; **24**(4):832–842.
19. Kamlangdee A, Kingstad-Bakke B, Anderson TK, Goldberg TL, Osorio JE. Broad protection against avian influenza virus by using a modified vaccinia Ankara virus expressing a mosaic hemagglutinin gene. *Journal of Virology* 2014; **88**(22):13 300–13 309.
20. Yusim K, Dilan R, Borducchi E, Stanley K, Giorgi E, Fischer W, Theiler J, Marcotrigiano J, Korber B, Barouch DH. Hepatitis C genotype 1 mosaic vaccines are immunogenic in mice and induce stronger T-cell responses than natural strains. *Clinical and Vaccine Immunology* 2013; **20**(2):302–305.
21. Fenimore PW, Muhammad MA, Fischer WM, Foley BT, Bakken RR, Thurmond JR, Yusim K, Yoon H, Parker M, Hart MK, Dye JM, Korber B, Kuiken C. Designing and testing broadly-protective filoviral vaccines optimized for cytotoxic T-lymphocyte epitope coverage. *PLoS ONE* 2012; **7**(10):e44 769.
22. Badamchi-Zadeh A, McKay PF, Korber BT, Barinaga G, Walters AA, Nunes A, Paulo Gomes J, Follmann F, Tregoning JS, Shattock RJ. Multi-component prime-boost vaccination regimen with a consensus MOMP antigen enhances *Chlamydia trachomatis* clearance. *Frontiers in Immunology* 2016; **7**:162.
23. Gag CTL/CD8+ Epitope Map. http://www.hiv.lanl.gov/content/immunology/maps/ctl/Gag.html (Date of access: 11/08/2016).
24. De Bruijn NG. A combinatorial problem. *Proceedings of Koninklijke Nederlandse Akademie van Wetenschappen. Series A* 1946; **49**(7):758–764.
25. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. *Proceedings of National Academy of Science* 2001; **98**:9748–9753.
26. Compeau PEC, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology* 2011; **29**:987–991.
27. Smith TF, Waterman MS. Identification of common molecular subsequences. *Journal of Molecular Biology* 1981; **147**: 195–197.
28. Giegerich R. A systematic approach to dynamic programming in bioinformatics. *Bioinformatics* 2000; **16**:665–677.
29. Garey M, Johnson D. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman: New York, 1979.
30. Chen J, Liu Y, Lu S, O'Sullivan B, Razgon I. A fixed-parameter algorithm for the directed feedback vertex set problem. *Journal of the ACM* 2008; **55**:21:1–21:19. DOI:10.1145/1411509.1411511.
31. Depth-first search and linear graph algorithms. *SIAM Journal of Computing* 1972; **1**:146–160.
32. Epigraph tool suite. http://www.hiv.lanl.gov/content/sequence/EPIGRAPH/epigraph.html (Date of access: 15/06/2016).
33. HIV alignments. http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html (Date of access: 17/02/2016).
34. Williamson AL, Rybicki EP. Justification for the inclusion of Gag in HIV vaccine candidates. *Expert Rev Vaccines* 2016; **15**(5):585–598.
35. Liu J, O'Brien KL, Lynch DM, Simmons NL, La Porte A, Riggs AM, Abbink P, Coffey RT, Grandpre LE, Seaman MS, Landucci G, Forthal DN, Montefiori DC, Carville A, Mansfield KG, Havenga MJ, Pau MG, Goudsmit J, Barouch DH. Immune control of an SIV challenge by a T-cell-based vaccine in rhesus monkeys. *Nature* 2009; **457**(7225):87–91.
36. Addo MM, Yu XG, Rathod A, Cohen D, Eldridge RL, Strick D, Johnston MN, Corcoran C, Wurcel AG, Fitzpatrick CA, Feeney ME, Rodriguez WR, Basgoz N, Draenert R, Stone DR, Brander C, Goulder PJ, Rosenberg ES, Altfeld M, Walker BD. Comprehensive epitope analysis of human immunodeficiency virus type 1 (HIV-1)-specific T-cell responses directed

against the entire expressed HIV-1 genome demonstrate broadly directed responses, but no correlation to viral load. *Journal of Virology* 2003; **77**(3):2081–2092.
37. Kaufmann DE, Bailey PM, Sidney J, Wagner B, Norris PJ, Johnston MN, Cosimi LA, Addo MM, Lichterfeld M, Altfeld M, Frahm N, Brander C, Sette A, Walker BD, Rosenberg ES. Comprehensive analysis of human immunodeficiency virus type 1-specific CD4 responses reveals marked immunodominance of Gag and Nef and the presence of broadly recognized peptides. *Journal of Virology* 2004; **78**(9):4463–4477.
38. Liu MK, Hawkins N, Ritchie AJ, Ganusov VV, Whale V, Brackenridge S, Li H, Pavlicek JW, Cai F, Rose-Abrahams M, Treurnicht F, Hraber P, Riou C, Gray C, Ferrari G, Tanner R, Ping LH, Anderson JA, Swanstrom R, CHAVI Core B, Cohen M, Karim SS, Haynes B, Borrow P, Perelson AS, Shaw GM, Hahn BH, Williamson C, Korber BT, Gao F, Self S, McMichael A, Goonetilleke N. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *Journal of Clinical Investigation* 2013; **123**(1):380–393.
39. McElrath MJ1, De Rosa SC, Moodie Z, Dubey S, Kierstead L, Janes H, Defawe OD, Carter DK, Hural J, Akondy R, Buchbinder SP, Robertson MN, Mehrotra DV, Self SG, Corey L, Shiver JW, Casimiro DR, Step Study Protocol Team. HIV-1 vaccine-induced immunity in the test-of-concept step study: a case-cohort analysis. *Lancet* 2008; **372**(9653): 1894–1905.
40. Hansen SG, Sacha JB, Hughes CM, Ford JC, Burwitz BJ, Scholz I, Gilbride RM, Lewis MS, Gilliam AN, Ventura AB, Malouli D, Xu G, Richards R, Whizin N, Reed JS, Hammond KB, Fischer M, Turner JM, Legasse AW, Axthelm MK, Edlefsen PT, Nelson JA, Lifson JD, Früh K, Picker LJ. Cytomegalovirus vectors violate CD8+ T cell epitope recognition paradigms. *Science* 2013; **340**(6135):1237 874.
41. Hansen SG, Piatak M Jr, Ventura AB, Hughes CM, Gilbride RM, Ford JC, Oswald K, Shoemaker R, Li Y, Lewis MS, Gilliam AN, Xu G, Whizin N, Burwitz BJ, Planer SL, Turner JM, Legasse AW, Axthelm MK, Nelson JA, Früh K, Sacha JB, Estes JD, Keele BF, Edlefsen PT, Lifson JD, Picker LJ. Immune clearance of highly pathogenic SIV infection. *Nature* 2013; **502**(7469):100–104.
42. Hansen SG, Wu HL, Burwitz BJ, Hughes CM, Hammond KB, Ventura AB, Reed JS, Gilbride RM, Ainslie E, Morrow DW, Ford JC, Selseth AN, Pathak R, Malouli D, Legasse AW, Axthelm MK, Nelson JA, Gillespie GM, Walters LC, Brackenridge S, Sharpe HR, López CA, Früh K, Korber BT, McMichael AJ, Gnanakaran S, Sacha JB, Picker LJ. Broadly targeted CD8 T cell responses restricted by major histocompatibility complex E. *Science* 2016; **351**(6274):714–720.
43. HIV sequence locator. http://www.hiv.lanl.gov/content/sequence/LOCATE/locate.html (Date of access: 16/08/2016).
44. Epitope location finder. http://www.hiv.lanl.gov/content/sequence/ELF/epitope_analyzer.html (Date of access: 16/08/2016).
45. Létourneau S, Im EJ, Mashishi T, Brereton C, Bridgeman A, Yang H, Dorrell L, Dong T, Korber B, McMichael AJ, Hanke H. Design and Pre-Clinical Evaluation of a Universal HIV-1 Vaccine Design and pre-clinical evaluation of a universal HIV-1 vaccine. *PLoS ONE* 2007; **2**(10):e984.
46. Kulkarni V, Valentin A, Rosati M, Alicea C, Singh AK, Jalah R, Broderick KE, Sardesai NY, Le Gall S, Mothe B, Brander C, Rolland M, Mullins JI, Pavlakis GN, Felber BK. Altered response hierarchy and increased T-cell breadth upon HIV-1 conserved element DNA vaccination in macaques. *PLoS ONE* 2014; **9**(1):e86 254.
47. Yang OO, Ali A, Kasahara N, Faure-Kumar E, Bae JY, Picker LJ, Park H. Short conserved sequences of HIV-1 are highly immunogenic and shift immunodominance. *Journal of Virology* 2015; **89**(2):1195–1204.
48. Ferguson AL, Mann JK, Omarjee S, Ndung'u T, Walker BD, Chakraborty AK. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* 2013; **38**(3):606–617.
49. Jaoko W, Nakwagala FN, Anzala O, Manyonyi GO, Birungi J, Nanvubya A, Bashir F, Bhatt K, Ogutu H, Wakasiaka S, Matu L, Waruingi W, Odada J, Oyaro M, Indangasi J, Ndinya-Achola J, Konde C, Mugisha E, Fast P, Schmidt C, Gilmour J, Tarragona T, Smith C, Barin B, Dally L, Johnson B, Muluubya A, Nielsen L, Hayes P, Boaz M, Hughes P, Hanke T, McMichael A, Bwayo J, Kaleebu P. Safety and immunogenicity of recombinant low-dosage HIV-1 A vaccine candidates vectored by plasmid pTHr DNA or modified vaccinia virus Ankara (MVA) in humans in East Africa. *Vaccine* 2008; **26**(22):2788–2795.
50. Gorse GJ, Baden LR, Wecker M, Newman MJ, Ferrari G, Weinhold KJ, Livingston BD, Villafana TL, Li H, Noonan E, Russell ND. Safety and immunogenicity of cytotoxic T-lymphocyte poly-epitope, DNA plasmid (EP HIV-1090) vaccine in healthy, human immunodeficiency virus type 1 (HIV-1)-uninfected adults. *Vaccine* 2008; **26**(2):215–223.
51. Macatangay BJ, Riddler SA, Wheeler ND, Spindler J, Lawani M, Hong F, Buffo MJ, Whiteside TL, Kearney MF, Mellors JW, Rinaldo CR. Therapeutic vaccination with dendritic cells loaded with autologous HIV type 1-infected apoptotic cells. *Journal of Infectious Diseases* 2016; **213**(9):1400–1409.