SUMMARY

Aspects of the disclosure provide a method and system for selecting a set of candidate elements for inclusion in a vaccine such that the likelihood that every member of a population has a positive response to the vaccine is maximized.

According to an aspect of the present disclosure, there is provided a computer-implemented method of selecting one or more amino acid sequences for inclusion in a vaccine from a set of predicted immunogenic candidate amino acid sequences, the method comprising: identifying an immune profile response value for each candidate amino acid sequence in respect of each one of a plurality of sample components of an immune profile, wherein the immune profile response value represents whether the candidate amino acid sequence results in an immune response for the sample component of an immune profile; retrieving a plurality of immune profiles for a population; generating a plurality of representative immune profiles for the population, wherein the representative immune profiles overlap with the sample components of an immune profile; and, selecting the one or more amino acid sequences for inclusion in the vaccine that

minimises a likelihood of no immune response for each representative immune profile, based on the immune profile response values,

wherein the step of selecting comprises applying a mathematical optimisation algorithm to minimise a maximum likelihood of no immune response for each of the representative immune profiles, and

the immune profile comprises a set of HLA alleles and the sample components of an immune profile comprise sample HLA alleles, and wherein the variables of the mathematical optimisation algorithm comprise:

(a) a binary indicator variable for each candidate amino acid sequence which indicates whether the candidate amino acid is included in a vaccine;

(b) a continuous variable for each representative immune profile which gives a log likelihood of no immune response;

(c) a continuous variable for each sample component of an immune profile which gives a log likelihood of no response; and,

(d) a continuous variable which gives a maximum log likelihood that any representative immune profile does not respond to the selected one or more amino acid sequences,

wherein the mathematical optimisation algorithm minimises the continuous variable which gives a maximum log likelihood that any representative immune profile does not respond to the selected one or more amino acid sequences.

Advantageously the proposed approach explicitly accounts for and optimizes with respect to a wide variety of components the make up an immune profile, in contrast to the approaches of the state of the art, and maximises the chances of a vaccine being a success across a given population. Where the population is representative of the global population, the approach can be considered to lead toward an optimal, universal vaccine, that is, that the chances of an immune response being caused by the combination of vaccine elements included in the

vaccine is maximised. For example, where the sample components are a plurality of sample HLA alleles, the proposed approach explicitly accounts for and is optimized with respect to all alleles.

In sum, the method of the above aspect of the disclosure formulates a vaccine design with respect to a specific population as an optimization problem in which the goal is to maximize the likelihood of response of each citizen.

The present technique may be thought of as an allele-based approach; however, unlike the methodology of the art, the current approach considers individual citizens rather than looking at the most frequently occurring alleles in a population and seeking to provide an average across that set. We note that in the art population coverage describes the fraction of a population for which the epitope based vaccine is theoretically effective.

The predicted immunogenic candidate amino acid sequences may be short or long peptide sequences, where a long peptide sequence may include multiple short peptide sequences. The set of predicted immunogenic candidate amino acid sequences are typically retrieved from a prediction engine which computes some sort of a score that a peptide will result in some immune response (e.g., binding, presentation, cytokine release, etc.). Examples of publically available databases and tools that may be used for such predictions include the Immune Epitope Database (IEDB) (https://www.iedb.org/), the NetMHC prediction tool (http://www.cbs.dtu.dk/services/NetMHC/) and the NetChop prediction tool (http://www.cbs.dtu.dk/services/NetChop/). Other techniques are disclosed in WO2020/070307 and WO2017/186959.

The score from the prediction engine associated with each sequence may be used to identify the immune response value. Alternatively the immune response value may be retrieved from a database populated using data in previous literature, for example, by extracting univariate response statistics.

The one or more predicted candidate amino acid sequences may be of a fixed length or of variable lengths. For example, when considering MHC Class I HLA alleles, epitope lengths of 8, 9, 10, 11 and 12 amino acids may be candidates and when considering MHC Class II HLA alleles, each epitope is typically 15 amino acids in length. Alternatively, the candidate amino acid sequences may be groups of sequences. Example, candidate amino acid sequences include: (1) short

peptide sequences, such as 9-mer amino acid sequences; (2) long peptide sequences, such as 27-mer amino acid sequence which may be based on a short peptide sequence and include flanking regions; (3) longer amino acid sequences which may include multiple short peptide sequences as well as the intervening, naturally-occurring sequence; and (4) entire protein sequences.

The step of selecting the one or more amino acid sequences for inclusion in the vaccine may also be based on a correspondence between the sample components of an immune profile and the components of the immune profile present in the respective representative immune profiles.

In certain examples the immune profile may comprise one or more selected from a group comprising: a set of HLA alleles; presence (or absence) of tumor infiltrating lymphocytes; presence (or absence) of immune checkpoint markers, such as PD1, PD-L1, or CTLA4; presence (or absence) of hypoxia markers, such as HIF-1a or BNIP3; presence (or absence) of chemokine receptors such as CXCR4, CXCR3, and CX3CR1; and, previous infection by human papillomavirus. Each of these features has been shown to contribute, positively or negatively, to the immune response of a particular epitope, or candidate vaccine element. Thus the immune response value associated with each candidate amino acid sequence may represent the contribution of how likely that candidate sequence is to produce an immune response with the particular variables in question.

In specific examples the sample components of an immune profile comprise a sample HLA allele, such that the immune profile response value comprises an HLA allele immune response value for each candidate amino acid sequence in respect of each one of a plurality of sample HLA alleles. The immune profiles for a population may comprise a plurality of HLA genotypes for a population. The step of generating a plurality of representative immune profiles may comprise generating a plurality of representative sets of HLA alleles for the population. The HLA alleles of the representative sets may overlap with the sample HLA alleles.

The sample HLA alleles of the immune profile may be a set of most frequently occurring alleles in a population or all alleles of a population. A degree of overlap

between the sample HLA alleles and the representative immune profiles may include: (1) that all sample HLA alleles occur within at least one representative immune profile; and/or (2) that all HLA alleles of the representative immune profiles occur within the sample HLA alleles. Preferably at least one allele for each representative immune profile needs to be in the set of sample HLA alleles. Preferably each of the sample HLA alleles should be present in at least one of the representative sets. Similar variations in degrees of overlap are contemplated between the components of the immune profile and the representative immune profiles.

In implementations, the candidate amino acid sequences are vaccine elements and each representative set is a simulated citizen of a given population.

The method may further comprise retrieving a set of predicted immunogenic candidate amino acid sequences. The retrieval may be from a local memory, database or remote data repository.

In preferred examples, the step of generating comprises: (i) creating a first distribution over the plurality of immune profiles; and, (ii) sampling the first distribution to create the plurality of representative immune profiles. In examples, the immune profiles may comprise HLA genotypes.

More preferably, the first distribution is a distribution over the plurality of immune profiles for each region of the population.

Each region may be a population group having an ethnic population group (e.g. Caucasian, Africa, Asian) or a geographical population group (e.g. Lombardy, Wuhan).

Even more preferably, the first distribution is a posterior distribution over genotypes in each region based on a prior distribution and observed genotypes from the plurality of immune profiles in each region of the population.

In certain specific implementations, the first distribution is a symmetric Dirichlet distribution, wherein the method further comprises the step of collecting all

The mathematical optimisation algorithm may be constrained by one or more predetermined thresholds. In examples, the amino acid sequences may be selected based on a particular vaccine delivery platform.

Typical algorithms may struggle with such computational complexity and so to provide efficiencies and improvements, the method may be configured to provide one or more surrogate variables for the mathematical optimisation algorithm. The surrogate variables may comprise a log likelihood of no response for a representative set. In a specific preferred implementation, variables of the mathematical optimisation algorithm comprise: (a) a binary indicator variable for each candidate amino acid sequence which indicates whether the candidate amino acid is included in a vaccine; (b) a continuous variable for each representative immune profile which gives a log likelihood of no immune response; (c) a continuous variable for each sample component which gives a log likelihood of no response; and, (d) a continuous variable which gives a maximum log likelihood that any representative immune profile does not respond to the selected one or more amino acid sequences, wherein the mathematical optimisation algorithm minimises the continuous variable which gives a maximum log likelihood that any representative immune profile does not respond to the selected one or more amino acid sequences.

Accordingly, in a certain examples, the immune profile may comprise a set of HLA alleles and the sample components of an immune profile may comprise sample HLA alleles. In these examples, optionally the variables of the mathematical optimisation algorithm may comprise: (a) a binary indicator variable for each candidate amino acid sequence which indicates whether the candidate amino acid is included in a vaccine; (b) a continuous variable for each representative immune profile which gives a log likelihood of no immune response; (c) a continuous variable for each sample component of an immune profile which gives a log likelihood of no response; and, (d) a continuous variable which gives a maximum log likelihood that any representative immune profile does not respond to the selected one or more amino acid sequences, wherein the mathematical optimisation algorithm minimises the continuous variable which gives a maximum

log likelihood that any representative immune profile does not respond to the selected one or more amino acid sequences.

An objective of the mathematical optimisation algorithm is to minimize variable (d). In examples, the setting of the binary variables corresponds to the optimal choice of amino acid sequences for the given population. Advantageously the mathematical optimisation algorithm is a mixed integer linear program.

In this way the optimisation can take advantages of the benefit of such programming since the decisions are binary, i.e. whether or not to include an amino acid sequence in the vaccine.

Choosing amino acid sequences for inclusion in a vaccine is not an unlimited exercise and selection is preferably constrained in some way. Preferably, the method further comprises: assigning a cost to each candidate amino acid sequence, wherein the step of selecting is constrained based on the cost assigned to each candidate amino acid sequence, such that the selected one or more amino acid sequences have a total cost below a predetermined threshold budget.

Accordingly, an amount of amino acid sequences to be included in the vaccine can be selected based on the practical realities of the chosen vaccine platform and the vaccine delivery method. Additionally, or alternatively, the step of selecting is constrained based on a maximum amount of amino acid sequences allowed in a vaccine delivery platform.

Optionally, this may be performed by assigning a cost of 1 to each amino acid sequence and a budget according to the number of amino acid sequences that can be included in the vaccine.

In addition to being considered an allele-based approach, a proposed example may also be thought of as a graph-based approach in which, the method further comprises creating a tripartite graph, wherein: a first set of nodes corresponds to the candidate amino acid sequences; a second set of nodes corresponds to the sample components of an immune profile; and, a third set of nodes corresponds to the representative immune profiles for the population, and

wherein: weights of edges between the first set of nodes and the second set of nodes are the immune response values; and, weights of edges between the second set of nodes and the third set of nodes represent correspondence between the sample components and each representative immune profile.

Thus the implementation may be thought of as a network flow problem through the graph in which a minimax problem is handled with the goal of choosing a set of vaccine elements which minimize the log likelihood of no response for each hypothetical citizen. Conventional graph-based approaches do not consider the population HLA background.

In preferred examples the immune response value is a log likelihood value based on amino acid sub-sequences of the candidate amino acid sequence.

The vaccine design approach is applicable for any approach which assigns a value for a log likelihood. Most short peptide prediction engines compute some sort of a score that a peptide will result in some immune response (e.g., binding, presentation, cytokine release, etc.), and this score generally takes into account a specific HLA allele. In some cases, this is already a probability, and in others, it can be converted into a probability using a transformation function, such as a logistic function. Additionally, the step of identifying comprises selecting a best likelihood value as the immune response value from a likelihood value for each amino-acid subsequence.

Thus, where the candidate amino acid sequences comprise multiple peptide sequences, the likelihood values can be determined based on a score for each short peptide sequence that goes into a long or longer peptide sequence.

In particularly preferred examples the one or more candidate amino acid sequences are comprised in one or more proteins of a coronavirus, preferably the SARS-CoV-2 virus.

In this way the approach is suitable for providing a universal, optimised vaccine design across a population of interest for the SARS-CoV-2 virus. In examples, the one or more candidate amino acid sequences may be one or more of the Spike

(S) protein, Nucleoprotein (N), Membrane (M) protein and Envelope (E) protein of a virus, as well as open reading frames, such as orf1ab. Thus, the method of the present invention may be applied to an entire virus proteome. This is particularly beneficial for the identification of candidate elements for vaccine design.

The method may further comprise synthesising one or more selected amino acid sequences.

The method may further comprise encoding the one or more selected amino acid sequences into a corresponding DNA or RNA sequence. Further, the method may comprise incorporating the DNA or RNA sequence into a genome of a bacterial or viral delivery system to create a vaccine.

Thus, according to an example there is provided a method of creating a vaccine, comprising: selecting one or more amino acid sequences for inclusion in a vaccine from a set of predicted immunogenic candidate amino acid sequences by a method according to any of the above aspects; and synthesising the one or more amino acid sequences or encoding the one or more amino acid sequences into a corresponding DNA or RNA sequence and/or incorporating the DNA or RNA sequence into a genome of a bacterial or viral delivery system to create a vaccine.

According to a further example there may be provided a computer-implemented method of selecting one or more amino acid sequences for inclusion in a vaccine from a set of predicted immunogenic candidate amino acid sequences, the method comprising: retrieving a set of predicted immunogenic candidate amino acid sequences; identifying an HLA allele immune response value for each candidate amino acid sequence in respect of each one of a plurality of sample HLA alleles, wherein the HLA allele immune response value represents if the candidate amino acid sequence results in an immune response for the sample HLA allele; retrieving a plurality HLA genotypes for a population; generating a plurality of representative sets of HLA alleles for the population, wherein the HLA alleles of the representative sets overlap with the sample HLA alleles; selecting the one or more amino acid sequences for inclusion in the

vaccine that minimises a likelihood of no immune response for each representative set of HLA alleles, based on the HLA allele immune response values and a correspondence between the sample HLA alleles and the HLA alleles present in the respective representative set of HLA alleles.

In accordance with a further aspect of the disclosure there is provided a system for selecting one or more amino acid sequences for inclusion in a vaccine from a set of predicted immunogenic candidate amino acid sequences, the system comprising at least one processor in communication with at least one memory device, the at least one memory device having stored thereon instructions for causing the at least one processor to perform a method according to any of the above aspects.

In accordance with a further aspect of the disclosure there is provided a computer readable medium having computer executable instructions stored thereon for implementing the method of any of the above aspects.

BRIEF DESCRIPTION OF THE DRAWINGS

Embodiments will now be described in detail, by way of example only, with reference to the accompanying figures, in which:

Figure 1 shows a schematic of a tripartite graph according to examples of the invention;

Figure 2 shows a high-level flowchart of the proposed approach;

Figure 3 shows an alternative schematic of a tripartite graph according to examples of the invention;

Figure 4 shows an example output; and,

Figure 5 shows a method according to an embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

According to certain embodiments described herein there is proposed a method and system for selecting a small set of candidate elements for inclusion in a vaccine such that the likelihood that every member of a population has a positive response to the vaccine is maximized. Specifically, there is a focus on epitope-