A guide to current methodology and usage of reverse vaccinology towards *in silico* vaccine discovery

Stephen J. Goodswen[1], Paul J. Kennedy[2], John T. Ellis[1,*]

[1] School of Life Sciences, University of Technology Sydney, 15 Broadway, Ultimo, NSW 2007, Australia

[2] School of Computer Science, Faculty of Engineering and Information Technology and the Australian Artificial Intelligence Institute, University of Technology Sydney, 15 Broadway, Ultimo, NSW 2007, Australia

Correspondence: Emeritus Professor John T. Ellis

Email address: John.Ellis@uts.edu.au

One sentence summary: The authors first present an introduction to a computational process named reverse vaccinology to predict vaccine candidates, and then describe in detail an up-to-date workflow of this process that can be followed and/or adapted for any pathogen having a genome sequence.

Keywords: reverse vaccinology; *in silico* vaccine discovery; immunoinformatics; subtractive proteomics; computational vaccinology.

Abstract

Reverse vaccinology (RV) was described at its inception in 2000 as an *in silico* process that starts from the genomic sequence of the pathogen and ends with a list of potential protein and/or peptide candidates to be experimentally validated for vaccine development. Twenty-two years later, this process has evolved from a few steps entailing a handful of bioinformatics tools to a multitude of steps with a plethora of tools. Other *in silico* related processes with overlapping workflow steps have also emerged with terms such as subtractive proteomics, computational vaccinology, and immunoinformatics. From the perspective of a new RV practitioner, determining the appropriate workflow steps and bioinformatics tools can be a time consuming and overwhelming task, given the number of choices. This review presents the current understanding of RV and its usage in the research community as determined by a comprehensive survey of scientific papers published in the last seven years. We believe the current mainstream workflow steps and tools presented here will be a valuable guideline for all researchers wanting to apply an up-to-date *in silico* vaccine discovery process.

Introduction

In October 2000, a novel process for vaccine discovery was first described by Rino Rappuoli in a landmark publication (Rappuoli, 2000). The process was named 'reverse vaccinology' (RV) to encapsulate the idea that the vaccine discovery process started *in silico* (on a computer) using genetic information rather than in a laboratory with the pathogen itself. RV's overriding goal is to identify potential protein and/or peptide candidates to be experimentally validated for vaccine development i.e., the hope is that these identified candidates are immunogenic. It must be accepted, nonetheless, that the output from an *in silico* process is fundamentally informed predictions. Experimental validation is the only way to be certain a predicted candidate is immunogenic.

The first study (Pizza *et al.*, 2000) accredited to have followed the RV process essentially had only two RV-related steps: 1) identifying open reading frames (ORFs) in unassembled DNA sequence fragments that potentially encoded surface-exposed or exported proteins; and 2) a phylogenetic analysis to distinguish from the identified proteins those conserved in sequence across a range of target strains. Like many novel processes, RV has evolved greatly over the last 22 years since its inception. An RV-inspired study can now typically have a multitude of computational steps with a choice of hundreds of bioinformatics resources to perform these steps. Other *in silico* related processes have also emerged, namely **subtractive genomics** and **proteomics**, **computational vaccinology**, and **immunoinformatics** (see Glossary). Conceptual boundaries between RV and these latest processes have blurred. Nonetheless, all these novel processes play an important role in this revolutionary era of identifying vaccine candidates *in silico.*

Fig. 1 shows the rise in number of scientific publications with 'reverse vaccinology' in its title since 2000. The total number of publications over this 21 year period is 180. Supplementary Table S1 lists the 180 publications. The increasing interest in RV did not occur until 2015, with over 133 (74%) of the 180 publications released in the last seven years. RV's current importance is exemplified by its escalating application to the greatest global health crisis of our age, the coronavirus COVID-19 pandemic. Seven papers with RV in the title and focusing on 'COVID-19' were published in 2020-21 (as of October 2022, a further seven have been published). The aim of this review is to present the current status of RV and its usage as revealed in the 133 publications of the last seven years. We present a comprehensive guideline of the most commonly used workflow and bioinformatics programs, given the current RV status. We believe this guideline will be a valuable resource for all RV practitioners wanting to apply an up-to-date *in silico* vaccine discovery process.

Principles of classical reverse vaccinology

To fully appreciate this review, the reader requires an understanding of the RV principles and influences. For example, why is RV now a reality, where does RV fit within the conventional approach to vaccine discovery; and from an RV perspective, what vaccine types can be discovered, what pathogen components are most likely to induce an immune response, and what are the main immune system players. Figures 2-7 are now presented to answer these questions and provide an introduction to RV. Table 1 shows comparisons between the conventional approach and classical RV in terms of antigen types that can be discovered, and time and financial factors impacting the discovery. Additional information on RV can be found in three reviews published between 2015-16 that are specific to bacteria (Heinson *et al.*, 2015), viruses (Bruno *et al.*, 2015), and ticks (Lew-Tabor & Valle, 2016).

Overview of the reverse vaccinology workflow

Pathogenic (see Glossary) organisms are composed of thousands of proteins. The central RV aim is to narrow down this number leaving only the most worthwhile candidates for laboratory investigation. This aim is achieved by predicting or gathering protein characteristics that support or oppose candidacy using bioinformatics programs or accessing biological databases, respectively (these characteristics are described later in depth). Reverse vaccinology tools (i.e., bioinformatics programs and biological databases) can be executed or accessed via three modes: web servers, application programming interfaces (APIs) to access tools over the internet, and standalone (i.e., tools installed on local computer). Each mode has its advantages and disadvantages. Web servers are by far the easiest to use but have restrictions on input data size and constraints on parsing the output. Using only web servers is essentially a one step at a time manual workflow. APIs and standalone programs allow for automated high-throughput workflows but require programming and computer administration skills. Standalone has a further disadvantage in that its installation becomes outdated because

most programs and databases are incrementally updated. Note that not every RV tool provides all three modes of operation.

Most predicted characteristics by a bioinformatics program are assigned a score (e.g., a probability that the protein contains a signal peptide), and most database derived characteristics belong to classifications (e.g., a protein has a subcellular localization classification of 'extracellular'). These scores and classifications are used to select candidates. The two main selection methods applied are filtering and ranking. Filtering is a manual process performed by the RV practitioner. It involves a series of workflow steps with conditional rule-based tests applied consecutively to each protein's characteristic scores or classifications to retain or discard it from the next workflow step e.g., retain protein if signal peptide probability is greater than a 0.5 threshold, and discard protein if subcellular localization is cytoplasm. The order of tests and threshold values applied are at the discretion of the RV practitioner. Ranking aims to assign only one score collectively representing all predicted characteristics, with the highest scoring proteins considered the most worthy candidates. Ranking can be achieved using ML (see later '**Machine learning specific to reverse vaccinology**').

Ideally, the RV workflow ends with selected candidates being tested for their immunogenicity in a laboratory experiment or animal model. Typically, however, most RV studies due to budget or other resource constraints rely on *in silico* techniques to verify their candidates. For example, a vaccine formulation can be modelled and assessed in a simulated immune system (these techniques are discussed further later).

A typical reverse vaccinology workflow

We have collated statistics from a survey sought to capture the current status, patterns and trends of RV usage. The survey source was all scientific publications after 2014 containing

121 'reverse vaccinology' in the title. Although the total number of publication titles for this
122 period was 133, 43 publications were excluded from the survey (21 publications were
123 reviews and/or did not contain RV workflows, 16 were not accessible, five did not specify
124 RV programs, and one was a duplicate publication but with a different DOI. Supplementary
125 Table S1 lists these 43 publications and the reason for their exclusion). Therefore, the RV
126 workflows from 90 publications provided the survey data. These 90 publications are referred
127 to henceforth as 'latest publications'. Supplementary Table S1 lists the survey questions and
128 results. Fig. 8 shows a graphical RV snapshot providing a status overview**.**

129 The following RV workflow is compiled from the most common steps presented in the latest
130 publications. With this in mind, we make no judgement as to what steps should or should not
131 define the RV scope. The common steps collectively entail a filtering workflow to discover a
132 multi-epitope vaccine against a pathogenic bacterium. There are essentially four stages: input
133 data gathering and preparation, predicting proteins naturally exposed to the immune system
134 (classical RV), predicting epitopes (immunoinformatics), and vaccine candidate verification.
135 The most popular bioinformatics program and/or database resource to achieve each step is
136 shown bold in brackets. Table 2 lists the main output of RV interest and where to access the
137 program or resource. Supplementary Information S1 describes these programs, including type
138 of input and output. Fig. 9 shows a schematic of the typical RV workflow as derived from
139 latest publications.

140 *Stage #1 – input data gathering and preparation*
141 The essential input data to the workflow are protein sequences. Every available sequence
142 pertaining to every available strain from the target species are the ultimate input data to attain
143 a **conserved vaccine** (see Glossary). Data can be downloaded from resources such as the
144 National Center for Biotechnology Information (NCBI) (Agarwala *et al.*, 2018) and UniProt
145 Knowledgebase (UniProtKB) (Bateman *et al.*, 2021)). If protein sequences are not available,

then genome sequences are the workflow commencement data. Thereby, predicting genes encoded in genomes would be the first step followed by coding sequence (CDS) translations to protein sequences.

Given sequences representative of entire proteomes from multiple strains, the aim is to find conserved proteins and compile them in one set to represent the common (core) proteome of a species (**CD–HIT** (Li & Godzik, 2006)). Conserved proteins tend to play an essential function. Next step is to remove the following from the core proteome: proteins homologous to those of the vaccine recipient (**BlastP)**, allergenic (**AllerTOP** (Dimitrov *et al.*, 2014)) and toxic (**ToxinPred** (Gupta *et al.*, 2013)) proteins.

Stage #2 – predicting proteins naturally exposed to the immune system

There is no consensus order for the next steps but the broad aim is to determine which of the remaining core proteins (i.e., proteins that are non-redundant, non-homologous, non-allergenic, and non-toxic) are naturally exposed to the immune system. This can be achieved by predicting informative protein characteristics such as antigenicity (**VaxiJen** (Doytchinova & Flower, 2007)), subcellular localisation **(PSORTb** (Yu *et al.*, 2010)**)**, transmembrane domains (**TMHMM** (Krogh *et al.*, 2001)), signal peptides (**signalP** (Teufel *et al.*, 2022)), virulence (**VFDB** (Chen *et al.*, 2005)), adhesion (**SPAAN** (Sachdeva *et al.*, 2005)**)**, protein function (**Pfam** (Mistry *et al.*, 2021)), and physical and chemical (physicochemical) properties (**ProtParam** (E. *et al.*, 2005)). User defined criteria is applied to prediction values to select proteins for the immunoinformatics workflow stage.

Stage #3 – predicting epitopes (immunoinformatics)

Whether cellular and/or humoral immune responses are required for protection is dependent on the target species' pathogenicity and virulence. The key here from an RV perspective is whether helper T-lymphocytes (HTLs), cytotoxic T lymphocytes (CTLs), and B-cell epitopes

are required as the basis of the protective immune response. The immunoinformatics stage involves predicting the required epitopes residing on selected proteins e.g., on filtered proteins expected to be exposed to the immune system (CTLs: **IEDB-MHC-I Binding** and HTLs: **IEDB-MHC-II Binding (Vita *et al.*, 2019)**, and B-cell epitope: **BepiPred** (Jespersen *et al.*, 2017))**.** Predicted epitopes here are small lengths of amino acids (peptides) from the selected proteins. Promiscuous epitopes with high binding affinity and broad population coverage (**IEDB-Population coverage** (Bui *et al.*, 2006)) are selected from epitope-rich proteins. The selected epitopes are connected with suitable **linkers** (see Glossary) and adjuvants to construct one sequence that represents the multi-epitope vaccine candidate (i.e., vaccine construct).

Stage #4 – verifying vaccine construct candidates

The aim of the final workflow stage is to verify by computational means whether the vaccine construct is potentially immunogenic and safe, which in effect is attempting to determine how the construct, represented essentially as a one dimensional digital sequence, might interact in the 3D real-world. The immunoinformatics and this final verification stage are expected to be iterative with different combinations of vaccine construct candidates i.e., different combinations of CTLs, HTLs and B-cell epitopes. Each candidate is checked for antigenicity (**VaxiJen**), allergenicity (**AllerTOP**), toxicity (**ToxinPred**), solubility (**SOLpro** (Cheng *et al.*, 2005)) and stability (**ProtParam**). Candidates predicted to be antigenic, non-allergic, non-toxic, soluble and highly stable are further verified by predicting secondary and tertiary structure (**PSIPRED** (Buchan & Jones, 2019) and **I-TASSER** (Zhang, 2008), respectively), epitopes on 3D structure (**ElliPro** (Ponomarenko *et al.*, 2008)), molecular docking with immune receptor (**PatchDock** (Duhovny *et al.*, 2002, Schneidman-Duhovny *et al.*, 2005), molecular dynamics simulation (**GROMACS** (Berendsen *et al.*, 1995) and **PyMOL** – a commercial product: https://pymol.org/2/)**,** binding free energy (**MM-PBSA** and **MM-GBSA**

(Miller *et al.*, 2012), codon optimization (**Java Codon Adaptation Tool** (Grote *et al.*, 2005)), *in silico* cloning (**SnapGene** – a commercial product: https://www.snapgene.com/), and immune simulation (**C-ImmSim** (Rapin *et al.*, 2010)).

Informative protein characteristics

This section presents the predicted or obtained protein characteristics from the latest publications. The main question to be answered here for each characteristic is why it is considered informative to the overall *in silico* vaccine discovery approach. Programs used to predict or obtain these characteristics, and reported in more than one publication, are named along with a usage percentage given the number of latest publications. For example, Database of essential genes (DEG) is used in the workflow of 20 of the 90 latest RV publications; therefore its usage is 22.2 % (20/90). Note, programs listed here with a strikethrough indicate that the published URL failed to access the site or no up-to-date URL could be found at the time of execution by the authors (November 2022). URLs and usage percentage for all programs are listed in Supplementary Table S1.

Conserved proteins (stage #1)

The level of a protein's conservancy between strains is an informative protein characteristic. An ideal workflow starting point towards attaining a conserved vaccine is to determine proteins present in all strains of the target organism i.e., determine conserved proteins representing the core proteome. If no protein sequences are available, then the starting point is to perform a pangenomic analysis to determine the core genome (i.e., a set of homologous genes present in all genomes of the target organism) for translation into protein sequences.

The core proteome can be obtained by measuring protein sequence identity i.e., the amount of characters which match exactly between two different sequences. A user-defined threshold is first applied to the identity of proteins from the same strain to filter out paralogous and

duplicated proteins, and then to the identity of proteins from all strains to select the core proteome. Conserved proteins contain amino acid residues that are vital to its function, which is manifested by fewer variations from evolutionary selection pressures (Rappuoli, 2007). From a vaccine development perspective, conserved proteins help address the challenge of antigen variability i.e., a vaccine will only have continued success if the antigens targeted are relatively conserved and do not undergo significant variability over time. It must be noted, however, that conserved proteins are not expected to be the most virulent in a strain and therefore by association are possibly less antigenic. For example, strains have varying degrees of virulence. Strain-specific proteins are considered the determining factor making one strain more virulent than others. Virulence-associated proteins, nonetheless, are more prone to antigenic variation due to an evolutionary balancing act to evade the immune system by varying their antigens but still retaining functionality (Ernst, 2017). A popular workflow step in the latest publications is to determine which of the conserved proteins are essential for pathogen survival within the host and, in effect, filter out non-essential proteins from the RV protocol e.g., determine conserved proteins with roles in adhesion, and entry and infection. Tools for conservation and/or essentiality analysis: database of essential genes (DEG) 20.0%, CD-HIT 12.2%, COGS 6.7%, orthoMCL 5.6%, BPGA 4.4%, PATRIC 4.4%, ConSurf 2.2%, Geptop 2.2%, OrthoFinder 2.2%, and MBGD 2.2%.

Sequence similarity analysis with the proteome of the vaccine recipient (stage #1)

To avoid the likelihood of an autoimmune response, the sequences of vaccine candidates should have no significant similarity with any proteins from the intended vaccine recipient species. Note that although significant similarity between two sequences can infer they are related by evolutionary changes from a common ancestral sequence (i.e., sequence homology), finding homologous sequences is not the objective. Chains of amino acids from similar sequences, irrespective of their ancestry, can fold to potentially become similar

biologically active proteins in their native 3D structures. This has the conceivable consequence that the immune system responds both to the 3D structure of the vaccine and undesirably to a similar 3D structure residing in the vaccine recipient. Similarity based search tools: BlastP 47.8%, PSI-BLAST 7.8%.

Toxicity (stage #1 and #4)

It is important to ensure that any potential vaccine candidate, protein or peptide, will not have a detrimental effect when administered to the intended vaccine recipient i.e., a measure of the candidate's potential toxicity is required. Differences in single and dipeptide amino acid compositions of toxic and non-toxic peptides has been shown to exist (Gupta *et al.*, 2013). These differences can be detected with ML. Tool: ToxinPred 25.6%.

Allergenicity (stage #1 and #4)

Allergen proteins or peptides need to be removed from vaccine candidacy to avoid host allergic reactions. Tools: AllerTOP 20.0%, AllergenFP 15.6%, AlgPred 12.2%, AllerCatPro 2.2%, and ~~SORTALLER~~ 2.2%.

Antigenicity (stage #2)

Predicting a protein's antigenicity potential is possibly the most highly desirable characteristic. No encoded signal within protein sequences has yet been detected that clearly indicates a protein is antigenic. Consequently, there are no known programs directly using protein sequences to predict antigenicity. However, VaxiJen (developed in 2007) (Doytchinova & Flower, 2007) and AntigenPro (developed in 2010) (Magnan *et al.*, 2010) predict antigenicity scores by applying ML methods to known protective and non-protective antigen training data based on physicochemical properties derived from protein sequences or a collection of sequence-based features, respectively. Tools: VaxiJen 68.9%, AntigenPro 13.3%, Protegen (database of protective antigens) 3.3%.

Subcellular localization (stage #2)

An important characteristic is where a protein resides in the pathogen i.e., a protein's subcellular localization (SCL). The main determinant of an SCL is the protein sequence (Horton *et al.*, 2007). SCL's of interest for classical RV are those accessible to the host immune system e.g., cell wall, extracellular, secreted, and surface-exposed. Tools: PSORTb 43.3%, CELLO 24.0%, ~~SurfG+~~ 7.8%, SOSUI-GramN 4.4%, Wolf PSORT 2.2%.

Secreted proteins (stage #2)

Proteins secreted to the outside of the pathogen are accessible to the immune system. One of the most well-known sorting signals is the secretory signal peptide (SP), which targets a protein to the secretory pathway via the endoplasmic reticulum. Note, however, that not all secretory proteins have SPs, or are necessarily secreted to the outside of the pathogen (Emanuelsson *et al.*, 2007). Tools: SignalP 25.6%, SecretomeP (non-classical secretion) 5.6%, Phobius 5.6%, TatP 2.2%.

Membrane-related proteins (stage #2)

Surface membranes of pathogens are exposed to the outside environment and are therefore in full view of a host's immune system surveillance. Consequently, membrane molecules, including proteins spanning or anchored to the membrane are likely to be antigenic (Krogh *et al.*, 2001). Tools: TMHMM 36.7%, HMMTOP 15.6%, Phobius 5.6%, CCTOP 3.3%, PRED-TMBB 3.3%, BOMP 2.2%, TMBETADISC-RBF 2.2%.

Virulence (stage #2)

Focusing on pathogen targets accessible to the host immune system (e.g., membrane-related and secreted proteins) is important because of their potential role as virulence factors aiding in host cell infection. Target proteins that are virulent are deemed more worthy of onward investigation than non-virulent proteins. Tools to predict or determine virulence in bacterial proteins: VFDB 20.0%, VirulentPred 11.1%, VICMpred 2.2%. Adhesion is a significant

293 virulence factor and adhesins are worthwhile candidates because of their surface exposure.
294 Tool for predicting adhesins: SPAAN 12.2%. Some bacteria have been found to have
295 pathogenicity islands (PAIs), which carry virulence factor genes (Dobrindt *et al.*, 2000).
296 GIPSy (4.4%) is a tool to predict if putative targets are on PAIs i.e., virulence-associated.

297 *Protein function (stage #2)*
298 Determining a protein's function can provide an indication of its potential interaction with the
299 immune system. The conjecture is that amino acids determine the structure, and the structure
300 defines the function of the mature protein in the pathogen. If annotation on protein function is
301 unavailable or limited for the target organism, homology searching can be used to find
302 annotated proteins in other organisms e.g., proteins with similar sequences frequently
303 perform similar functions (program: BlastP).

304 Protein function is a multifaceted concept with complex mutually overlapping and
305 intertwined levels such as biochemical, cellular, organism-mediated, developmental and
306 physiological (Rost *et al.*, 2003, Clark & Radivojac, 2011). For instance, two proteins with
307 the same annotated molecular function may be involved in drastically different biological
308 processes, and conversely, a set of proteins associated with the same biological process may
309 have different molecular functions. It is also well-known that proteins can have more than
310 one function (Clark & Radivojac, 2011) e.g., **moonlighting proteins** (see Glossary)
311 (Henderson & Martin, 2011, Wang *et al.*, 2014). Several classification systems have been
312 proposed to standardize functional annotation, although not strictly specific to immunology
313 terms. One such classification system is Gene Ontology (GO) (Ashburner *et al.*, 2000,
314 Carbon *et al.*, 2021).

315 Proteins are typically composed of one or more building blocks, called **domains** (see
316 Glossary). Domain sequences can be classified in accordance to degrees of similarity. If a

region of protein sequence has a highly significant match to a particular domain, then it is likely to share similar structures and functions. Functionally important residues are also expected to be highly conserved. Tools: KEGG 11.1%, CDD 6.7%, ~~CELLO2go~~ 6.7%, Pfam 6.7%, InterProScan 5.6%, UniProt 4.4%, GO 3.3%, and eggNOG-mapper 2.2%.

B-cell epitopes (stage # 3)

The majority (~90%) of B-cell epitopes are **discontinuous** (or conformational) and the remaining 10% are **continuous** (or linear) (Korber *et al.*, 2006) (see Glossary). The main point to emphasize is that the specific interaction between B-cells and epitopes (in their folded state) occur at a 3D level. A challenge to the RV practitioner is that at least one epitope is predicted on any given protein. Therefore, selecting proteins for candidacy based on whether or not it contains an epitope is unfeasible. A common practice in the RV selection process is to use a metric based on a protein's epitope density. For example, B-cell epitope ratio (the numbers of amino acids of all epitopes divided into all amino acids of protein) (Oprea & Antohe, 2013), and mature epitope density (the number of 9-mer epitopes) (Santos *et al.*, 2013). Continuous predictors: ~~BCPred~~ 24.4% , BepiPred 23.3%, ABCpred 20.0%, IEDB B-cell epitopes 10.0%, ~~FBCPred~~ 4.4%; discontinuous predictors (predicted from 3D structure): ElliPro 18.9%, DiscoTope 4.4%; and Epitope mapping: Pepitope 3.3%.

T-cell epitopes (stage # 3)

T-cell epitopes are typically short linear peptides (Hanada *et al.*, 2004) and are predicted via an indirect method (see Fig. 10). Major histocompatibility complex (MHC) molecules are inherited and unique to an individual. They bind peptides exhibiting specific sequence patterns i.e., allele sequences. Therefore, MHC alleles vary within the species of the target host. This is associated with an individual's susceptibility or resistance to infection (Juliarena *et al.*, 2008), and why vaccine efficacy may differ between individuals. A judicious approach towards developing a vaccine that protects a broader target population would be to identify

conserved epitopes that bind to multiple MHC alleles (i.e., promiscuous epitopes) *and* bind to promiscuous MHCs. Note that there is no guarantee that a protein predicted to contain peptides that bind to a particular MHC allele will be presented by antigen-presenting cells and/or recognised by cognate T-cell receptors and/or is immunogenic.

Similar to B-cell epitopes, a previous study (Goodswen *et al.*, 2014) reported that every protein from the eukaryotic pathogens tested were predicted to contain at least one peptide binding with a high-affinity to at least one of the known human MHC alleles. This finding suggests that selecting a protein for vaccine candidacy on the basis it contains a high-affinity peptide is impractical. Proposed solutions are to identify **immunological hotspots** (see Glossary) and use density ratio metrics such as MHC I or II binding site ratios (Oprea & Antohe, 2013) (similar to B-cell epitope ratios) and an ML-derived probability to encapsulate all peptide-MHC binding scores from a protein into one score (Goodswen *et al.*, 2014). Tools: IEDB MHC-II Binding 31.1%, IEDB MHC-I Binding 25.6%, ProPred 12.2%, NetCTL 11.1%, NetMHCpan 8.9%, MHCPred 7.8%, NetMHCII 7.8%, NetMHCIIpan 7.8%, NetMHC 5.6%, ~~MHC2Pred~~ 4.4%, CTLPred 3.3%, SYFPEITHI 3.3%, MHCcluster 2.2%, NetCTLpan 2.2%, RANKPEP 2.2%, Vaxitop 2.2%. IFNepitope (17.8%) and IL10Pred (4.4%) can predict the nature of an MHC class-II epitope as either an IFN-γ or IL-10 inducer, respectively.

Conservancy of epitopes (stage # 3)

It is desirable for a conserved epitope-based vaccine to contain epitopes conserved across multiple strains or even species than epitopes unique to only one strain. Conserved epitopes tend to evolve slowly, even under immune pressure, because they typically have a critical protein function (Ernst, 2017). One method to determine the degree of epitope conservation is to appropriately align the epitope to a set of homologous protein sequences representing the desired scope of multiple strains. Note that sequence conservation does not guarantee that the

epitope will be recognized by the immune system or be cross-reactive. This is mainly because of differences in residues flanking the conserved epitope on different antigens (Ernst, 2017) e.g., T-cell epitopes need to be presented via MHC molecules to be recognised, and the flanking sequences influence this presentation. Also, B-cell epitope conformation is influenced by the entire 3D antigen shaped by the flanking sequences. Tools related to peptide conservation analysis: IEDB Population coverage 12.2%, IEDB Epitope conservancy tool 10.0%, IEDB-clustering analysis 5.6%, and BLAT (sequence similarity based search tool) 3.0%.

Chemical and physical properties of vaccine construct (stage # 4)

The vaccine construct comprising peptides, adjuvants, and linkers at the time of delivery will be a folded 3D structure presented to the immune system. This means that exposed peptides of the construct as opposed to buried peptides are more important in determining the immunogenic capacity. This is because only exposed amino acids can interact with T- and/or B-cells. Predicting different physicochemical properties of the construct can help assess its potential interactions in a 3D environment. Preferable vaccine construct properties are hydrophilic, stable, good water solubility, high **thermostability** (see Glossary), and not too large for purification (Enayatkhani *et al.*, 2021, Goodarzi *et al.*, 2021). The following properties can be deduced from the construct sequence: molecular weight (smaller size vaccines are easier to purify during experimental studies (Allemailem, 2021)), theoretical isoelectric point (pI) (the pH at which construct has a neutral charge), instability index (an estimate of the construct stability in a solution), aliphatic index (indicates the relative volume occupied by **aliphatic** side chains (see Glossary) and is an indicator of thermostability), and hydropathicity index (a number representing the hydrophobic or hydrophilic properties). Tool for predicting physicochemical characteristics ProtPram 52.2% (predicts molecular weight,

391 pI, instability, aliphatic, and hydropathicity indexes). Tools for predicting solubility: SOLpro
392 13.3%, Protein-sol 6.7%, Innovagen 2.2%, and ~~PROSO II~~ 2.2%.

393 *Tertiary structure of vaccine construct (stage # 4)*
394 Theoretically, a protein sequence contains all the information needed to make structural
395 predictions. Unlike genetic code, however, there is no known code that can be used to
396 definitively predict the folded structure of a protein. There are mainly two prediction
397 methods: comparative modelling (when the input protein sequence significantly matches with
398 a known structure), and *de novo.* Viewing a 3D structure to assess its immunogenic potential,
399 or even its correctness, requires expert knowledge. Therefore, predicted 3D models are used
400 in subsequent workflow steps towards computationally validating a vaccine construct.
401 Models are defined with coordinates, typically in a Protein Data Bank (PDB) file format.
402 Tools: I-TASSER 16.7%, Phyre2 12.2%, RaptorX 12.2%, PEP-FOLD 10.0%, SWISS-
403 MODEL 10.0%, Modeller 7.8%, Robetta 6.7%, 3DPro 4.4%, MHOLline 4.4%, CABS-fold
404 2.2%, and trRosetta 2.2%.

405 *Protein-protein interactions* (stage # 4)
406 Proteins function by interacting with other proteins. The interactions create protein
407 complexes and networks (Aguttu *et al.*, 2021). Understanding candidate protein interactions
408 with closely related proteins may help reveal the candidate's function and its immunogenic
409 potential. This understanding can be achieved by first determining candidate and intra-species
410 protein interactions; and then performing a functional enrichment analysis on the resulting
411 interactions network. Tools: STRING 18.9%, GalaxyPepDock 3.3%.

412 *Protein structure analysis (stage # 4)*
413 The accuracy and reliability of most predicted 3D models remains in question. Consequently,
414 independent programs have been developed for recognition of errors and/or model refinement
415 given predicted 3D coordinates. These programs typically provide a type of conformational

416 correctness score (e.g., Template modelling (TM) score and Root-Mean-Square Deviation
417 (RMSD) (Ahmad *et al.*, 2017), and/or a Ramachandran plot of residues, where residues
418 located in a specific region indicate a reliable 3D model. Verification tools: ProSA 21.1%,
419 UCSF Chimera 20.0%, ERRAT 15.6%, PROCHECK 15.6%, ~~RAMPAGE~~ 8.9%, PDBsum
420 6.7%, Mod Refiner 3.3%, QMEAN 3.3%, MolProbity 2.2%. Refinement tools: GalaxyRefine
421 20.0% (ProSA also performs refinement).

422 *Molecular docking of vaccine constructs (stage # 4)*
423 Microbial signatures, such as bacterial cell wall components, are recognized by host innate
424 immune receptors (Ishii *et al.*, 2008) e.g., Toll-like receptor (TLR) cells. These receptors
425 trigger innate immune activation and regulate subsequent adaptive immune responses
426 (Medzhitov, 2007). An expectation is that an effective vaccine construct will present
427 microbial signatures. The best 3D predicted models of candidate constructs are used in
428 molecular docking (MD) programs to assess their binding conformation and interactions with
429 host immune receptors e.g., TLRs. If sufficient binding affinity and presentation ability with
430 host receptors are observed in simulated docking then it supports the possibility of a construct
431 induced immune response in the real-world. The type of host receptor used for MD is
432 dependent on the target pathogen i.e., it needs to be established, possibly through the
433 Literature, whether the receptor naturally plays a role in a host's immune response, which
434 conversely equates to whether the vaccine candidate is a potential agonist to the chosen
435 receptor. Furthermore, the receptor choice is dependent on availability of its 3D model.

436 Most MD programs predict the best docked intermolecular conformations e.g., where the
437 construct (the ligand molecule) and receptor molecule have the highest number of favourable
438 interactions. Construct-receptor complexes with low global binding energy scores are
439 considered favourable (see *Estimation of binding free energy* later). PDB codes or files in
440 PDB format of the ligand and receptor molecules are the only input data required. MD tools:

PatchDock 15.6%, AutoDock Vina 10.0%, HADDOCK 10.0%, ClusPro 8.9%, Discovery studio 5.6%, HawkDock 3.3%, and CPORT 2.2%. MD refinement and analysis tools: FireDock 12.2% and CPPTRAJ 4.4%.

Molecular dynamics simulation (stage # 4)

The best-scored construct-receptor complexes are subjected to molecular dynamics simulation i.e., simulating Newtonian equations of motion. Simulation programs use **force fields** (see Glossary), and the result of simulations are **trajectories** (see Glossary). The simulation objective here is to check docking binding stability and residual flexibility with metrics such as RMSD and root mean square fluctuations (RMSF)(Ahmad *et al.*, 2018), respectively. Lower RMSD and RMSF values indicate more stable complexes (Allemailem, 2021). Tools: PyMOL 12.2%, GROMACS 12.2%, AMBER 10.0%, iMods 7.8%, VMD 6.7%, and MDWeb 2.2%.

Estimation of binding free energy (stage # 4)

Solvation (see Glossary) and associated binding free energies produced as an outcome of interactions between the bound construct and receptor complex in an aqueous solvent are calculated i.e., the sum of all the energy released due to the intermolecular interactions of the construct (ligand) and immune receptor (protein) is estimated. Negative binding free energy is an indicator of high construct-receptor binding affinity. The binding free energy is calculated by taking frames from the molecular dynamics simulation trajectories. Tools: MM-GBSA 4.4% and MM-PBSA 4.4%.

Immune system simulation (stage # 4)

A considered Holy Grail for the *in silico* vaccine discovery approach is to predict immunogenicity of the vaccine construct in a simulated immune system i.e.; perform verification experiments *in silico*. Over the last 30 years predominantly two modelling techniques have been attempted to simulate the immune system: equation-based and **agent-**

based modelling (ABM) (see Glossary) (Shinde & Kurhekar, 2018). ABM appears to be the trending technique with several publications reporting programs implementing ABM techniques: Reactive Animation (2005) (Efroni *et al.*, 2005), SIMISYS (2006) (Kalita *et al.*, 2006), synthetic immune system (2007) (Mata & Cohn, 2007), IMMUNOGRID (2009) (Pappalardo *et al.*, 2009), C-ImmSim (2010) (Rapin *et al.*, 2010). However, the published URLs to access these programs are no longer valid, and Google searches conducted in November 2022 found no internet access to these or equivalent programs. The one exception is C-ImmSim, which may reflect why it is the only simulation program used in the latest publications (20.0% usage). This ABM program performs *in silico* experiments by simulating vaccine injections (represented by a vaccine sequence) administered at different time intervals. The output is a vaccine immune response profile with results such as antibody production in response to antigen injections. C-ImmSim still relies on epitope predictions prior to the ABM simulation with rules incorporating working theories on the immune system. From a user perspective, the challenge is ascertaining reliability of the output profile without performing an *in vivo* validation. We could find no study comparing a C-ImmSim's output with the real *in vivo* vaccine immune response.

Another program referred to in the latest publications to predict immunogenicity was IEDB Class I Immunogenicity (Calis *et al.*, 2013) (2.2% usage). This program provides a score indicating the probability of a peptide eliciting an immune response when presented on a MHC I molecule.

Codon optimization of vaccine sequence (stage # 4)

A vaccine development goal is to express the vaccine construct (represented by a sequence of amino acids) in an expression organism at levels to allow production and future purification for vaccine efficacy studies. A variety of protein expression organisms are currently available e.g. bacteria (*Escherichia coli* is the most popular*)* and eukaryotic hosts (e.g., mammalian

cells, yeast, and insect cells) (Tripathi & Shrivastava, 2019). The choice of expression organism dictates the type of expression vector containing the gene of interest, which are commonly either plasmids (propagated in bacterial cells) or viruses (engineered to infect eukaryotic cells). Each expression organism has strengths and weaknesses (Rosano & Ceccarelli, 2014, Gutierrez & Lewis, 2015, Baghban *et al.*, 2019, Tripathi & Shrivastava, 2019) and its selection may ultimately be governed by the vaccine construct's DNA sequence. For example, specific codon usage of different genes in some organisms relate to their rate of expression (Gouy & Gautier, 1982). This may require selecting the optimum DNA coding sequence for the vaccine construct from the vast number of possible coding sequences, given there are multiple codons coding for the same amino acid. As an illustration, the arginine codon AGA is a common codon in eukaryotic genes but is particularly rare in *E. coli* (Calderone *et al.*, 1996). The usage of rare codons for arginine in *E. coli* can provoke translational errors of amino acids (Sorensen *et al.*, 1989). Therefore, certain codons in some organisms used for expression of foreign genes are considered optimal for minimising errors.

The workflow step is to back-translate the vaccine construct sequence to generate a DNA sequence, and then optimise/adapt the codon usage to achieve high expression in the intended expression organism e.g., *E. coli*. Tools: Codon Adaptation (JCAT) tool (Grote *et al.*, 2005) 27.8% and Gene Designer software (commercial product) 2.2%.

In silico cloning of the codon optimised vaccine sequence in an expression organism (stage # 4)

The final workflow step is to confirm cloning and expression of the optimized final vaccine sequence in a suitable expression organism. This can be achieved by *in silico* cloning, which is essentially simulating experimental methods to assemble recombinant DNA molecules and to direct their replication within host organisms e.g., restriction enzyme digestion, PCR

primer design, PCR amplification, and ligation. Currently, the most popular program is SnapGene (20.0% usage), which is a commercial product.

Reverse vaccinology pipelines

To automate and facilitate the RV process of predicting protective antigens, software pipelines have been developed and made freely available since 2006. There are currently 11 known RV-related pipelines and listed here in the order of their release year: NERVE (Vivona *et al.*, 2006), VaxiJen (Doytchinova & Flower, 2007), Vaxign (Xiang & He, 2008), AntigenPro (Magnan *et al.*, 2010), Vacceed (Goodswen *et al.*, 2014), VacSol (Rizwan *et al.*, 2017), Antigenic (Rahman *et al.*, 2019), PanRV (Naz *et al.*, 2019), ReVac (D'Mello *et al.*, 2019), Vaxign-ML (Ong *et al.*, 2020), Vax-ELAN (Rawal *et al.*, 2021). These pipelines can be categorised according to their methodology for selecting candidates given protein characteristics (e.g., filtering or ranking), type of protein characteristics used in candidate selection (e.g., biological and/or physiochemical), mode of operation (e.g., web server and/or standalone), and organism type for which the pipeline has been designed (e.g., bacteria and/or eukaryotic parasite). Table 3 shows different attributes and categories of the 11 pipelines. A study by Dalsass *et al.* (Dalsass *et al.*, 2019) in 2019 compared pipelines designed for bacterial vaccines from years 2006 to 2017 (e.g., NERVE, VaxiJen 1.0, Vaxign, and VacSol but excluding AntigenPro). The study also included an ML method (Bowman *et al.*, 2011) and a revised Bowman ML method (Heinson *et al.*, 2017), which was not made available as a pipeline. VaxiJen 1.0 also uses ML but with a smaller training dataset. Dalsass *et al.* concluded from an evaluation with a benchmark dataset that the predicted vaccine candidates from each pipeline/method were in poor agreement suggesting that users should not rely on a single RV pipeline. The Bowman-Heinson method, nonetheless, performed the overall best in terms of the evaluation measures. Note that almost all known RV pipelines that perform

candidate ranking use ML for this purpose (the exception is ReVac that uses feature-based scoring).

Machine learning specific to reverse vaccinology

In this section we make the distinction between the internal or hidden use of ML within the bioinformatics programs and the application of ML by the RV practitioner. Machine learning is now a critical component in practically every bioinformatics program used to predict RV-related protein characteristics. Surprisingly, however, ML is not directly applied in the typical RV workflow. For example, the workflow for selecting candidates in 87.8% of the latest publications is a consecutive filtering process not involving ML. This process essentially entails predicting a score or classification for a protein characteristic via a Web server, and then retaining or discarding proteins based on a rule-based selection threshold for the next Web server in the workflow. A major disadvantage of a series of filtering steps is that a potential candidate can inadvertently be discarded due to only one erroneous characteristic prediction and/or a marginally below threshold value. Ideally, all predicted protein characteristic scores and classifications should be simultaneously considered during candidate selection. This ideal has been approached by ML i.e., the RV pipelines that rank candidates implement ML with the generalised goal of collectively representing all predicted protein characteristics in a single score indicative of a protective antigen. One advantage is that ML-derived ranking scores are not severely compromised by one or two erroneous protein characteristics, unlike the filtering workflow. The ML methods used are binary classifiers such as support vector machines (learning models with associated learning algorithms), k-nearest neighbors algorithm, and random forest algorithm. These supervised algorithms *learn* from training data to classify unseen input data as 1 (positive) or 0 (negative) e.g., vaccine or non-vaccine candidate. Training data comprises one dataset

representing examples of positives and another one representing negatives. Quantity and quality of training data are paramount to the ML algorithm's performance.

Ideal training data would be sourced from proteins that were observed in a host to induce a protective immune response (positives) or observed to be non-immunogenic (negatives). Currently, there are insufficient numbers of known proteins meeting these ideal requirements. This raises a fundamental cyclic conundrum that currently limits the ML potential for RV candidate selection. That is, a sufficient number of verified protective antigens are required in the training data to predict protective antigens. The present strategy to tackle the conundrum is to build a sufficient quantity of training data using verified *and* 'likely' protective antigens. 'Likely' antigens are those published to induce an immune response *in vitro* or in an animal model, and those proteins experimentally shown to be naturally exposed to the immune system. The strategy can be statistically evaluated by predicting the outcome of known verified antigens not used in the training data. We have successfully followed this strategy in a recent study against *Babesia bovis (Goodswen et al., 2021a, Goodswen et al., 2021b)*. Finding 'likely' antigens can still be a time-consuming task for many pathogens, especially eukaryotic parasites. The only known repository distinguishing proteins with immunogenic potential is Protegen (Yang *et al.*, 2011) (November 2022: contains 1548 protective antigens, with 167 unique to parasites).

An ongoing but significant challenge in training data preparation is how best to represent the collection of biological and/or physiochemical characteristics, predicted from protein sequences of varying length, as a fixed length of features appropriate for ML input. For example, VaxiJen has faced this challenge by using auto cross covariance (ACC) to transform physicochemical properties of varying length amino acid sequences into uniform equal-length vectors (Doytchinova & Flower, 2007). We describe in a previous study (Goodswen *et al.*, 2013) a methodology to convert a collection of biological characteristics,

predicted by seven bioinformatics programs, to a fixed set of features representing the ML training data.

VaxiJen, which uses ML for candidate ranking, is used in 68.9% of the workflows described in the latest publications and is by far the most popular RV pipeline. Interestingly, however, VaxiJen is essentially used in these publications to predict an antigenicity score as one step in a filtering workflow e.g., programs such as PSORTb and/or TMHMM programs are still used before or after to filter VaxiJen results.

Concluding remarks

Reverse vaccinology remains a dynamic evolving process that can still be regarded as one in its infancy due to limitations still to overcome. In a nutshell, these limitations are bioinformatics tools and their biological input and output data with various levels of inaccuracies; lack of an accepted standard as to what steps constitute an RV workflow or an agreed set of tools to complete these steps; and inadequate numbers of experimentally validated vaccine candidates to provide examples for prediction targets, ML training and testing data. Taken together, the accumulated impact of these limitations makes it difficult to quantify how close RV is from reaching its full potential. This section first presents the constraints of the review itself and then proceeds with the authors' observations, opinions, and proposed solutions on RV's current status having conducted the review research. Table 4 summarises the outstanding RV issues and proposed solutions.

The review constraints

To capture current understanding of RV and its usage in the scientific community, all published papers from the last seven years with 'reverse vaccinology' in their title were manually reviewed (133 papers in total, source: Web of Science). There were, however, 490 additional papers from the same period with RV in the abstract or keywords but not in the

613 title. A question that arises is whether the 133 reviewed papers truly represent current RV
614 status. Five of the 490 papers (Dixit, 2021, Fadaka *et al.*, 2021, Goethel *et al.*, 2021,
615 Wisnewski *et al.*, 2021, Yousafi *et al.*, 2021) were randomly selected and reviewed, given the
616 impracticality of reviewing every RV-related paper. We propose that the trends in RV
617 methodology and usage revealed in the 133 papers would not change significantly given
618 more RV-related papers from the same time period. A further challenge in capturing current
619 RV status is the unknown number of papers using an *in silico* vaccine discovery approach but
620 with no reference to RV in the title or abstract e.g., three such papers (Pourseif *et al.*, 2019,
621 Dong *et al.*, 2020, Mahmud *et al.*, 2021) use an RV approach in their overall workflow.
622 Added to this challenge is the non-standardised usage of terminology in publications, which
623 we believe reflects the scientific community's disputed understanding of what constitutes an
624 RV workflow step. For example, similarities and differences in steps described by such terms
625 as RV, subtractive proteomics, computational vaccinology, predictive vaccinology, and
626 immunoinformatics are debatable. Nonetheless, there exists a common goal in all reviewed
627 papers irrespective of terms used, which is to identify vaccine candidates *in silico*.

628 The Web of Science reports 171 'subtractive proteomics', 228 'computational
629 vaccinology', and 1047 immunoinformatics publications (as of November 2022 when using a
630 Topic search i.e., searching title, abstract, and keywords). We acknowledge that it remains
631 undetermined whether the presented current understanding of RV correlates to current
632 understanding of *in silico* vaccine discovery, given the unrealistic task of reviewing all
633 publications.

634 *Proposed unified term to encapsulate in silico vaccine discovery*
635 Given the latest publications as a guideline, the *in silico* steps can be categorised into four
636 consecutive stages: 1) input data gathering and preparation; 2) predicting proteins naturally
637 exposed to the immune system (classical RV); 3) predicting epitopes (immunoinformatics);

and 4) computational candidate verification. We propose that these four stages are unified under the term '*in silico* vaccine discovery'. Put simply, any workflow step that takes place on a computer can be encapsulated in this one term. Ideally, '*in silico* vaccine discovery' should be consistently used in titles, abstracts, and/or keywords in future publications. One consistent term will retain that important searchable link between all publications in the field.

Challenges presented by bioinformatics tools

Bioinformatics tools are a primary reason why *in silico* vaccine discovery is now a reality (see Fig. 2.) However, the tools in themselves contribute to RV challenges. First, the number of available bioinformatics tools to perform the workflow steps is almost overwhelming now and continues to rise e.g., 283 different tools were used in one or more of the workflows of the latest publications. The challenge is in selecting the best tool to use for each step, especially when choices are for tools performing the same task. There is no agreed common set of tools or workflow for *in silico* vaccine discovery. Without actually evaluating the tools, it is difficult to determine which tool is best for the task at hand. To critically evaluate and compare tools, one would need to find experimentally validated test data specific to the tools and establish appropriate test measures to justify 'the best tool', notwithstanding the fact one would need to install the latest tools (if needed), learn how to use them, determine comparative parameter settings, and extract/interpret results for evaluation. Due to extensive logistics of evaluating so many tools and the potential for subjectivity, we make no judgement here as to the quality of the tools.

It is clear from their frequency of use, however, that some tools are vastly more popular than others performing the same task. One could speculate that popular programs must be judged by the community to be comparatively of higher quality. Conversely, programs may increase in popularity simply because they are chosen on this reputation. We recommend using several tools performing the same task in order to prioritise/value results

that are in agreement, rather than trust one set of results from the most popular program. One common feature for all popular tools, including RV-related pipelines, is their accessibility through a graphical user interface (GUI). We concede that this review runs the risk of further encouraging the selection of popular tools by quantifying their popularity. It must be emphasized that popularity of a tool does not correlate necessarily with its quality. Older well-established tools are more likely to be used or cited when there may be better, more modern alternatives yet to gain popularity (see 'Future directions' for examples).

Second, all bioinformatic tools have various levels of inaccuracies e.g., there is always an *unknown* percentage of erroneous predictions. Ideally, every tool performing the same task needs to be independently evaluated on experimentally validated task-specific test/benchmark data using a standard set of testing protocols. Protocols such as using consistent empirical evaluation measures e.g., comparing program predictions with known actual results and deriving metrics like accuracy, specificity, sensitivity, and error rate. Realistically, it would be a monumental challenge for any one organisation to perform these proposed benchmark evaluations for the purpose of making the metrics readily accessible to the public, especially considering the ever-growing number of new tools and new versions of existing ones (see section later on proposed new website).

Third, the increasing broad range and complexity of the task-specific tools also presents a challenge to an RV practitioner. Often, the methods behind the tools are hidden from the user or too computationally sophisticated to fully understand. We conjecture that many users accept the tool output at face value without necessarily knowing how it was derived. If all tools implemented perfect methods with perfect accuracy then this black box mentality would not be an issue. Blindly choosing tools on popularity or simply due to lack of choice may hinder the required progression for new or improved tools.

Fourth, computational prediction of biological phenomenon (e.g., immune response cellular interactions) is unlikely ever to be perfect. Computer algorithms can be used to predict phenomena at a fundamental level with informative levels of accuracy (e.g., predicting a signal encoded in a protein sequence), but this accuracy decreases as systems grow. For example, there are separate rules at play at the atomic, biomolecular, subcellular and cellular levels etcetera. Each level adds a layer of complexity to the overall parent system. Chance interactions also contribute to complexity. The consequence of this complexity is an increase in variables, which generates more possibilities that are less predictable. A dynamic model of the immune system interacting with vaccine formulations is in principle feasible, but realistically there are still many hurdles to overcome.

Challenges presented by input data

Protein sequences are the key starting input data for the RV workflow. This immediately presents a challenge if none are available for the target organism. The compromise is to predict genes encoded within the genome sequence of the target organism. Except, for some pathogen species there are no complete genome sequences e.g., the genome sequence availability for eukaryotic and multicellular pathogens is limited when compared to the viral and bacterial pathogens. There are more than 100,000 prokaryotic genomes in public archives (Sommer & Salzberg, 2021) ranging from draft to high-quality sequences. Each generation of genome sequencing techniques has greatly improved sequence quality and cost-effectiveness.

The majority of protein sequences in public databases are deduced from predicted genes. Relative to eukaryotic genomes, prokaryotic genomes are small, structurally simple, have no introns, and most of their DNA (≈ 80-90%) encodes protein-coding genes. Current prokaryotic gene finders have a high sensitivity (≈ 99%) to known genes using species-specific gene models (Sommer & Salzberg, 2021), nevertheless, they also predict multiple novel but questionable genes (Dimonaco *et al.*, 2022) that are typically annotated

'hypothetical protein'. A recent evaluation of prokaryotic gene predictors (Dimonaco *et al.*, 2022) found that their performance was dependent on the genome being analysed, which effectively means a user should cautiously select a gene predictor appropriate to the target organism. *Ab initio* gene predictors for eukaryotic genomes are inaccurate in the absence of experimental evidence (Goodswen *et al.*, 2012), especially the precise recognition of exon-intron structures. To exacerbate this inaccuracy, the gene predictions are typically from poor quality eukaryotic genomes. For example, a recent study (Berna *et al.*, 2021) reveals misassembly, karyotype differences, and chromosomal rearrangements of the *Toxoplasma gondii* genome following a re-evaluation. This is disconcerting considering that *T. gondii* is an important model system for the phylum Apicomplexa, which includes *Plasmodium falciparum*, the cause of malaria. Taken together, inaccuracies in genome sequences and gene predictions, the prediction accuracy of protein characteristics is compromised given protein sequences deduced from gene predictions.

Underutilisation of automated and/or high-throughput workflows

A surprising 95.6% of workflows in the latest publications rely on RV tools online, despite restrictions on input data size and constraints on parsing the output. This implies that the typical workflow is not automated and/or high-throughput. We speculate that the alternatives of having to install a standalone program and/or adapt an API are a major disincentive to RV practitioners limited with time and/or programming and computer administration skills. The RV pipelines developed so far mainly perform stage #2 of the '*in silico* vaccine discovery' workflow i.e., predict proteins naturally exposed to the immune system. We propose that there is a need for an automated, high-throughput '*in silico* vaccine discovery' pipeline. The ideal pipeline would entail: an input filtering stage to obtain core proteins that are essential, non-redundant, non-homologous, non-allergenic, and non-toxic; a subsequent stage incorporating an ML selection process for proteins naturally exposed to the

737 immune system; and then an iterative third and fourth stage. The third stage involves
738 predicting from epitope-rich proteins, promiscuous epitopes with high binding affinity and
739 broad population coverage. These epitopes are used to construct different combinations of
740 candidate vaccine sequences. The final fourth stage is to computationally verify candidates
741 for immunogenicity and safety. Each workflow step within each pipeline stage would be
742 performed by a collection of bioinformatics tools to obtain a consensus, rather than a reliance
743 on one tool. APIs and similar internet access tools are the key to achieving high-throughput
744 automation. The ideal pipeline would need to provide a user-friendly GUI without
745 programming or third party installation requirements i.e., the pipeline is delivered as a
746 complete standalone package with pre-installed or pre-programmed access to third party
747 bioinformatics tools. This ideal could be achieved with software container technology e.g.,
748 Docker (Piccolo & Frampton, 2016, Kadri *et al.*, 2022).

749 *The need for in vivo validation*

750 Possibly the most important question to pose concerning RV is whether it is a
751 successful process for identifying vaccine candidates. The preeminent measure of success is
752 the manufacture of the vaccine candidate discovered by RV. The only known RV-inspired
753 commercialised vaccine is BEXSERO, which provides protection against meningococcal
754 disease caused by the bacterium *Neisseria meningitidis* serogroup B (Masignani *et al.*, 2019).
755 Progressing to the manufacturing stage is a long, complex process. It is difficult to assess if
756 any candidates identified in the latest publications will reach the manufacturing stage. An
757 expectation is that if significant validation results were obtained for the *in silico* identified
758 candidates, then a patent application would ensue e.g., a patent was applied and granted for a
759 candidate related to the BEXSERO vaccine (patent: US-8398999-B2). None of the latest
760 publications could be associated with patent applications. Perhaps a more interim success
761 measure is whether an RV-derived candidate induces a protective response in an animal

762 model. Currently, only 12.2% of the latest publications report tests on animal models. It is
763 unclear whether the vaccine candidates computationally or *in vitro* verified in 57.8% and
764 7.8% of the latest publications, respectively, will undergo future investigation or encourage
765 further grant funding to pursue the vaccine. Moreover, there is no known study that has
766 collectively quantified the prediction outcomes from RV studies i.e., it is not known how
767 many false positive and negative candidates have been erroneously proposed or excluded for
768 experimental validation.

769 We speculate that the limited use of animal model validation is due to time, financial,
770 and/or legal constraints, but paradoxically, *in silico* vaccine discovery without *in vivo*
771 validation could be considered an unfinished endeavour. Even with *in vivo* validation, a
772 candidate may only elicit its true potential in the context of other critical interdependent
773 vaccine design factors e.g., a perfect candidate might be identified, but any wrong decision in
774 the type of adjuvant and/or antigen display method and/or vaccine delivery route could
775 negate its immunogenic potential.

776 *A proposed new website for in silico vaccine discovery*

777 To help address the many challenges presented so far, we propose the creation of a
778 new website dedicated to *in silico* vaccine discovery. The premise is to provide a platform for
779 the research community to discuss and address challenges. In particular, the underlying goals
780 would be to establish, through community input, standards for '*in silico* vaccine discovery'
781 workflows and recommended tools, including data repositories for experimentally validated
782 candidates as examples of prediction targets, and task-specific benchmarking data for tool
783 evaluation and ML training data.

784 The choice of bioinformatic tools is not static. New or updated tools are constantly
785 being made available, whilst older or even new unsuccessful tools can disappear from public

access. Evaluating and keeping up-to-date with new tools, versions, URLs, and methodology would be a substantial challenge for any website curator or organisation. We propose that the website adopts a 'product review-type' model, such as those universally used by a community of consumers to make better purchasing decisions. In the new website, however, a registered scientific community will have the capacity to add/update new tools, versions, URLs, and importantly, add prescribed program appraisals.

Future directions

Protein sequences are the primary data that drives RV. Sequences are essentially a one dimensional abstraction, but yet host-pathogen interactions within an immune system are 3D. Transformation of the current one dimensional RV ideology to a 3D one is beginning to happen (e.g., molecular docking with immune receptor and molecular dynamics simulation) but requires continued encouragement. One exciting new development is AlphaFold (Jumper *et al.*, 2021), which is designed as a deep learning system for the prediction of 3D models of protein structures. In 2020, this program won the 14th Critical Assessment of Structural Prediction competition (CASP14) by a substantial margin. The newly upgraded AlphaFold 2 is producing predictions that approach the accuracy of an experimentally predicted structure. The code for AlphaFold is freely available at https://github.com/deepmind/alphafold/, and the AlphaFold database (https://alphafold.ebi.ac.uk/) provides open access to over 200 million protein structure predictions. AlphaFold is expected to accelerate research in nearly every field of biology, including *in silico* vaccine discovery.

AlphaFold is a product of artificial intelligence (AI). The impact of AI to every industry, including vaccine development, is expected to be so great it has potential to rival that of the internet. Multiple recent reviews (Alimadadi *et al.*, 2020, Arshadi *et al.*, 2020, Lalmuanawma *et al.*, 2020, Vaishya *et al.*, 2020, Arora *et al.*, 2021, Lv *et al.*, 2021) focus on

the application of AI towards drug and vaccine discovery, particularly for COVID-19 (relevance to protozoal infectious diseases is discussed in another review (Hu *et al.*, 2022)).

Machine learning is a core subfield under AI. Given the copious volumes of biological data relevant to RV that can be gathered or predicted, if may now be humanly impossible to detect vaccine candidates without ML. Applying ML to candidate decision making, rather than user-defined filtering criteria, is expected to grow. Reliable ML decisions, however, are completely dependent on quality input and training data. Poor quality protein sequences (mainly predicted) and limited protein vaccine examples are obstructing the decision making potential. Consequently, it is vital that the ML algorithms receive iterative cycles of experimental feedback for training. A rapid, inexpensive, high-throughput screening assay is greatly needed.

A long term aspiration is a 3D host immune system simulator that computationally predicts a vaccine candidate's efficacy. A simulator that was not used or cited in the latest publications is the Universal Immune System Simulator (UISS). UISS seems to be gaining prominence as a human immune system simulator with several validation studies (Pappalardo *et al.*, 2018, Pappalardo *et al.*, 2020, Russo *et al.*, 2020, Maleki *et al.*, 2022). There are also known simulation models that could provide useful ways to explore the interaction of different immunological components. A recent review describes examples of simulation models for immunologists (Handel *et al.*, 2020a). We were unable, however, to find publicly available programs that implement these described models. An R package called Dynamical Systems Approaches to Immune Response Modelling (DSAIRM) is a tool to learn about modelling in immunology (Handel, 2020b). This might be a worthwhile starting point for RV practitioners, with limited programming experience, to develop and use simulation models specific to their research.

A goal to strive towards is the acceptance, by vaccine regulatory agencies, of evidence generated from *in silico* trials designed to evaluate safety and efficacy of *in silico*-derived candidates. Contributions to this goal are ongoing. So far, proposed protocols for *in silico* trials have been validated for the efficacy evaluation of *in silico* developed vaccines (and for existing vaccines to determine dosages for improved efficacy) (Pappalardo *et al.*, 2019, Viceconti *et al.*, 2021, Russo *et al.*, 2022).

How is the future of RV envisioned? *In silico* vaccine discovery needs to be and is becoming a totally holistic approach. RV, as it currently stands, plays only a small but important part. Classical RV primarily focuses on genomics. Other high-throughput cutting-edge omics technologies are beginning to contribute to the holistic approach, such as transcriptomics, proteomics, metabolomics, interactomics (study of interactions between and among proteins), and immunomics. Rationally, RV can no longer be an approach used in isolation of other emerging approaches. It is even expected the term 'reverse vaccinology' may shortly be one of the past. New terms like '*in silico* vaccine discovery' perhaps now better encompasses the epitome of a holistic approach. Furthermore, solutions for identifying candidates *in silico* may not necessarily come from understanding of the biology and in the domain of biologists. To truly achieve a holistic approach requires a collaboration of interdisciplinary experts from unconventional areas e.g., spatial and hydrodynamics engineers to adapt their programs that compute area and volume of irregular 3D shapes such as antibodies and their antigens.

The COVID-19 virus pandemic has changed the world of immunology. It has fast-tracked vaccine technologies such as producing an RNA vaccine in record time. It is expected that RV methodology may change to exploit the new or matured technologies motivated by the COVID-19 urgency (e.g., RNA vaccines, viral vectors, and protein-based vaccines with potent adjuvants) (Rappuoli *et al.*, 2021).

One unquestionable reality is that the world will continue to be challenged by established, unknown, neglected tropical diseases, emerging, and re-emerging infectious disease threats. Vaccination is considered the most efficient tool for preventing these threats (Delany *et al.*, 2014). Reverse vaccinology, an integral stage of *in silico* vaccine discovery, will clearly help save time, cost and effort by reducing the number of false candidates assigned for laboratory validation.

Funding

This work was supported by the Australian Research Council [DP180102584].

Glossary

Agent-based modelling – a computational approach for simulating the actions and interactions of self-governing agents (e.g., immune cells) in order to understand the behaviour and outcomes of a system (e.g., the immune system).

Adjuvant – an agent that has no specific antigenic effect on its own but stimulates the immune system when used with other components.

Aliphatic – a group of organic chemical compounds in which the carbon atoms are linked in open chains.

Amphipathic – a hydrophobic side facing the major histocompatibility complex molecule and a hydrophilic side interacting with the T-cell receptor.

Antigenicity –the capacity of epitopes on proteins to bind specifically with T- and B-cell receptors from the adaptive immune system.

Attenuated vaccine – contains a live, attenuated (or weakened) micro-organism i.e., a 'whole pathogen' living vaccine or infectious vaccine.

Discontinuous (or conformational) B-cell epitope – amino acids are brought together spatially in the folded antigen to form the epitope i.e., binding site motifs are not encoded by a contiguous primary sequence.

Domains – protein domains are generally considered as independently-folding units of structure.

Computational vaccinology – an interdisciplinary field addressing scientific and clinical questions in vaccinology using computational and informatics approaches, which overlaps fields such as immunoinformatics, reverse vaccinology, vaccinomics, literature mining, and systems vaccinology.

Continuous (or linear) B-cell epitope – a continuous stretch of amino acids in a protein sequence.

Conserved vaccine – a vaccine that provides broad protection across multiple strains.

Force field –a computational method used in molecular dynamics simulation to estimate the forces between atoms within molecules and also between molecules.

Immunoinformatics – the application of tools of computation and analysis to the capture and interpretation of immunological data.

Immunological hotspot – a region with a certain density of epitopes within a given protein sequence.

Killed vaccine – contains a killed (or inactivated), but previously virulent, micro-organism i.e., a 'whole pathogen' non-living vaccine or non-infectious vaccine).

Linker – an added sequence in a vaccine construct that plays a vital role in making the construct more stable e.g., produces extended conformation (flexibility), protein folding, and separation of functional domains.

Moonlighting proteins – examples of multifunctional proteins e.g., these protein types are typically classified as cytoplasmic and lack sequence motifs commonly found in known secreted or surface-exposed proteins, but they additionally have the ability to localise on the cell surface to contribute to virulence.

Pathogenic – ability of an organism to cause disease.

Solvation – the interaction of a solvent with dissolved molecules.

Subtractive proteomics – a computation process starting with entire proteome that undergoes a sequential subtraction process to narrow down the number of proteins to a few vaccine candidates e.g., the process involves a step by step removal of unwanted proteins from the pathogen and host proteomes to leave a set of protein candidates that are essential for the pathogen but absent in the host. Subtractive genomics is a process identical to 'subtractive proteomics' but applied to genomes and genes.

Subunit vaccine – comprises antigenic components of a micro-organism i.e., a non-living, non-infectious vaccine or 'acellular' vaccine. The vaccine formulation needs other ingredients such as adjuvants.

Thermostability – indicates resistant to irreversible change at a high relative temperature.

Trajectories – sequential snapshots (frames) of a simulated molecular system which represents atomic coordinates at specific time periods.

Virulence – the degree of pathogenicity within a group or species.

References

Agarwala R, Barrett T, Beck J*, et al.* (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **46**: D8-D13.

Aguttu C, Okech BA, Mukisa A & Lubega GW (2021) Screening and characterization of hypothetical proteins of Plasmodium falciparum as novel vaccine candidates in the fight against malaria using reverse vaccinology. *Journal of Genetic Engineering and Biotechnology* **19**.

Ahmad S, Raza S, Uddin R & Azam SS (2017) Binding mode analysis, dynamic simulation and binding free energy calculations of the MurF ligase from Acinetobacter baumannii. *Journal of Molecular Graphics & Modelling* **77**: 72-85.

Ahmad S, Raza S, Uddin R & Azam SS (2018) Comparative subtractive proteomics based ranking for antibiotic targets against the dirtiest superbug: Acinetobacter baumannii. *Journal of Molecular Graphics & Modelling* **82**: 74-92.

Alimadadi A, Aryal S, Manandhar I, Munroe PB, Joe B & Cheng X (2020) Artificial intelligence and machine learning to fight COVID-19. *Physiological Genomics* **52**: 200-202.

Allemailem KS (2021) A Comprehensive Computer Aided Vaccine Design Approach to Propose a Multi-Epitopes Subunit Vaccine against Genus Klebsiella Using Pan-Genomics, Reverse Vaccinology, and Biophysical Techniques. *Vaccines* **9**.

Arora G, Joshi J, Mandal RS, Shrivastava N, Virmani R & Sethi T (2021) Artificial Intelligence in Surveillance, Diagnosis, Drug Discovery and Vaccine Development against COVID-19. *Pathogens* **10**.

Arshadi AK, Webb J, Salem M*, et al.* (2020) Artificial Intelligence for COVID-19 Drug Discovery and Vaccine Development. *Frontiers in Artificial Intelligence* **3**.

Ashburner M, Ball CA, Blake JA*, et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**: 25-29.

Baghban R, Farajnia S, Rajabibazl M, Ghasemi Y, Mafi A, Hoseinpoor R, Rahbarnia L & Aria M (2019) Yeast Expression Systems: Overview and Recent Advances. *Molecular Biotechnology* **61**: 365-384.

Bateman A & Martin MJ & Orchard S*, et al.* (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**: D480-D489.

Berendsen HJC, Vanderspoel D & Vandrunen R (1995) GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* **91**: 43-56.

Berna L, Marquez P, Cabrera A, Greif G, Francia ME & Robello C (2021) Reevaluation of the Toxoplasma gondii and Neospora caninum genomes reveals misassembly, karyotype differences, and chromosomal rearrangements. *Genome Research* **31**: 823-833.

Bhasin M & Raghava GPS (2004) Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine* **22**: 3195-3204.

Bowman BN, McAdam PR, Vivona S*, et al.* (2011) Improving reverse vaccinology with a machine learning approach. *Vaccine* **29**: 8156-8164.

Bruno L, Cortese M, Rappuoli R & Merola M (2015) Lessons from Reverse Vaccinology for viral vaccine design. *Current Opinion in Virology* **11**: 89-97.

Buchan DWA & Jones DT (2019) The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Research* **47**: W402-W407.

Bui H-H, Sidney J, Dinh K, Southwood S, Newman MJ & Sette A (2006) Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *Bmc Bioinformatics* **7**.

Calderone TL, Stevens RD & Oas TG (1996) High-level misincorporation of lysine for arginine at AGA codons in a fusion protein expressed in Escherichia coli. *Journal of Molecular Biology* **262**: 407-412.

Calis JJA, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, Kesmir C & Peters B (2013) Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. *Plos Computational Biology* **9**.

Carbon S & Douglass E & Good BM*, et al.* (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research* **49**: D325-D334.

Chen LH, Yang J, Yu J, Ya ZJ, Sun LL, Shen Y & Jin Q (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Research* **33**: D325-D328.
Cheng J, Randall AZ, Sweredoski MJ & Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Research* **33**: W72-W76.
Clark WT & Radivojac P (2011) Analysis of protein function and its prediction from amino acid sequence. *Proteins-Structure Function and Bioinformatics* **79**: 2086-2096.
D'Mello A, Ahearn CP, Murphy TF & Tettelin H (2019) ReVac: a reverse vaccinology computational pipeline for prioritization of prokaryotic protein vaccine candidates. *Bmc Genomics* **20**.
Dalsass M, Brozzi A, Medini D & Rappuoli R (2019) Comparison of Open-Source Reverse Vaccinology Programs for Bacterial Vaccine Antigen Discovery. *Frontiers in Immunology* **10**.
Delany I, Rappuoli R & De Gregorio E (2014) Vaccines for the 21st century. *Embo Molecular Medicine* **6**: 708-720.
Dimitrov I, Bangov I, Flower DR & Doytchinova I (2014) AllerTOP v.2-a server for in silico prediction of allergens. *Journal of Molecular Modeling* **20**.
Dimonaco NJ, Aubrey W, Kenobi K, Clare A & Creevey CJ (2022) No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study. *Bioinformatics* **38**: 1198-1207.
Dixit NK (2021) Design of Monovalent and Chimeric Tetravalent Dengue Vaccine Using an Immunoinformatics Approach. *International Journal of Peptide Research and Therapeutics* **27**: 2607-2624.
Dobrindt U, Janke B, Piechaczek K, Nagy G, Ziebuhr W, Fischer G, Schierhorn A, Hecker M, Blum-Oehler G & Hacker J (2000) Toxin genes on pathogenicity islands: impact for microbial evolution. *International Journal of Medical Microbiology* **290**: 307-311.
Dong R, Chu Z, Yu F & Zha Y (2020) Contriving Multi-Epitope Subunit of Vaccine for COVID-19: Immunoinformatics Approaches. *Frontiers in Immunology* **11**.
Doytchinova IA & Flower DR (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *Bmc Bioinformatics* **8**.
Duhovny D, Nussinov R & Wolfson HJ (2002) Efficient unbound docking of rigid molecules. *Algorithms in Bioinformatics, Proceedings,* Vol. 2452 (Guigo R & Gusfield D, eds.), p.^pp. 185-200.
E. G, C. H, A. G, S. D, M.R. W, R.D. A & A. B (2005) Protein Identification and Analysis Tools on the ExPASy Server. *The Proteomics Protocols Handbook,*(Walker JM, ed.) p.^pp. 571-607 Humana Press.
Efroni S, Harel D & Cohen IR (2005) Reactive animation: Realistic modeling of complex dynamic systems. *Computer* **38**: 38-+.
Emanuelsson O, Brunak S, von Heijne G & Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols* **2**: 953-971.
Enayatkhani M, Hasaniazad M, Faezi S, Guklani H, Davoodian P, Ahmadi N, Einakian MA, Karmostaji A & Ahmadi K (2021) Reverse vaccinology approach to design a novel multi-epitope vaccine candidate against COVID-19: an in silico study. *Journal of Biomolecular Structure & Dynamics* **39**: 2857-2872.
Ernst JD (2017) Antigenic Variation and Immune Escape in the MTBC. *Strain Variation in the Mycobacterium Tuberculosis Complex: Its Role in Biology, Epidemiology and Control,* Vol. 1019 (Gagneux S, ed.) p.^pp. 171-190.
Fadaka AO, Sibuyi NRS, Martin DR, Goboza M, Klein A, Madiehe AM & Meyer M (2021) Immunoinformatics design of a novel epitope-based vaccine candidate against dengue virus. *Scientific Reports* **11**.
Flower DR, Macdonald IK, Ramakrishnan K, Davies MN & Doytchinova IA (2010) Computer aided selection of candidate vaccine antigens. *Immunome research* **6 Suppl 2**: S1-S1.
Goethel M, Listek M, Messerschmidt K, Schloer A, Hoenow A & Hanack K (2021) A New Workflow to Generate Monoclonal Antibodies against Microorganisms. *Applied Sciences-Basel* **11**.

Goodarzi NN, Bolourchi N, Fereshteh S & Badmasti F (2021) Introduction of novel putative immunogenic targets against Proteus mirabilis using a reverse vaccinology approach. *Infection Genetics and Evolution* **95**.
Goodswen SJ, Kennedy PJ & Ellis JT (2012) Evaluating High-Throughput Ab Initio Gene Finders to Discover Proteins Encoded in Eukaryotic Pathogen Genomes Missed by Laboratory Techniques. *Plos One* **7**.
Goodswen SJ, Kennedy PJ & Ellis JT (2013) A novel strategy for classifying the output from an in silico vaccine discovery pipeline for eukaryotic pathogens using machine learning algorithms. *Bmc Bioinformatics* **14**.
Goodswen SJ, Kennedy PJ & Ellis JT (2014) Enhancing In Silico Protein-Based Vaccine Discovery for Eukaryotic Pathogens Using Predicted Peptide-MHC Binding and Peptide Conservation Scores. *Plos One* **9**.
Goodswen SJ, Kennedy PJ & Ellis JT (2014) Vacceed: a high-throughput in silico vaccine candidate discovery pipeline for eukaryotic pathogens based on reverse vaccinology. *Bioinformatics* **30**: 2381-2383.
Goodswen SJ, Kennedy PJ & Ellis JT (2021a) Applying Machine Learning to Predict the Exportome of Bovine and Canine Babesia Species That Cause Babesiosis. *Pathogens* **10**.
Goodswen SJ, Kennedy PJ & Ellis JT (2021b) Predicting Protein Therapeutic Candidates for Bovine Babesiosis Using Secondary Structure Properties and Machine Learning. *Frontiers in Genetics* **12**.
Gouy M & Gautier C (1982) Codon usage in bacteria - correlation with gene expressivity. *Nucleic Acids Research* **10**: 7055-7074.
Grote A, Hiller K, Scheer M, Munch R, Nortemann B, Hempel DC & Jahn D (2005) JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Research* **33**: W526-W531.
Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Raghava GPS & Open Source Drug D (2013) In Silico Approach for Predicting Toxicity of Peptides and Proteins. *Plos One* **8**.
Gutierrez JM & Lewis NE (2015) Optimizing eukaryotic cell hosts for protein production through systems biotechnology and genome-scale modeling. *Biotechnology Journal* **10**: 939-949.
Hanada K, Yewdell JW & Yang JC (2004) Immune recognition of a human renal cancer antigen through post-translational protein splicing. *Nature* **427**: 252-256.
Handel A (2020b) A software package for immunologists to learn simulation modeling. *Bmc Immunology* **21**.
Handel A, La Gruta NL & Thomas PG (2020a) Simulation modelling for immunologists. *Nature Reviews Immunology* **20**: 186-195.
Heinson AI, Woelk CH & Newell M-L (2015) The promise of reverse vaccinology. *International Health* **7**: 85-89.
Heinson AI, Gunawardana Y, Moesker B*, et al.* (2017) Enhancing the Biological Relevance of Machine Learning Classifiers for Reverse Vaccinology. *International Journal of Molecular Sciences* **18**.
Henderson B & Martin A (2011) Bacterial Virulence in the Moonlight: Multitasking Bacterial Moonlighting Proteins Are Virulence Determinants in Infectious Disease. *Infection and Immunity* **79**: 3476-3491.
Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ & Nakai K (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Research* **35**: W585-W587.
Hu RS, Hesham A & Zou Q (2022) Machine Learning and Its Applications for Protozoal Pathogens and Protozoal Infectious Diseases. *Frontiers in Cellular and Infection Microbiology* **12**.
Ishii KJ, Koyama S, Nakagawa A, Coban C & Akira S (2008) Host innate immune receptors and beyond: Making sense of microbial infections. *Cell Host & Microbe* **3**: 352-363.
Jardetzky TS, Brown JH, Gorga JC, Stern LJ, Urban RG, Strominger JL & Wiley DC (1996) Crystallographic analysis of endogenous peptides associated with HLA-DR1 suggests a common, polyproline II-like conformation for bound peptides. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 734-738.

Jespersen MC, Peters B, Nielsen M & Marcatili P (2017) BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Research* **45**: W24-W29.
Juliarena MA, Poli M, Sala L, Ceriani C, Gutierrez S, Dolcini G, Rodriguez EM, Marino B, Rodriguez-Dubra C & Esteban EN (2008) Association of BLV infection profiles with alleles of the BoLA-DRB3.2 gene. *Animal Genetics* **39**: 432-438.
Jumper J, Evans R, Pritzel A*, et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**: 583-+.
Kadri S, Sboner A, Sigaras A & Roy S (2022) Containers in Bioinformatics: Applications, Practical Considerations, and Best Practices in Molecular Pathology. *The Journal of molecular diagnostics : JMD*.
Kalita JK, Chandrashekar K, Hans R, Selvam P & Newell MK (2006) Computational modelling and simulation of the immune system. *International journal of bioinformatics research and applications* **2**: 63-88.
Korber B, LaBute M & Yusim K (2006) Immunoinformatics comes of age. *Plos Computational Biology* **2**: 484-492.
Krogh A, Larsson B, von Heijne G & Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology* **305**: 567-580.
Lalmuanawma S, Hussain J & Chhakchhuak L (2020) Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos Solitons & Fractals* **139**.
Lew-Tabor AE & Valle MR (2016) A review of reverse vaccinology approaches for the development of vaccines against ticks and tick borne diseases. *Ticks and Tick-Borne Diseases* **7**: 573-585.
Li W & Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658-1659.
Lundegaard C, Hoof I, Lund O & Nielsen M (2010) State of the art and challenges in sequence based T-cell epitope prediction. *Immunome research* **6 Suppl 2**: S3-S3.
Lv H, Shi L, Berkenpas JW, Dao F-Y, Zulfiqar H, Ding H, Zhang Y, Yang L & Cao R (2021) Application of artificial intelligence and machine learning for COVID-19 drug discovery and vaccine design. *Briefings in Bioinformatics* **22**.
Magnan CN, Zeller M, Kayala MA, Vigil A, Randall A, Felgner PL & Baldi P (2010) High-throughput prediction of protein antigenicity using protein microarray data. *Bioinformatics* **26**: 2936-2943.
Mahmud S, Rafi MO, Paul GK*, et al.* (2021) Designing a multi-epitope vaccine candidate to combat MERS-CoV by employing an immunoinformatics approach. *Scientific Reports* **11**.
Maleki A, Russo G, Palumbo GAP & Pappalardo F (2022) In silico design of recombinant multi-epitope vaccine against influenza A virus. *Bmc Bioinformatics* **22**.
Masignani V, Pizza M & Moxon ER (2019) The Development of a Vaccine Against Meningococcus B Using Reverse Vaccinology. *Frontiers in Immunology* **10**.
Mata J & Cohn M (2007) Cellular automata-based modeling program: synthetic immune system. *Immunological Reviews* **216**: 198-212.
Medzhitov R (2007) Recognition of microorganisms and activation of the immune response. *Nature* **449**: 819-826.
Miller BR, III, McGee TD, Jr., Swails JM, Homeyer N, Gohlke H & Roitberg AE (2012) MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation* **8**: 3314-3321.
Mistry J, Chuguransky S, Williams L*, et al.* (2021) Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**: D412-D419.
Naz K, Naz A, Ashraf ST, Rizwan M, Ahmad J, Baumbach J & Ali A (2019) PanRV: Pangenome-reverse vaccinology approach for identifications of potential vaccine candidates in microbial pangenome. *Bmc Bioinformatics* **20**.

Ong E, Wang H, Wong MU, Seetharaman M, Valdez N & He Y (2020) Vaxign-ML: supervised machine learning reverse vaccinology model for improved prediction of bacterial protective antigens. *Bioinformatics* **36**: 3185-3191.
Oprea M & Antohe F (2013) Reverse-vaccinology strategy for designing T-cell epitope candidates for Staphylococcus aureus endocarditis vaccine. *Biologicals* **41**: 148-153.
Pappalardo F, Russo G, Tshinanu FM & Viceconti M (2019) In silico clinical trials: concepts and early adoptions. *Briefings in Bioinformatics* **20**: 1699-1708.
Pappalardo F, Russo G, Pennisi M, Palumbo GAP, Sgroi G, Motta S & Maimone D (2020) The Potential of Computational Modeling to Predict Disease Course and Treatment Response in Patients with Relapsing Multiple Sclerosis. *Cells* **9**.
Pappalardo F, Russo G, Pennisi M, Sgroi G, Alessandro G, Palumbo P, Motta S & Fichera E (2018) An agent based modeling approach for the analysis of tuberculosis - immune system dynamics. p.^pp. 1386-1392. Madrid, SPAIN.
Pappalardo F, Halling-Brown MD, Rapin N*, et al.* (2009) ImmunoGrid, an integrative environment for large-scale simulation of the immune system for vaccine discovery, design and optimization. *Briefings in Bioinformatics* **10**: 330-340.
Piccolo SR & Frampton MB (2016) Tools and techniques for computational reproducibility. *Gigascience* **5**.
Pizza M, Grandi G, Telford JL & Rappuoli R (2002) Reverse vaccinology: A genome-based approach to vaccine development. *Chimica Oggi-Chemistry Today* **20**: 32-36.
Pizza M, Scarlato V, Masignani V*, et al.* (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **287**: 1816-1820.
Ponomarenko J, Bui H-H, Li W, Fusseder N, Bourne PE, Sette A & Peters B (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes. *Bmc Bioinformatics* **9**.
Pourseif MM, Yousefpour M, Aminianfar M, Moghaddam G & Nematollahi A (2019) A multi-method and structure-based in silico vaccine designing against Echinococcus granulosus through investigating enolase protein. *Bioimpacts* **9**: 131-144.
Rahman MS, Rahman MK, Saha S, Kaykobad M & Rahman MS (2019) Antigenic: An improved prediction model of protective antigens. *Artificial Intelligence in Medicine* **94**: 28-41.
Rammensee HG, Friede T & Stevanovic S (1995) MHC ligands and peptide motifs: first listing *Immunogenetics* **41**: 178-228.
Rapin N, Lund O, Bernaschi M & Castiglione F (2010) Computational Immunology Meets Bioinformatics: The Use of Prediction Tools for Molecular Binding in the Simulation of the Immune System. *Plos One* **5**.
Rappuoli R (2000) Reverse vaccinology. *Current Opinion in Microbiology* **3**: 445-450.
Rappuoli R (2007) Bridging the knowledge gaps in vaccine design. *Nature Biotechnology* **25**: 1361-1366.
Rappuoli R, De Gregorio E, Del Giudice G, Phogat S, Pecetta S, Pizza M & Hanon E (2021) Vaccinology in the post-COVID-19 era. *Proceedings of the National Academy of Sciences of the United States of America* **118**.
Rawal K, Sinha R, Abbasi BA*, et al.* (2021) Identification of vaccine targets in pathogens and design of a vaccine using computational approaches. *Scientific Reports* **11**.
Rizwan M, Naz A, Ahmad J, Naz K, Obaid A, Parveen T, Ahsan M & Ali A (2017) VacSol: a high throughput in silico pipeline to predict potential therapeutic targets in prokaryotic pathogens using subtractive reverse vaccinology. *Bmc Bioinformatics* **18**.
Rosano GL & Ceccarelli EA (2014) Recombinant protein expression in Escherichia coli: advances and challenges. *Frontiers in Microbiology* **5**.
Rost B, Liu J, Nair R, Wrzeszczynski KO & Ofran Y (2003) Automatic prediction of protein function. *Cellular and Molecular Life Sciences* **60**: 2637-2650.

Russo G, Di Salvatore V, Sgroi G, Palumbo GAP, Reche PA & Pappalardo F (2022) A multi-step and multi-scale bioinformatic protocol to investigate potential SARS-CoV-2 vaccine targets. *Briefings in Bioinformatics* **23**.
Russo G, Pennisi M, Fichera E, Motta S, Raciti G, Viceconti M & Pappalardo F (2020) In silico trial to test COVID-19 candidate vaccines: a case study with UISS platform. *Bmc Bioinformatics* **21**.
Sachdeva G, Kumar K, Jain P & Ramachandran S (2005) SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinformatics* **21**: 483-491.
Santos AR, Pereira VB, Barbosa E, Baumbach J, Pauling J, Roettger R, Turk MZ, Silva A, Miyoshi A & Azevedo V (2013) Mature Epitope Density - A strategy for target selection based on immunoinformatics and exported prokaryotic proteins. *Bmc Genomics* **14**.
Schneidman-Duhovny D, Inbar Y, Nussinov R & Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Research* **33**: W363-W367.
Shinde SB & Kurhekar MP (2018) Review of the systems biology of the immune system using agent-based models. *Iet Systems Biology* **12**: 83-92.
Sommer MJ & Salzberg SL (2021) Balrog: A universal protein model for prokaryotic gene prediction. *Plos Computational Biology* **17**.
Sorensen MA, Kurland CG & Pedersen S (1989) Codon usage determines translation rate in *Escherichia coli*. *Journal of Molecular Biology* **207**: 365-377.
Teufel F, Almagro Armenteros JJ, Johansen AR, Gislason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G & Nielsen H (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*.
Tripathi NK & Shrivastava A (2019) Recent Developments in Bioprocessing of Recombinant Proteins: Expression Hosts and Process Development. *Frontiers in Bioengineering and Biotechnology* **7**.
Vaishya R, Javaid M, Khan IH & Haleem A (2020) Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome-Clinical Research & Reviews* **14**: 337-339.
Viceconti M, Pappalardo F, Rodriguez B, Horner M, Bischoff J & Tshinanu FM (2021) In silico trials: Verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products. *Methods* **185**: 120-127.
Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A & Peters B (2019) The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research* **47**: D339-D343.
Vivona S, Bernante F & Filippini F (2006) NERVE: New Enhanced Reverse Vaccinology Environment. *Bmc Biotechnology* **6**.
Vivona S, Gardy JL, Ramachandran S, Brinkman FSL, Raghava GPS, Flower DR & Filippini F (2008) Computer-aided biotechnology: from immuno-informatics to reverse vaccinology. *Trends in Biotechnology* **26**: 190-200.
Wang G, Xia Y, Cui J, Gu Z, Song Y, Chen YQ, Chen H, Zhang H & Chen W (2014) The Roles of Moonlighting Proteins in Bacteria. *Current Issues in Molecular Biology* **16**: 15-22.
Wang P, Sidney J, Dow C, Mothe B, Sette A & Peters B (2008) A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *Plos Computational Biology* **4**.
Wisnewski AV, Redlich CA, Liu J*, et al.* (2021) Immunogenic amino acid motifs and linear epitopes of COVID-19 mRNA vaccines. *Plos One* **16**.
Xiang Z & He Y (2008) Vaxign: a web-based vaccine target design program for reverse vaccinology. Vol. 1 p.^pp. 23-29. Boston, MA.
Yang B, Sayers S, Xiang Z & He Y (2011) Protegen: a web-based protective antigen database and analysis system. *Nucleic Acids Research* **39**: D1073-D1078.
Yousafi Q, Amin H, Bibi S, Rafi R, Khan MS, Ali H & Masroor A (2021) Subtractive Proteomics and Immuno-informatics Approaches for Multi-peptide Vaccine Prediction Against Klebsiella oxytoca and Validation Through In Silico Expression. *International Journal of Peptide Research and Therapeutics* **27**: 2685-2701.

1219 Yu NY, Wagner JR, Laird MR*, et al.* (2010) PSORTb 3.0: improved protein subcellular localization
1220 prediction with refined localization subcategories and predictive capabilities for all prokaryotes.
1221 *Bioinformatics* **26**: 1608-1615.
1222 Zhang Y (2008) I-TASSER server for protein 3D structure prediction. *Bmc Bioinformatics* **9**.

1223

Programs and biological databases for *in silico* vaccine discovery

This document is a supplement to the article '**A guide to current methodology and usage of reverse vaccinology towards in silico vaccine discovery'**. Its purpose is to present a brief introduction and portal to the main bioinformatics tools (programs and biological databases) mentioned in the article. A typical reverse vaccinology (RV) workflow, as followed in the latest publications from the last seven years (2015 to 2021), can be conceptually viewed in four stages: stage #1 – input data gathering and preparation, stage #2 – predicting proteins naturally exposed to the immune system (classical RV), stage #3 – predicting epitopes (immunoinformatics), and stage #4 – vaccine candidate verification. Bioinformatics tools perform the steps within these stages.

Table of Contents

Conserved proteins (stage #1)

DEG [1] is a **database of essential genes**. DEG hosts records of currently available essential genomic elements, such as protein-coding genes and non-coding RNAs, among bacteria, archaea and eukaryotes.

URL = http://tubic.tju.edu.cn/deg

First released = 2004 (and last updated Dec 18th 2017)

Latest version = 15.2

Method = all information is stored and operated by using an open-source database management system, MySQL. The essential genes in DEG are extracted from 36 publications.

Input = protein sequence in FASTA format for a local BLASTP.

Note: 48 Bacteria and nine eukaryote species are recorded in DEG. The eukaryote species are *Arabidopsis thaliana*, *Aspergillus fumigatus*, *Caenorhabditis elegans*, *Danio rerio*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe* 972h-

Clustering (stage #1)

CD-HIIT [2] is a very widely used program for clustering and comparing protein or nucleotide sequences. Used in the reverse vaccinology workflow to find the common (core) proteome of a species.

URL = http://cd-hit.org/

First released = 2001 (and last updated 01 Mar 2019)

Latest version = 4.8.1

Method = the algorithm behind cd-hit is short word filtering, which can determine that the similarity between two sequences is below a certain value without performing an actual sequence alignment.

Input = protein sequences in FASTA format

Example output

```
Sorted Clusters
>Cluster 0
0       561aa, >TGME49_210678... *
1       486aa, >TGME49_323700... at 99.59%
2       219aa, >TGME49_207650... at 98.63%
3       486aa, >TGME49_323800... at 99.59%
4       486aa, >TGME49_323600... at 99.59%
```

```
5       219aa, >TGME49_237894... at 99.54%
6       486aa, >TGME49_237900... at 99.59%
>Cluster 1
0       568aa, >TGME49_322010... *
1       568aa, >TGME49_242240... at 93.13%
2       249aa, >TGME49_323330... at 99.60%
3       199aa, >TGME49_323300... at 98.49%
>Cluster 2
0       181aa, >TGME49_328000... at 100.00%
```

Sequence similarity analysis with the proteome of the vaccine recipient

To avoid the likelihood of an autoimmune response, the sequences of vaccine candidates should have no significant similarity with any proteins from the intended vaccine recipient species. The immune system targets cells and proteins for destruction that it considers "non-self".

BLASTp can be used to identify sequence similarity.

PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) [3] derives a position-specific scoring matrix (PSSM) or profile from the multiple sequence alignment of sequences detected above a given score threshold using NCBI protein–protein BLAST (BLASTp). PSI-BLAST provides a means of detecting distant relationships between proteins of the target organism and a human. Search database non-redundant protein sequences (nr) using PSI-BLAST. The aim with respect to reverse vaccinology is to exclude human homolog proteins as candidates.

Example output

```
# blastp

# Iteration: 2

# Query:

# RID: UE7MVUDJ016

# Database: nr

# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches,
gap opens, q. start, q. end, s. start, s. end, evalue, bit score, % positives

# 501 hits found

Query_21782,TKC41514.1,90.870,471,42,1,1,471,26,495,0.0,866,94.48

Query_21782,XP_012499284.1,91.083,471,42,0,1,471,1,471,0.0,864,95.12

Query_21782,XP_020145984.1,90.234,471,46,0,1,471,1,471,0.0,863,94.27

…
```

Allergenicity (stage #1 and #4)

AllerTOP [4] predicts the allergenicity of a protein

URL = https://www.ddg-pharmfac.net/AllerTOP/

Latest version = 0.2

Method = based on auto cross covariance (ACC) transformation of protein sequences into uniform equal-length vectors. ACC is a protein sequence mining method that has been applied to quantitative structure-activity relationships (QSAR) studies of peptides with different length. The principal properties of the amino acids were represented by five E descriptors: amino acid hydrophobicity, molecular size, helix-forming propensity, relative abundance of amino acids, and β-strand forming propensity.

The proteins are classified by k-nearest neighbor algorithm (kNN,k=1) based on training set containing 2427 known allergens from different species and 2427 non-allergens.

Input = protein sequence in FASTA format (only one sequence at a time)

Output = a description of whether the sequence is an allergen

Example output

```
Your sequence is:
PROBABLE NON-ALLERGEN
The nearest protein is:
UniProtKB accession number Q9NZN5
defined as non-allergen
```

Other programs that predict allergenicity:

AlgPred [5] predicts allergenic proteins given a primary sequence. Main output = a probability and statement of protective antigen or non-antigen, according to a predefined threshold; method = user choice of Random Forest (RF) based on amino-acid composition or a hybrid approach (RF + BLAST + MERCI). MERCI (Motif - EmeRging and with Classes – Identification) is a program used to locate motifs in sets of sequences that represent positive and negatives [6].

Example output

```
AlgPred (Random Forest based on amino-acid composition)

Subject     ML Score    Prediction
test1 0.996       Allergen

ML Score = predicted scored from Random Forest
Note: Amino Acid Composition (AAC): It is a 20 length vector where each
element represents the fraction of each amino acid present in the protein
sequence.

AlgPred (a hybrid approach  based on RF + BLAST + MERCI)

Subject     ML Score    MERCI Score BLAST Score Hybrid Score
      Prediction
test1 1.0         0.5         0.5         2.0         Allergen

Hybrid Score is a combination of scores generated from machine learning
(RF), MERCI, and BLAST
```

AllergenFP.v1.0 (http://ddg-pharmfac.net/Allergen).

Toxicity (stage #1 and #4)

ToxinPred [7] predicts highly toxic regions in a given protein sequence

URL = http://crdd.osdd.net/raghava/toxinpred/

Method = models based on support vector machines (SVM) and quantitative matrix using various properties of toxic and non-toxic peptides/proteins obtained from Swiss-Prot and TrEMBL.

Input = protein sequence in FASTA format (only one sequence at a time)

Output = a table showing toxicity prediction and physicochemical properties of peptides within a given protein sequence

Example output

Peptide Sequence	SVM score	Prediction	Hydrophobicity	Hydropathicity	Hydrophilicity	Charge	Mol wt
CPKILKKCRC	-0.97	Non-Toxin	-0.38	-0.20	0.54	4.00	1191.71
PKILKKCRCS	-0.60	Non-Toxin	-0.40	-0.53	0.67	4.00	1175.65
KILKKCRCSI	-0.65	Non-Toxin	-0.32	0.08	0.49	4.00	1191.70
ILKKCRCSIR	-0.34	Non-Toxin	-0.39	0.02	0.49	4.00	1219.71
LKKCRCSIRI	-0.41	Non-Toxin	-0.39	0.02	0.49	4.00	1219.71
KKCRCSIRIC	-0.31	Non-Toxin	-0.44	-0.11	0.57	4.00	1209.68
KCRCSIRICM	0.02	Toxin	-0.30	0.47	0.14	3.00	1212.70

Subcellular localization (stage #2)

PSORTb [8] predicts bacterial protein subcellular localization (SCL) scores for five major localizations for Gram-negative bacteria (cytoplasmic, inner membrane, periplasmic, outer membrane and extracellular) and four localizations for Gram-positive bacteria (cytoplasmic, cytoplasmic membrane, cell wall and extracellular).

URL = www.psort.org/psortb/

Latest version = 3.0.3

Method = support vector machines (SVM) (contains 13 SVMs, one for each of the localization sites (five Gram-negative, four Gram-positive and four archaeal).

Input = protein sequence in FASTA format.

Output = a score and SCL associated with highest score.

Note: Web display mode is limited to the analysis of approximately 100 proteins. For larger analyses, the user must enter email address (results of up to 5000 per submission returned by email) or for even larger analyses a standalone version is recommended

Example output

```
SeqID Localization      Score
SAK_BPP42   Extracellular     9.98
```

Where ‘Score’ = a probability for the subcellular localization

TMHMM [9] predicts transmembrane helices in proteins.

URL = https://services.healthtech.dtu.dk/service.php?TMHMM-2.0

Latest version = 2.0

Method = hidden Markov model.

Input = protein sequence in FASTA format.

Main output = the number of predicted transmembrane helices.

Note: At most 10,000 sequences and 4,000,000 amino acids per submission; and each sequence should not be more than 8,000 amino acids.

Example output

```
COX2_BACSU
len=278
ExpAA=68.69
First60=39.89
PredHel=3
Topology=i7-29o44-66i87-109o
```

Where:

len=": the length of the protein sequence.

"ExpAA=": The expected number of amino acids intransmembrane helices

"First60=": The expected number of amino acids in transmembrane helices in the first 60 amino acids of the protein (see above).

"PredHel=": The number of predicted transmembrane helices by N-best.

"Topology=": The topology predicted by N-best. The topology shows the position of the transmembrane helices, where 'i' denotes the loop is on the inside, and 'o' on the outside. The above example 'i7-29o44-66i87-109o' means that it starts on the inside and has a predicted TMH at position 7 to 29, then a TMH at position 44-66 on the outside, and then a TMH at position 87-109 on the inside.

Antigenicity (stage #2 and #4)

VaxiJen [10] is an alignment-free approach for antigen prediction, which is based on auto cross covariance (ACC) transformation of protein sequences into uniform vectors of principal amino acid properties i.e., antigen classification solely based on the physicochemical properties of proteins without recourse to sequence alignment.

URL = https://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html

Latest version = 3.0.3

Method = ACC and two-class discriminant analysis by partial least squares (DA-PLS) [11].

Input = protein sequences in FASTA format.

Main output = a probability and statement of protective antigen or non-antigen, according to a predefined threshold.

Note: Jobs containing >100 proteins need to contact creators. The models discriminate between immunoprotective antigens and non-antigens without considering explicitly the presence or absence of T-cell or/and B-cell epitopes

Example output

```
Overall Prediction for the Protective Antigen = 0.5752 ( Probable ANTIGEN )
```

Signal peptides (stage #2)

SignalP [12] predicts the presence of signal peptides and the location of their cleavage sites in proteins from Archaea, Gram-positive Bacteria, Gram-negative Bacteria and Eukarya.

URL = https://services.healthtech.dtu.dk/service.php?SignalP-6.0

Latest version = 6.0

Method = based on a transformer protein language model with a conditional random field for structured prediction.

Input = protein sequences in FASTA format.

Output = long (with graphics) or short (no graphics) formats

Example output (short format)

```
GLR1_DROME_Glutamate_receptor_1_OS_Drosophila_melanogaster_GN_GluRIA_PE_1_SV_2

Prediction: Signal Peptide (Sec/SPI)

Cleavage site between pos. 27 and 28. Probability 0.949258

Protein type Other  Signal Peptide (Sec/SPI)

Likelihood   0.0013 0.9987
```

Virulence (stage #2)

The virulence factor database (VFDB) [13] is an online resource for curating information about virulence factors of bacterial pathogens**.**

URL = http://www.mgc.ac.cn/VFs/

First released = 2004 (last updated March 18th 2022)

Usage = can search VFDB by browsing each genus or by typing keywords. A BLAST search tool against all known VF-related genes is also available.

Adhesion (stage #2)

SPAAN [14] **(**can be accessed through Vaxign or NERVE) is a software program for prediction of adhesins and adhesin-like proteins using neural networks.

URL = https://sourceforge.net/projects/adhesin/files/SPAAN/

First released = 2005 (last updated 2013)

Method = uses a non-homology method using 105 compositional properties combined with artificial neural networks (ANNs) to identify adhesins and adhesin-like proteins in species belonging to a wide phylogenetic spectrum

Input = protein sequences in FASTA format.

Output = probability of a protein being an adhesion

Example output (SPAAN used from Vaxign)

```
"#","Protein Accession","Protein Name","Gene Accession","Gene
Symbol","Locus Tag","Adhesin Probability"
"1","SAK_BPP42","","-","-","-","0.662"
```

Protein function (stage #2)

CELLO2go [15] is a web server for protein subCELlular LOcalization prediction with functional Gene Ontology annotation

URL = http://cello.life.nctu.edu.tw/cello2go

First released = 2014

Method = provides brief and/or detailed annotations of GO terms related to homologs of a query protein found by BLAST searching in combination with a CELLO-predicted subcellular localization(s) for the queried protein

Input = protein sequences in FASTA format.

Output = Pie charts and Tables

Example output (Table only)

CELLO predictor for Gram- model:

Localization	Score
Extracellular	0.037
Outermembrane	0.018
Periplasmic	0.044
Innermembrane	0.197
Cytoplasmic	6.704

VICMpred is an SVM-based method for the prediction of functional proteins of gram-negative bacteria using amino acid patterns and composition [16].

Example output

Score of Different Functional Class	
Function	Score
cellular Process	**0.86543937**
Information Molecule	-0.23647597
Metabolism	-0.33623458
Virulence factors	-1.8177894

CDD (Conserved Domain Database) is a resource for the annotation of functional units in proteins. Its collection of domain models includes a set curated by NCBI, which utilizes 3D structure to provide insights into sequence/structure/function relationships [17].

Physical and chemical properties (stage #2 and #4)

ProtParam [18] (available from Expasy – the Swiss Bioinformatics Resource Portal [19]) allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL or for a user entered protein sequence. The computed parameters using the input sequence include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY)

URL = https://web.expasy.org/protparam/

First released = 2005

Input = protein sequence in FASTA format (only one sequence at a time).

Output = physical and chemical parameters for between selected endpoints on the input sequence or for the entire sequence

Note: No standalone version but ProtParam is a sub-module of Seq.Utils.

Example output

```
ProtParam

KPC1_DROME (P05130)
Protein kinase C, brain isozyme (EC 2.7.11.13) (PKC) (dPKC53E(BR))
Drosophila melanogaster (Fruit fly)

The computation has been carried out on the complete sequence (679 amino
acids).
Warning: All computation results shown below do not take into account
any annotated post-translational modification.

 References and documentation are available.
Number of amino acids: 679
Molecular weight: 77694.95
Theoretical pI: 6.77
Amino acid composition:
Ala (A)   28    4.1%
Arg (R)   26    3.8%
Etc …
Total number of negatively charged residues (Asp + Glu): 96
Total number of positively charged residues (Arg + Lys): 94
Atomic composition:
Carbon      C           3477
```

```
Hydrogen    H           5374
Nitrogen    N            922
Oxygen      O           1018
Sulfur      S             41

Formula: C3477H5374N922O1018S41
Total number of atoms: 10832

Extinction coefficients:
Extinction coefficients are in units of  M-1 cm-1, at 280 nm measured in
water.
Ext. coefficient    81135
Abs 0.1% (=1 g/l)   1.044, assuming all pairs of Cys residues form cystines
Ext. coefficient    79760
Abs 0.1% (=1 g/l)   1.027, assuming all Cys residues are reduced
Estimated half-life:
The N-terminal of the sequence considered is M (Met).
The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).
                            >20 hours (yeast, in vivo).
                            >10 hours (Escherichia coli, in vivo).
Instability index:
The instability index (II) is computed to be 37.98
This classifies the protein as stable.
Aliphatic index: 70.60
Grand average of hydropathicity (GRAVY): -0.517
```

Cytotoxic T lymphocytes epitopes (stage #3)

IEDB MHC-I [20] **(MHC-I Binding predictors)** are tools from the Immune Epitope Database (IEDB) analysis resource for predicting peptide binding to MHC class I molecules

URL = http://tools.iedb.org/mhci

Latest version = 2.24

Method = prediction method is chosen by the user. Prediction methods are: Artificial neural network (ANN), Stabilized matrix method (SMM), SMM with a Peptide:MHC Binding Energy Covariance matrix (SMMPMBEC), Scoring Matrices derived from Combinatorial Peptide Libraries (Comblib_Sidney2008), Consensus, NetMHCpan, NetMHCcons, PickPocket and NetMHCstabpan.

Input = protein sequences in FASTA format.

Main output = Table (see below)

Example output

allele	seq_num	start	end	length	peptide	core	icore	score	rank
HLA-A*01:01	1	92	100	9	CSANNSHHY	CSANNSHHY	CSANNSHHY	0.826691	0.06
HLA-A*01:01	2	197	205	9	ALTDLGLLY	ALTDLGLLY	ALTDLGLLY	0.774194	0.07
HLA-A*01:01	2	232	240	9	QSSINISGY	QSSINISGY	QSSINISGY	0.617697	0.13
HLA-A*01:01	1	417	425	9	ITEMLRKDY	ITEMLRKDY	ITEMLRKDY	0.559896	0.16
HLA-A*01:01	1	217	225	9	TTWCSQTSY	TTWCSQTSY	TTWCSQTSY	0.512549	0.18
HLA-A*01:01	1	233	241	9	RTWENHCTY	RTWENHCTY	RTWENHCTY	0.508887	0.19
HLA-A*01:01	1	162	170	9	FNNGITIQY	FNNGITIQY	FNNGITIQY	0.423562	0.25
HLA-A*01:01	2	487	495	9	YEDKVWDKY	YEDKVWDKY	YEDKVWDKY	0.422743	0.25

Helper T-lymphocyte epitopes (stage #3)

IEDB MHC-II [20] **(MHC-II Binding predictors)** are tools from the Immune Epitope Database (IEDB) analysis resource for predicting peptide binding to MHC class II molecules.

URL = http://tools.iedb.org/mhcii/

Method = prediction method is chosen by the user. Prediction methods are: IEDB recommended, Consensus method, Combinatorial library, NN-align-2.3 (netMHCII-2.3), NN-align-2.2 (netMHCII-2.2), SMM-align (netMHCII-1.1), Sturniolo, NetMHCIIpan-3.1, and NetMHCIIpan-3.2.

Input = protein sequences in FASTA format.

Main output = Table (see below)

Example output

allele	seq_num	start	end	length	method	peptide	percentile_rank	adjusted_rank	comblib_core	comblib_score
HLA-DRB1*01:01	1	482	496	15	Consensus (comb.lib./smm/nn)	ALTFLAVGGVLLFLS	0.1	0.1	FLAVGGVLL	0.01
HLA-DRB1*01:01	1	481	495	15	Consensus (comb.lib./smm/nn)	IALTFLAVGGVLLFL	0.1	0.1	FLAVGGVLL	0.01
HLA-DRB1*01:01	1	483	497	15	Consensus (comb.lib./smm/nn)	LTFLAVGGVLLFLSV	0.1	0.1	FLAVGGVLL	0.01
HLA-DRB1*01:01	1	479	493	15	Consensus (comb.lib./smm/nn)	RSIALTFLAVGGVLL	0.1	0.1	FLAVGGVLL	0.01
HLA-DRB1*01:01	1	480	494	15	Consensus (comb.lib./smm/nn)	SIALTFLAVGGVLLF	0.1	0.1	FLAVGGVLL	0.01
HLA-DRB1*01:01	1	485	499	15	Consensus (comb.lib./smm/nn)	FLAVGGVLLFLSVNV	0.91	0.91	FLAVGGVLL	0.01
HLA-DRB1*01:01	1	484	498	15	Consensus (comb.lib./smm/nn)	TFLAVGGVLLFLSVN	0.91	0.91	FLAVGGVLL	0.01
HLA-DRB1*01:01	1	447	461	15	Consensus (comb.lib./smm/nn)	GGAFRSLFGGMSWIT	1.8	1.8	FRSLFGGMS	0.06
HLA-DRB1*01:01	1	358	372	15	Consensus (comb.lib./smm/nn)	VNPFVSVATANAKVL	1.8	1.8	FVSVATANA	0.02
HLA-DRB1*01:01	1	446	460	15	Consensus (comb.lib./smm/nn)	FGGAFRSLFGGMSWI	2	2	FRSLFGGMS	0.06

Other predictors for cell-mediated epitopes:

NetMHC predicts binding of peptides to MHC class I molecules [21].

TepiTool [22] provides prediction of peptides binding to MHC class I and class II molecules. Input = protein sequence.

Example output

```
Seq #,Peptide start,Peptide end,Peptide,Percentile rank,Allele
1,29,37,SSFDKGKYK,0.01,HLA-A*11:01
1,74,82,FPIKPGTTL,0.01,HLA-B*35:01
1,74,82,FPIKPGTTL,0.01,HLA-B*07:02
```

IL17eScan [23], the similarity search module maps all experimentally validated epitopes in IEDB database that induce IL-17 response onto the similar sequences present in the input peptide/protein sequences (performs Smith Watermann search of query sequence in database of experimentally validated IL-17 inducing epitopes).

Example output

```
SIMSEARCH
The best scores are:                                        s-w bits E(388)
IL17eScan:77_IEDB                                ( 20)   47 21.8       1
IL17eScan:164_IEDB                               ( 15)   43 20.6     1.7
IL17eScan:195_IEDB                               ( 16)   43 20.5     1.9
```

IFNepitope [24] is a web server to predict and design IFN-gamma inducing peptides.

Example output

Serial No.	Epitope Name	Sequence	Method	Result	Score
1	Epitope_1	GVQQKWDATATELNN	MERCI	POSITIVE	5
2	Epitope_2	FAGIEAAASAIQGNV	MERCI	POSITIVE	2
3	Epitope_3	MTEQQWNFAGIEAAA	SVM	POSITIVE	0.99934

Linear B-cell epitopes (stage #3)

BCPred [25] predicts fixed length linear B-cell epitopes using string kernels

URL = http://ailab-projects1.ist.psu.edu:8080/bcpred/

Latest version = BCPREDS Server 1.0

Method = String kernels, which are a class of kernel methods that have been successfully used in many sequence classification tasks. In these tasks, a protein sequence is viewed as a string defined on a finite alphabet of 20 amino acids. In BCPred, the subsequence kernel and support vector machines (SVM) are used in predicting linear B-cell epitopes

Input = protein sequence in plain format.

Output = Table

Example output

Position	Epitope	Score
34	SRDANSSDASNWTIDGENRT	0.994
447	TLGKQQSEETCTDNINTVNE	0.989
425	QAGQNKDSKEDAEPTDNDCS	0.98
301	REPGSYTGRRTMQSISNEQK	0.937
199	VWTISVGVSMPIPVFGLQDD	0.796

Other Linear B-cell epitope predictors:

FBCPred [26] predicts flexible length linear B-cell epitopes.

BepiPred [27] predicts B-cell epitopes from a protein sequence, using a Random Forest algorithm trained on epitopes and non-epitope amino acids determined from crystal structures

Example output

```
Entry,Position,AminoAcid,Exposed/Buried,RelativeSurfaceAccessilibity,HelixP
robability, SheetProbability,CoilProbability,EpitopeProbability
5H2A_CRIGR,1,M,E,0.745,0.003,0.003,0.994,0.303
5H2A_CRIGR,2,E,E,0.592,0.052,0.084,0.864,0.371333333333
5H2A_CRIGR,3,I,E,0.39,0.056,0.142,0.802,0.442888888889
5H2A_CRIGR,4,L,E,0.48,0.018,0.088,0.893,0.510444444444
5H2A_CRIGR,5,C,E,0.482,0.018,0.088,0.893,0.587777777778
```

Conformational (discontinuous) B-cell epitopes (stage #3)

ElliPro [28] is a web-based tool for the prediction of antibody epitopes (linear and discontinuous) in protein antigens of a given sequence or structure (AUC value of 0.732).

URL = http://tools.iedb.org/ellipro

First released = 2008

Method = represents the protein structure as an ellipsoid and calculates protrusion indexes for protein residues outside of the ellipsoid. The method is based on geometrical properties of protein structure and does not require training.

Input = Protein Data Bank (PDB) ID(s) or upload PDB file

Output = Table with links to 3D views

Example output

```
Input Sequences: 5LYM
Chain:
A
1       KVFGRCELAA AMKRHGLDNY RGYSLGNWVC AAKFESNFNT QATNRNTDGS TDYGILQINS
61      RWWCNDGRTP GSRNLCNIPC SALLSSDITA SVNCAKKIVS DGNGMNAWVA WRNRCKGTDV
121     QAWIRGCRL
Predicted Linear Epitope(s):

No. Chain Start End Peptide Number of residues Score 3D structure
1   A     45    50  RNTDGS  6                  0.78
2 A 112 129 RNRCKGTDVQAWIRGCRL 18 0.771
3 A 100 103 SDGN               4  0.76
4 A 64  81  CNDGRTPGSRNLCNIPCS 18 0.666
5 A 1   7   KVFGRCE            7  0.597
6 A 13  23  KRHGLDNYRGY        11 0.574
7 A 85  88  SSDI               4  0.504

Predicted Discontinuous Epitope(s):
```

No.	Residues	Number of residues	Score	3D structure
1	A:S100, A:D101, A:G102, A:N103, A:N106	5	0.727	
2	A:K1, A:V2, A:F3, A:G4, A:R5, A:C6, A:E7, A:F38, A:N39, A:T40, A:Q41, A:A42, A:S85, A:S86, A:D87, A:I88, A:R112, A:N113, A:R114, A:C115, A:K116, A:G117, A:T118, A:D119, A:Q121, A:A122, A:I124, A:R125, A:G126, A:C127, A:R128, A:L129	32	0.657	
3	A:R45, A:N46, A:T47, A:D48, A:G49, A:S50, A:N59, A:S60, A:R61, A:W62, A:W63, A:C64, A:N65, A:D66, A:G67, A:R68, A:T69, A:P70, A:G71, A:S72, A:R73, A:N74, A:L75, A:C76, A:N77, A:I78, A:P79, A:S81	28	0.648	
4	A:A10, A:K13, A:R14, A:G16, A:L17, A:D18, A:N19, A:Y20, A:R21, A:G22, A:Y23, A:S24	12	0.564	

Prediction of conservancy of linear and conformational B cell epitopes:

Epitope Conservancy database [29] analyses the variability or conservation of epitopes (linear and discontinuous).

Input = epitope sequences and protein sequences from target organism. Note format for discontinuous epitopes.

Epitope population coverage (stage #3)

IEDB population coverage [30] calculates the fraction of individuals predicted to respond to a given epitope set on the basis of HLA genotypic frequencies and on the basis of MHC binding and/or T cell restriction data.

URL = http://tools.iedb.org/population/

First released = 2006

Method = the Allele Frequency database provides allele frequencies for 115 countries and 21 different ethnicities grouped into 16 different geographical areas.

Input = one epitope-allele combination per line

e.g., FMKAVCVEV HLA-A*02:01,HLA-A*02:02,HLA-A*02:03,HLA-A*02:06,HLA-A*68:02

Output = For each population coverage, the tool computes the following: (1) projected population coverage, (2) average number of epitope hits / HLA combinations recognized by the population, and (3) minimum number of epitope hits / HLA combinations recognized by 90% of the population (PC90).

Example output

Population Coverage Calculation Result

population/area	Class combined		
	coverage[a]	average_hit[b]	pc90[c]
World	98.47%	2.82	1.72
Average	**98.47**	**2.82**	**1.72**
Standard deviation	**0.0**	**0.0**	**0.0**

[a] projected population coverage
[b] average number of epitope hits / HLA combinations recognized by the population
[c] minimum number of epitope hits / HLA combinations recognized by 90% of the population

Solubility (stage #4)

SOLpro [31] predicts the propensity of a protein to be soluble upon overexpression in *E. coli* using a two-stage SVM architecture based on multiple representations of the primary sequence.

URL = http://scratch.proteomics.ics.uci.edu/explanation.html#SOLpro

Method = each classifier of the first layer takes as input a distinct set of features describing the sequence. A final SVM classifier summarizes the resulting predictions and predicts if the protein is soluble or not as well as the corresponding probability.

Input = protein sequence in plain format i.e., no header (only one sequence at a time)

Output = the results are sent to an e-mail address

Note: SOLpro is provided with the **Scratch protein predictor,** which is a server for predicting protein tertiary structure and structural features. It includes predictors for secondary structure, relative solvent accessibility, disordered regions, domains, disulfide bridges, single mutation stability, residue contacts versus average, individual residue contacts and tertiary structure [31].

Predict protein-protein interactions (stage #4)

STRING is a database that aims to integrate all known and predicted associations between proteins, including both physical interactions as well as functional associations [32].

CATH/Gene3D provides information on the evolutionary relationships of protein domains [33, 34]. CATH identifies domains in protein structures from wwPDB and classifies these into evolutionary superfamilies, thereby providing structural and functional annotations. Gene3D uses profile-Hidden Markov Models built from the CATH domain sequences to predict structural domains for proteins.

Secondary structure (stage #4)

PSIPRED [35] is a secondary structure prediction method

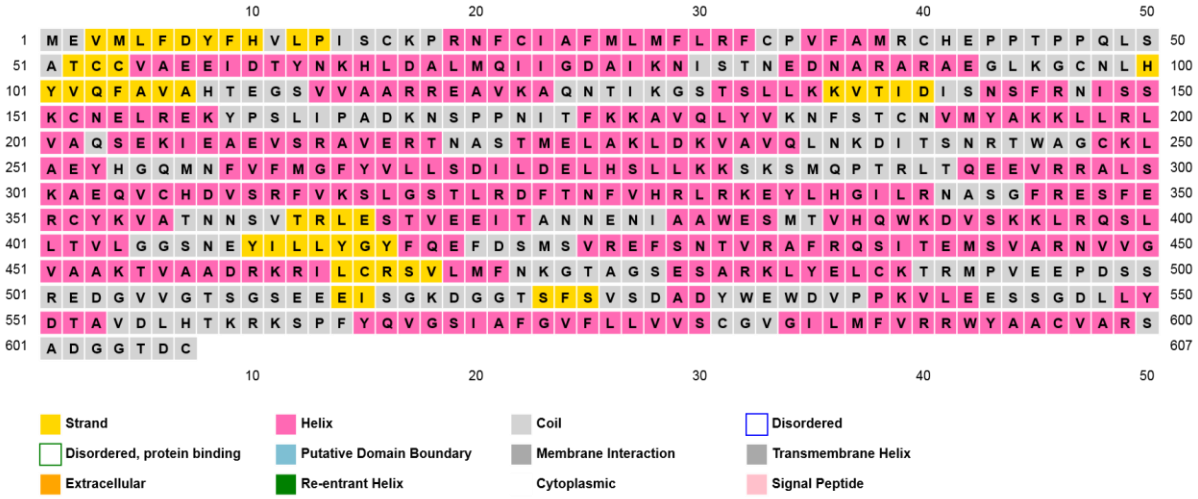
URL = http://bioinf.cs.ucl.ac.uk/psipred/

Method = incorporates two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST (Position Specific Iterated - BLAST).

Input = protein sequence in FASTA format (only one sequence at a time)

Main output = a cartoon

Example output



Tertiary structure (stage #4)

I-TASSER (Iterative Threading ASSEmbly Refinement) [36] is a hierarchical approach to protein structure prediction and structure-based function annotation.

URL = https://zhanggroup.org/I-TASSER/

Method = first identifies structural templates from the PDB by multiple threading approach LOMETS, with full-length atomic models constructed by iterative template-based fragment assembly simulations. Function insights of the target are then derived by re-threading the 3D models through protein function database BioLiP.

Input = protein sequence in FASTA format (only one sequence at a time)

Output = Secondary Structure, Solvent Accessibility, normalized B-factor, top 10 threading templates used by I-TASSER, top 5 final models predicted by I-TASSER, proteins structurally close to the target in the PDB (as identified by TM-align), predicted function using COFACTOR and COACH

Molecular docking (stage #4)

PatchDock [37, 38] is an algorithm for molecular docking based on shape complementarity principles

URL = https://zhanggroup.org/I-TASSER/

First released = 2002

Method = the algorithm is inspired by object recognition and image segmentation techniques used in Computer Vision. Docking can be compared to assembling a jigsaw puzzle e.g., matching two pieces by picking one piece and searching for the complementary one. Given two molecules, their surfaces

are divided into patches according to the surface shape. These patches correspond to patterns that visually distinguish between puzzle pieces. Once the patches are identified, they can be superimposed using shape matching algorithms. The algorithm has three major stages: Molecular Shape Representation, Surface Patch Matching, and Filtering and Scoring.

Input = two molecules of any type: proteins, DNA, peptides, drugs (requires PDB codes of receptor and ligand molecules or can upload files in PDB format)

Output = a list of potential complexes sorted by shape complementarity criteria.

Molecular dynamics simulation (stage #4)

GROMACS [39] is a package to perform molecular dynamics i.e., simulate the Newtonian equations of motion for systems with hundreds to millions of particles. It is primarily designed for biochemical molecules like proteins, lipids and nucleic acids that have a lot of complicated bonded interactions.

URL = https://www.gromacs.org/

Latest release = 2021.5 (first released in 1995)

Method = performs molecular dynamics simulation of (bio)macromolecules in a solvent, using classical equations of motion and force fields based on variable non-bonded interactions, and fixed bonded interactions. The system is coupled to an external bath of constant temperature and/or pressure. Rectangular periodic conditions are allowed. Bond lengths (and angles) can be constrained. External forces and force field terms related to experimental constraints can be added [39].

Input = protein databank file (PDB)

Output = trajectory file of a simulation. It contains all the coordinates, velocities, forces and energies.

Binding free energy (stage #4)

MM-PBSA (Molecular Mechanics Poisson Boltzmann Surface Area) and its complementary method **MM-GBSA** (Molecular Mechanics-Generalized Born Solvation Area) **[40]** are post-processing end-state methods to calculate free energies of molecules in solution.

URL = http://ambermd.org/

Release = source code can be downloaded at http://ambermd.org/ with AmberTools

Method = a program written in Python for streamlining end-state free energy calculations using ensembles derived from molecular dynamics (MD) or Monte Carlo (MC) simulations.

Input = solvated and unsolvated topology files

Output = file containing calculated free energies

Immune simulation (stage #4)

C-ImmSim [41] is an agent-based simulator of the immune response. It consists of a three dimensional (3D) stochastic cellular automaton in which the major classes of cells of both the lymphoid (T helper lymphocytes (Th), cytotoxic T lymphocytes (CTL), B lymphocytes, and antibody producer plasma cells, PLB) and the myeloid lineage (macrophages (Mw) and dendritic cells (DC))

are represented. All these entities interact with each other according to a set of rules that describe the different phases of the recognition and response processes of the immune system against a pathogen.

URL = https://kraken.iac.rm.cnr.it/C-IMMSIM/?page=0

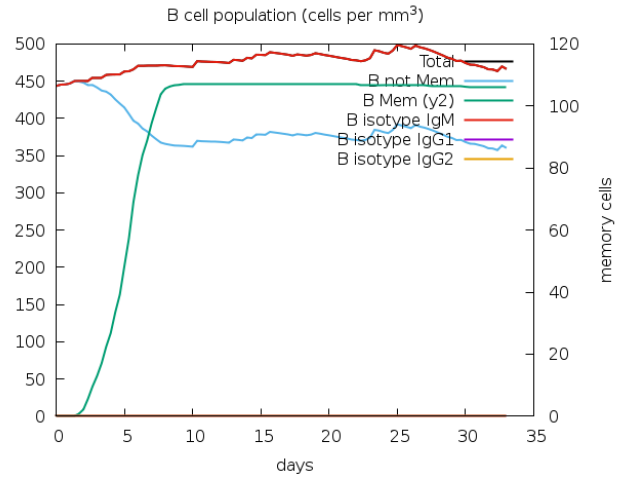
Last updated = 2010 (main logic behind C-ImmSim originates from 1991)

Method = a bit-string polyclonal lattice model. Bit-string refers to the way in which the molecules are represented, polyclonal indicates that the lymphocytes have genetic variation in their receptors, and lattice signifies that a discrete lattice is used to represent the space.

Input = protein sequences in a FASTA format

Output = graphs representing the vaccine immune response profile

Example output



Codon optimization (stage #4)

JCAT (Java Codon Adaptation Tool) [42] provides a method to adapt the Codon Usage to most sequenced prokaryotic organisms and selected eukaryotic organisms. The codon adaptation plays a major role in cases where foreign genes are expressed in hosts and the codon usage of the host differs from that of the organism where the gene stems from.

URL = http://www.jcat.de/

First released = 2005

Method = adaptation is based on Codon Adaptation Index (CAI) values proposed by Sharp,P.M. and Li,W.H. (1987). The CAI-values were calculated by applying an algorithm from Carbone,A., Zinovyev,A. and Kepes,F. (2003). The mean codon usage for a certain organism is derived by summing over all CAI-values of all genes of this organism (except genes without an amino acid sequence, e.g. RNAs) divided by the number of genes.

Input = protein or DNA sequence

Output = Results in a table and graph presentation e.g., Codon Adaptation Index (CAI) values given for the pasted sequence and the newly adapted sequence.

Example output

```
CAI-Value of the improved sequence:  GC-Content of the improved sequence:
0.9560192581582391                   65.34870950027457
```

Codon	Relative Adaptiveness (wij)
AUG	0.993603411513859
GAG	0.978678038379531
GUG	0.963752665245203
AUG	0.993603411513859
CUG	0.987206823027719

```
GC-Content of Homo sapiens:
40.892862223204
```

in silico cloning (stage #4)

SnapGene is a commercial product that enables a way to plan, visualize, and document everyday molecular biology procedures. With a graphical user interface, the software enables DNA sequence visualization, sequence annotation, sequence editing, cloning, protein visualization, and simulating common cloning methods.

URL = https://www.snapgene.com/

Latest release = 6.0

Input and output = SnapGene can read and write to the following common file formats:

Alignment Formats, ApE, CLC Bio, Clone Manager, DNA Strider, DNADynamo, DNASIS DNAssist, DNASTAR Lasergene®, DS Gene, EMBL (ENA), EnzymeX, GenBank / DDBJ Gene Construction Kit®, Geneious, GeneTool, Genome Compiler, Jellyfish, MacVector, pDRAW32, Sequencher, Serial Cloner, Swiss-Prot, Vector NTI®, Visual Cloning

References

1. Zhang R, Ou HY, Zhang CT. DEG: a database of essential genes, Nucleic Acids Research 2004;32:D271-D272.
2. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 2006;22:1658-1659.
3. Altschul SF, Madden TL, Schaffer AA et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Research 1997;25:3389-3402.
4. Dimitrov I, Bangov I, Flower DR et al. AllerTOP v.2-a server for in silico prediction of allergens, Journal of Molecular Modeling 2014;20.
5. Sharma N, Patiyal S, Dhall A et al. AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes, Briefings in Bioinformatics 2021;22.
6. Vens C, Rosso M-N, Danchin EGJ. Identifying discriminative classification-based motifs in biological sequences, Bioinformatics 2011;27:1231-1238.

7. Gupta S, Kapoor P, Chaudhary K et al. In Silico Approach for Predicting Toxicity of Peptides and Proteins, Plos One 2013;8.
8. Yu NY, Wagner JR, Laird MR et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics 2010;26:1608-1615.
9. Krogh A, Larsson B, von Heijne G et al. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes, Journal of Molecular Biology 2001;305:567-580.
10. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines, Bmc Bioinformatics 2007;8.
11. Doytchinova IA, Flower DR. Identifying candidate subunit vaccines using an alignment-independent method based on principal amino acid properties, Vaccine 2007;25:856-866.
12. Teufel F, Almagro Armenteros JJ, Johansen AR et al. SignalP 6.0 predicts all five types of signal peptides using protein language models, Nature Biotechnology 2022.
13. Chen LH, Yang J, Yu J et al. VFDB: a reference database for bacterial virulence factors, Nucleic Acids Research 2005;33:D325-D328.
14. Sachdeva G, Kumar K, Jain P et al. SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks, Bioinformatics 2005;21:483-491.
15. Yu C-S, Cheng C-W, Su W-C et al. CELLO2GO: A Web Server for Protein subCELlular LOcalization Prediction with Functional Gene Ontology Annotation, Plos One 2014;9.
16. Saha S, Raghava GPS. VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition, Genomics, proteomics & bioinformatics 2006;4:42-47.
17. Lu S, Wang J, Chitsaz F et al. CDD/SPARCLE: the conserved domain database in 2020, Nucleic Acids Research 2020;48:D265-D268.
18. E. G, C. H, A. G et al. Protein Identification and Analysis Tools on the ExPASy Server. In: Walker J. M. (ed) The Proteomics Protocols Handbook. Humana Press, 2005, 571-607
19. Duvaud S, Gabella C, Lisacek F et al. Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users, Nucleic Acids Research 2021;49:W216-W227.
20. Vita R, Mahajan S, Overton JA et al. The Immune Epitope Database (IEDB): 2018 update, Nucleic Acids Research 2019;47:D339-D343.
21. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system, Bioinformatics 2016;32:511-517.
22. Paul S, Sidney J, Sette A et al. TepiTool: A Pipeline for Computational Prediction of T Cell Epitope Candidates, Current protocols in immunology 2016;114:18.19.11-18.19.24.
23. Gupta S, Mittal P, Madhu MK et al. IL17eScan: A Tool for the Identification of Peptides Inducing IL-17 Response, Frontiers in Immunology 2017;8.
24. Dhanda SK, Vir P, Raghava GPS. Designing of interferon-gamma inducing MHC class-II binders, Biology Direct 2013;8.
25. El-Manzalawy Y, Dobbs D, Honavar V. Predicting linear B-cell epitopes using string kernels, Journal of Molecular Recognition 2008;21:243-255.
26. El-Manzalawy Y, Dobbs D, Honavar V. Predicting flexible length linear B-cell epitopes, Computational systems bioinformatics. Computational Systems Bioinformatics Conference 2008;7:121-132.
27. Jespersen MC, Peters B, Nielsen M et al. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes, Nucleic Acids Research 2017;45:W24-W29.
28. Ponomarenko J, Bui H-H, Li W et al. ElliPro: a new structure-based tool for the prediction of antibody epitopes, Bmc Bioinformatics 2008;9.
29. Bui H-H, Sidney J, Li W et al. Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines, Bmc Bioinformatics 2007;8.

30. Bui H-H, Sidney J, Dinh K et al. Predicting population coverage of T-cell epitope-based diagnostics and vaccines, Bmc Bioinformatics 2006;7.
31. Cheng J, Randall AZ, Sweredoski MJ et al. SCRATCH: a protein structure and structural feature prediction server, Nucleic Acids Research 2005;33:W72-W76.
32. Szklarczyk D, Gable AL, Nastou KC et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets, Nucleic Acids Research 2021;49:D605-D612.
33. Lewis TE, Sillitoe I, Dawson N et al. Gene3D: Extensive prediction of globular domains in proteins, Nucleic Acids Research 2018;46:D435-D439.
34. Sillitoe I, Bordin N, Dawson N et al. CATH: increased structural coverage of functional space, Nucleic Acids Research 2021;49:D266-D273.
35. Buchan DWA, Jones DT. The PSIPRED Protein Analysis Workbench: 20 years on, Nucleic Acids Research 2019;47:W402-W407.
36. Zhang Y. I-TASSER server for protein 3D structure prediction, Bmc Bioinformatics 2008;9.
37. Duhovny D, Nussinov R, Wolfson HJ. Efficient unbound docking of rigid molecules. In: Guigo R., Gusfield D. eds). Algorithms in Bioinformatics, Proceedings. 2002, 185-200.
38. Schneidman-Duhovny D, Inbar Y, Nussinov R et al. PatchDock and SymmDock: servers for rigid and symmetric docking, Nucleic Acids Research 2005;33:W363-W367.
39. Berendsen HJC, Vanderspoel D, Vandrunen R. GROMACS: A message-passing parallel molecular dynamics implementation, Computer Physics Communications 1995;91:43-56.
40. Miller BR, III, McGee TD, Jr., Swails JM et al. MMPBSA.py: An Efficient Program for End-State Free Energy Calculations, Journal of Chemical Theory and Computation 2012;8:3314-3321.
41. Rapin N, Lund O, Bernaschi M et al. Computational Immunology Meets Bioinformatics: The Use of Prediction Tools for Molecular Binding in the Simulation of the Immune System, Plos One 2010;5.
42. Grote A, Hiller K, Scheer M et al. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host, Nucleic Acids Research 2005;33:W526-W531.

Table 1 Comparison between conventional and reverse vaccinology approaches to subunit vaccine discovery

	Conventional	Reverse vaccinology
Type of vaccine components	Capacity to identify all types of pathogen components known to induce immunity including proteins, carbohydrates and lipids.	Limited to proteins only.
Protein availability for identification	Incapacity to identify all potential antigens because proteins expressed by a parasite may be different *in vitro* than those antigens expressed during infection *in vivo*, in a particular life cycle stage, or under different environmental conditions and stimuli.	All proteins can theoretically be identified because the genome holds the entire repertoire of genes, which the pathogen can potentially express, irrespective of life cycle stages and environmental stimuli.
Types of protein antigens identified	Laboratory techniques tend to capture the more abundant antigens or those that can be purified in quantities suitable for vaccine testing.	Allows for the discovery of both conventional vaccine targets (e.g., secreted or membrane-associated proteins) *and* novel protective antigens owing to the potential to analyse every single possible protein that can be expressed
Types of pathogens	Some pathogens are too difficult and/or dangerous to cultivate in the laboratory	Subunit vaccines can potentially be developed for any pathogen that has a genome sequence
Cost	Cost of laboratory setup, chemicals, and technicians is expensive	Relatively inexpensive – only requires a computer.
Timing	Time consuming laboratory procedures	Generating a list of potential antigens for laboratory testing can typically take only days, when given the appropriate computer analysis.
Accuracy	Identification of antigens is experimentally observed	Proteins are predominately translated from *predicted* genes encoded in the pathogen's genome sequence. Protein antigens are *predicted* by bioinformatics programs. Accuracy is dependent on quality of genome sequences and accuracy of programs.
Antigen verification	Experimental verification is an integral part of the conventional approach	Computational verification, however, only laboratory testing can verify that predicted antigens are truly antigenic

Table 2: Popular bioinformatics programs and biological databases used in a typical reverse vaccinology workflow

Stage	Name	Prediction	Mode	Organism	Access address
1	DEG	conservation	W	A,B,E	http://tubic.tju.edu.cn/deg/
1	CD–HIT	clustering	S,W	N	http://cd-hit.org/
1	BlastP	homology	A,S,W	N	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins
1+4	AllerTOP	allergenicity	W	N	https://www.ddg-pharmfac.net/AllerTOP/
1+4	ToxinPred	toxicity	S,W	N	https://webs.iiitd.edu.in/raghava/toxinpred2/
2	PSORTb	subcellular localization	S, W	A,B	www.psort.org/psortb/
2	TMHMM[a]	transmembrane domains	S, W	B, E	https://services.healthtech.dtu.dk/service.php?TMHMM-2.0
2+4	VaxiJen	antigenicity	W	B, E, F,V	https://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html
2	signalP	signal peptide	S,W	B, E, F,V	https://services.healthtech.dtu.dk/service.php?SignalP-6.0
2	VFDB	virulence	W	B	http://www.mgc.ac.cn/VFs/
2	SPAAN[b]	adhesion	S,W	B	https://sourceforge.net/projects/adhesin/files/SPAAN/
2	Pfam	protein function	W	N	https://pfam.xfam.org/[d]
3	IEDB MHC-I	CTL epitopes	S,W	N	http://tools.iedb.org/mhci/
3	IEDB MHC-II	HTL epitopes	S,W	N	http://tools.iedb.org/mhcii/
3	BepiPred	B-cell epitopes	S,W	N	https://services.healthtech.dtu.dk/service.php?BepiPred-3.0
3	ElliPro	Epitopes from 3D	S,W	N	http://tools.iedb.org/ellipro/
3	IEDB population coverage	population coverage	S,W	N	http://tools.iedb.org/population/
4	SOLpro	solubility	W	N	http://scratch.proteomics.ics.uci.edu/explanation.html#SOLpro
4	PSIPRED	secondary structure	A,S,W	N	http://bioinf.cs.ucl.ac.uk/psipred/
4	I-TASSER	tertiary structure	S,W	N	https://zhanggroup.org/I-TASSER/
4	PatchDock	molecular docking	S,W	N	https://bioinfo3d.cs.tau.ac.il/PatchDock/
4	GROMACS	molecular dynamics simulation	S	N	https://www.gromacs.org/
4	MM-PBSA/MM-GBSA	binding free energy	S	N	http://ambermd.org/
4	C-ImmSim	immune simulation	W	N	https://kraken.iac.rm.cnr.it/C

					-IMMSIM/?page=0
4	JCAT	codon optimization	S,W	B,E	http://www.jcat.de/
4	SnapGene[c]	*in silico* cloning	W	N	https://www.snapgene.com/

Stage = stage number of reverse vaccinology (RV) workflow: 1 (input data gathering and preparation, 2 (predicting proteins naturally exposed to the immune system – classical RV), 3 (predicting epitopes – immunoinformatics), and 4 (vaccine candidate verification).

Name = program or database name: IEDB MHC-I and IEDB-MHC-I (tools from the Immune Epitope Database (IEDB) analysis resource for predicting peptide binding to MHC class I and MHC class II molecules, respectively). [a]TMHMM-2.0 is outdated. DeepTMHMM has been released and is available at https://services.healthtech.dtu.dk/service.php?DeepTMHMM.
[b]SPAAN can be accessed through the web server Vaxign or NERVE.[c]SnapGene is a commercial product.

Prediction = main output from bioinformatics tool that is of interest to reverse vaccinology: CTL (cytotoxic T lymphocytes), HTL (helper T-lymphocyte).

Mode = modes of operating program: A (application programming interface (API)), B (Batch facility), D (download data from database), S (Standalone program), and W (Web Server).

Organism = type of organism for which the program has been designed: A (Archaea), B (Bacteria), E (Eukaryotes), F (Fungi), P (Plant), V (Viruses), N (type of organism not specified).

Access address = internet address for Web server or access to program (last viewed: February 2023). [d]The Pfam website was decommissioned in January 2023 (InterPro offers the same functionality and data https://www.ebi.ac.uk/interpro/).

Table 3: Freely available reverse vaccinology pipelines

Pipeline Name	Year	Usage (%)	Select.	Char.	Mode	Org.	Access address [Last viewed February 2023]
NERVE	2006	0.0	Filter	B	S	B	http://www.bio.unipd.it/molbinfo/
VaxiJen	2007	68.9	Rank	P	W	B,E,F, T,V	https://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html
Vaxign	2008	27.8	Filter	B	W	B	https://violinet.org/vaxign/
AntigenPro	2010	13.3	Rank	B + P	W	B	http://scratch.proteomics.ics.uci.edu/
Vacceed	2014	2.2	Rank	B	S	E	https://github.com/goodswen/vacceed
VacSol	2017	2.2	Filter	B	S.	B	https://sourceforge.net/projects/vacsol/
Antigenic	2019	0.0	Rank	P	W	B,E	https://github.com/srautonu/Antigenic
PanRV	2019	1.1	Filter	P	S	B	https://sourceforge.net/projects/panrv2/
ReVac	2019	0.0	Rank	P	S	B	https://github.com/admelloGithub/ReVac-package
Vaxign-ML	2020	4.4	Rank	B + P	W + S	B,V,E	https://violinet.org/vaxign/vaxign-ml/
Vax-ELAN	2021	0.0	Rank	B	W	B,E	https://vac.kamalrawal.in/vaxelan/v2

Year = year first released; Usage = percentage of publications since 2015 that have used the program; Select. = methodology for selecting candidates, where Filter denotes a rule-based filtering selection method comprising a series of conditional tests applied to each characteristic score or classification of a protein, and Rank denotes ranking candidates based on one single score collectively representing all predicted characteristic scores and classifications per protein; Char. = type of protein characteristics used in candidate selection, where B denotes biological characteristics e.g., subcellular location, transmembrane domains and P denotes physiochemical properties of amino acids e.g., charge, hydrophobicity; Mode = manner by which pipeline can be executed, where W denotes web server and S denotes standalone (i.e., pipeline installed on local computer); Org. = type of organism for which the pipeline has been designed: B (Bacteria), E (Eukaryote parasite), F (Fungi), T (Tumour protein), V (Viruses).

Table 4 Summary of outstanding reverse vaccinology issues and proposed solutions

#	Issue	Proposed solution
1	Usage of the term 'Reverse Vaccinology' to depict various workflow steps is inconsistent in publications	For universal understanding of the term 'Reverse Vaccinology', workflow steps should be restricted to those in classical RV, and RV acknowledged as one stage in the *in silico* approach to identifying vaccine candidates (see issue #2).
2	Commonly used workflow steps under the banner of RV have overlapped with other *in silico* approaches such as subtractive proteomics, computational vaccinology, and immunoinformatics	All related *in silico* approaches have the same end goal, which is to computationally identify vaccine candidates. A unified term of '*in silico* vaccine discovery' should be consistently used, especially in titles, abstracts, and/or keywords in future publications
3	All bioinformatics prediction programs have various levels of inherent inaccuracies.	Use several programs that perform the same task and derive a consensus.
4	The choice of bioinformatics programs to perform specific workflow tasks may occasionally be governed by its popularity or lack of choice, rather than on its merit.	An independent test using the same input data with known outcomes is sought for each existing and newly introduced program performing the same task.
5	Using a series of filtering workflow steps has the potential to inadvertently discard a true candidate due to only one erroneous characteristic prediction and/or a marginally below threshold value	All predicted protein characteristic can be simultaneously considered during candidate selection using ML, which is not severely compromised by one or two erroneous characteristics.
6	For most pathogens, there are insufficient numbers of verified protective antigens to use for ML training.	Use verified and 'likely' protective antigens from the target or related species. 'Likely' antigens are those published to induce an immune response *in vitro* or in an animal model, and those proteins experimentally shown to be naturally exposed to the immune system.
7	ML algorithms require positive and negative examples from training data. It is not clear what type of protein is truly a negative example e.g., only experimental testing can conclusively show a protein/peptide will not induce a protective immune response.	A repository is required for experimentally validated negative examples
8	There are possibly thousands of publications reporting immunogenicity results from *in vitro* and *in vivo* experiments. This is a vast unexploited resource for ML training data.	A single online repository, similar to Protegen, is required to record protective antigens from all past and future publications
9	Over reliance of RV tools online. This restricts achieving automated, high-throughput '*in silico* vaccine discovery.'	APIs and similar internet access tools are the key to achieving high-throughput automation. Program developers should be encouraged to provide this functionality
10	RV-related pipelines that have high-throughput capacity require third party installations, which can be challenging to users with limited computer administration skills.	RV pipeline developers should be encouraged to use software container technology.
11	Most predicted vaccine candidates are computationally verified, and their true	Currently, testing in an animal model is the recommended method to establish a protective

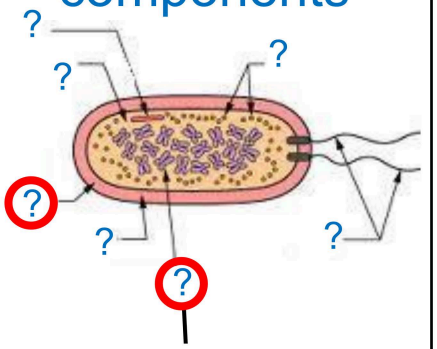
	protective efficacy is seldom established.	antigen. If not feasible, then evaluating the *in silico* process by predicting known protective antigens provides probabilities of protection when predicting anonymous candidates.
12	Difficult to quantify (or compare with published candidates) the contribution made by a protein/peptide to the overall vaccine efficacy due to different vaccine formulation variables (e.g., adjuvant, dose) and environmental variables (e.g. mouse model, challenge strain).	Research community needs to establish a standard protocol of candidate evaluation or at least determine a strategy for comparing study results.
13	High-throughput methods to perform *in silico* verification experiments on host–vaccine candidate interactions remain a tantalising goal.	Immune system simulators, such as C-ImmSim, show promise, but no correlation between predicted and the real *in vivo* vaccine immune responses have been evaluated.

API = application programming interfaces; ML = machine learning; RV = reverse vaccinology

Genome

Bioinformatics

Identify components



Potential antigens

Reverse Vaccinology

Vaccine candidates

Laboratory validation

Animal testing

Vaccine development