

METHOD, DEVICE AND STORAGE MEDIUM FOR IDENTIFYING PEPTIDES CAPABLE OF BINDING TO MHC-I/HLA-I AND BEING RECOGNIZED BY TCR

TECHNICAL FIELD

[0001]The present invention relates to the field of vaccines, and in particular to a method, a device, and a storage medium for identifying peptides capable of binding to MHC-I/HLA-I and being recognized by T-cell receptor (TCR).

BACKGROUND

[0002]Antigen screening technologies are mainly used to screen out antigens capable of eliciting the strongest immune responses during vaccine preparation. In the antigen screening process, it is necessary to evaluate antigens from different sources and then determine optimal combinations of antigens. Commonly used antigen screening technologies include biochip-based screening technologies and protein analysis-based screening technologies, and the like. Tumor vaccines are currently a research focus in the field of vaccines. The earliest tumor vaccines were reported in 1994, and in recent years, advances in technology have enabled tumor vaccines to achieve targeted treatment of patient-specific mutant antigens. By introducing peptides or mRNA of tumor-associated antigens or tumor-specific antigens into patients, the immune system of the patients can be activated. Breakthroughs in personalized tumor vaccine technologies critically depend on the development in the detection of neoantigens and prediction technologies for the preparation of neoantigen vaccines. Predicting the binding potential between specific mutant peptides and histocompatibility complexes is an important part of the detection of neoantigens, because tumor-specific antigens must be recognized by major histocompatibility complex (MHC) and presented to the cell surface before they can be further recognized by T cell receptors and ultimately lead to tumor cell killing. Therefore, it is of great significance to predict the binding between peptides and MHC-I/HLA-I, as well as whether the resulting pMHC complexes can maintain stable binding, be presented to the cell

surface and recognized by TCR.

[0003]In recent years, the field of neoantigens has developed rapidly, leading to the generation of a wide range of large-scale databases that can be used for training affinity prediction models for peptides binding to MHC-I/HLA-I. At the same time, a variety of algorithms have been applied to predict the binding potential between peptides and MHC-I/HLA-I, including traditional machine learning methods and deep learning methods with rapid development. Artificial neural networks (ANNs), convolutional neural networks (CNNs), stabilized matrix method (SMM), random forest methods, and other approaches have all been used to construct affinity prediction models. Although these algorithms have been reported to exhibit good performance when tested on test datasets, their performance is often only moderate when tested on completely independent datasets. The affinity prediction for peptide binding to MHC-I/HLA-I still suffers from relatively high error rates, particularly high false positive rates, which can have serious consequences in clinical applications. The causes of these issues mainly include the following aspects: first, it is difficult to obtain comprehensive and sufficiently large datasets for model training, and limited datasets make it challenging to learn more informative features; second, conventional algorithms applied to prediction models rarely take into account intrinsic features of protein sequences and sequence characteristics that may play important roles during protein–protein sequence interactions; third, both the representation of the full-length sequence of MHC-I/HLA-I and the representation of the pseudo-sequence limit the ability of models to specifically learn features of key sites involved in peptide binding, thereby adversely affecting model performance; fourth, most existing methods focus only on pMHC presentation, while features of mutant peptides, such as binding positions and the types of amino acid residues, are considered in isolation without further accounting for pMHC binding stability, pMHC presentation capability, and central tolerance mechanisms that play important roles in TCR (T-cell receptor) recognition, thus affecting the accurate selection of neoantigen mutant peptides.

SUMMARY

[0004]To address the above-mentioned problems, the present invention provides a method, a device, and a storage medium for identifying peptides capable of binding to MHC-I/HLA-I and

being recognized by TCR. The method enables more targeted identification of key features involved in recognition, transport, presentation, and TCR recognition of tumor-specific peptides, and addresses the limitation of insufficient data volume for model construction by applying novel modeling strategies for model training, feature extraction, and classification, thereby improving the accuracy of identifying peptides capable of binding to MHC-I/HLA-I and as well as the accuracy of neoantigen identification.

[0005]Embodiments of the present invention provide a method for identifying peptides capable of binding to MHC-I/HLA-I and being recognized by TCR, including following operations:

[0006]acquiring protein sequences, MHC-I/HLA-I sequences, and peptide sequences;

[0007]training a protein sequence tokenizer using the protein sequences;

[0008]training a protein representation model using the protein sequences tokenized by the protein sequence tokenizer to obtain a trained protein representation pre-trained model;

[0009]fine-tuning the trained protein representation pre-trained model by concatenating the MHC-I/HLA-I sequences and the peptide sequences tokenized by the protein sequence tokenizer to obtain a binary classification model;

[0010]identifying target wild-type/mutant peptides and classifying the mutant peptides using the binary classification model to obtain candidate neoantigen mutant peptides that are positively correlated with TCR recognition.

[0011]The binary classification model can be applied to predict peptides capable of binding to MHC-I/HLA-I, and candidate neoantigens can be predicted by applying the binary classification model in combination with the positions at which mutations occur.

[0012]Embodiments of the present invention provide a method for identifying peptides capable of binding to MHC-I/HLA-I and being recognized by TCR, including following operations:

[0013]acquiring protein sequences, MHC-I/HLA-I sequences, and peptide sequences;

[0014]training a protein sequence tokenizer using the protein sequences;

[0015]training a protein representation model using the protein sequences tokenized by the

protein sequence tokenizer to obtain a trained protein representation pre-trained model;

[0016]fine-tuning the trained protein representation pre-trained model by concatenating the MHC-I/HLA-I sequences and the peptide sequences tokenized by the protein sequence tokenizer to obtain a binary classification model; and

[0017]identifying peptides to be evaluated using the binary classification model to obtain peptides capable of binding to MHC-I/HLA-I.

[0018]The MHC-I refers to major histocompatibility complex class I (MHC class I), the HLA-I refers to major histocompatibility complex class I (MHC class I) in humans, and the TCR refers to a T cell receptor.

[0019]Preferably, the peptide sequences include positive peptide sequences and negative peptide sequences, in which the positive peptide sequences are the sequences of peptides capable of binding to MHC-I/HLA-I and are labeled as 1, and the negative peptide sequences are peptide sequences randomly generated from protein sequences based on features of the positive peptide sequences and are labeled as 0.

[0020]Preferably, the protein sequences include protein sequences collected from a protein database. Accordingly, collected protein sequences are used to train a tokenizer for high-frequency repeated amino acid residue combination motifs, to obtain the protein sequence tokenizer.

[0021]Preferably, training the protein representation model using the protein sequences tokenized by the protein sequence tokenizer to obtain the trained protein representation pre-trained model includes following operations:

[0022]inputting unlabeled protein sequences into the protein sequence tokenizer for tokenization, to obtain protein sequences segmented into functional amino acid residue combination motifs;

[0023]converting the protein sequences segmented into the functional amino acid residue combination motifs into numerically represented protein sequences; and

[0024]training the protein representation model using the numerically represented protein

sequences to obtain the trained protein representation pre-trained model.

[0025] Preferably, fine-tuning the trained protein representation pre-trained model using the MHC-I/HLA-I sequences and the peptide sequence tokenized by the protein sequence tokenizer to obtain the binary classification model includes following operations:

[0026] inputting labeled MHC-I/HLA-I sequences into the protein sequence tokenizer for tokenization, to obtain MHC-I/HLA-I sequences segmented into functional amino acid residue combination motifs, and converting the MHC-I/HLA-I sequences segmented into functional amino acid residue combination motifs into numerically represented MHC-I/HLA-I sequences;

[0027] converting labeled peptide sequences into numerically represented peptide sequences;

[0028] concatenating converted sequences, and fine-tuning the trained protein representation pre-trained model using concatenated sequences to obtain the binary classification model.

[0029] In specific embodiments, fine-tuning may include inheriting the original pre-trained model and adding a classification layer. In addition, on the basis of adding the classification layer, additional layers, such as fully connected layers, attention mechanism layers, or other suitable task layers may be further added.

[0030] In some embodiments, the concatenating is performed by direct concatenation, that is, directly concatenating the numerically represented MHC-I/HLA-I sequences with the numerically represented peptide sequences.

[0031] In another embodiments, the concatenating is performed by concatenating the numerically represented MHC-I/HLA-I sequences and the numerically represented peptide sequences, as well as additional numerically represented features related to binding between the peptides and MHC-I/HLA-I. Those skilled in the art may select an appropriate manner for the concatenating according to the objects to be concatenated. Exemplary manners for the concatenating include directly concatenating the numerically represented sequences and related features using torch.cat(), or performing addition, subtraction, multiplication, division, or other operations on feature tensors according to a customized concatenation manner.

[0032] In specific embodiments, the additional features related to binding between the peptides

and the MHC-I/HLA-I are one or more selected from a probability of cleavage of the protein sequences, a binding stability between the MHC-I/HLA-I and the peptides, and a presentation likelihood of complexes formed by the MHC-I/HLA-I and the peptides.

[0033]The probability of cleavage of the protein sequences may be obtained using suitable software by inputting the peptide sequences. Those skilled in the art will appreciate that any suitable software and any appropriate version thereof may be used, provided that the probability of cleavage of the protein sequences can be obtained based on the peptide sequences. For example, software such as pepsickle and iPCPS may be used for this purpose.

[0034]The binding stability between MHC-I/HLA-I and the peptides may be obtained using suitable software by inputting the peptide sequences and the MHC-I/HLA-I sequences. Those skilled in the art will appreciate that any suitable software with appropriate version thereof may be used, provided that the binding stability between MHC-I/HLA-I and the peptides can be obtained based on the peptide sequences and the MHC-I/HLA-I sequences. For example, software such as LKMHCstablepan and TLStab may be used for this purpose.

[0035]The presentation likelihood of the complex formed by MHC-I/HLA-I and the peptides may be obtained using suitable software by inputting the peptide sequences. Those skilled in the art will appreciate that any suitable software with appropriate version thereof may be used, provided that the presentation likelihood of the complex formed by MHC-I/HLA-I and the peptides can be obtained based on the peptide sequences and, optionally, the MHC-I/HLA-I sequences. For example, software such as DeepTAP, TAPPred, and TAPREG may be used for this purpose.

[0036]Preferably, prior to the fine-tuning, the method of the present invention further includes calculating other features related to binding between the peptides and MHC-I/HLA-I using the protein sequences, the MHC-I/HLA-I sequences, and the peptide sequences.

[0037]Preferably, identifying target wild-type/mutant peptides using the binary classification model includes following operations:

[0038]inputting the target wild-type/mutant peptides into the binary classification model, and outputting attention scores for each position of mutant peptides predicted as positive and attention scores of functional amino acid residue combination motifs of the MHC-I/HLA-I

sequences;

[0039]determining anchor positions of binding between the MHC-I/HLA-I sequences and the positive peptides as well as key representative positions of the MHC-I/HLA-I sequences, based on the attention scores for each position of the mutant peptides predicted as positive and the attention scores of the MHC-I/HLA-I sequences motifs, and classifying positive mutant peptides based on the anchor positions of binding between the MHC-I/HLA-I sequences and the peptides and mutation positions of the peptides.

[0040]Preferably, classifying the positive mutant peptides based on the anchor positions of binding between the MHC-I/HLA-I sequences and the positive mutant peptides as well as mutation positions relative to wild-type peptides includes following operations:

[0041]classifying the positive mutant peptides based on the anchor positions of binding between the MHC-I/HLA-I sequences and the peptides in combination with the mutation positions of, and determining, based on classification results, whether the positive mutant peptides are suitable for selection as candidate neoantigen mutant peptides that are positively correlated with TCR recognition.

[0042]Preferably, determining, based on the classification results, whether the positive mutant peptides are suitable for selection as candidate neoantigen mutant peptides that are positively correlated with TCR recognition includes following operations:

[0043]determining that the peptides are selectable as neoantigen mutant peptides when the wild-type peptides exhibit weak binding to MHC-I/HLA-I and weak TCR recognition (in general, wild-type peptides are weak in TCR recognition), and the mutant peptides have mutations at anchor positions for MHC-I/HLA-I and are strong in MHC-I/HLA-I binding and TCR recognition;

[0044]determining that the peptides are selectable as neoantigen mutant peptides when the wild-type peptides exhibit weak binding to MHC-I/HLA-I and weak TCR recognition (in general, wild-type peptides exhibit weak TCR recognition), and the mutant peptides have mutations at non-anchor positions for MHC-I/HLA-I and are strong in MHC-I/HLA-I binding and TCR recognition;

[0045]determining that the peptides are selectable as neoantigen mutant peptides when the wild-type peptides exhibit strong binding to MHC-I/HLA-I and weak TCR recognition (in general, wild-type peptides exhibit weak TCR recognition), and the mutant peptides have mutations at non-anchor positions for MHC-I/HLA-I and are strong in MHC-I/HLA-I binding and TCR recognition (herein, “strong TCR recognition” indicates that a sequence exposed to TCR is changed);

[0046]determining that the peptides are not selectable as neoantigen mutant peptides when the wild-type peptides exhibit strong binding to MHC-I/HLA-I and weak TCR recognition (in general, wild-type peptides exhibit weak TCR recognition), and the mutant peptides have mutations at anchor positions for MHC-I/HLA-I and are strong in MHC-I/HLA-I binding and weak in TCR recognition (herein, “weak TCR recognition” indicates that a sequence exposed to TCR remains unchanged).

[0047]The beneficial effects of the present invention are that, by applying deep learning algorithms based on natural language processing, we can identify the recognition, transport, and presentation of wild-type/mutant peptides with histocompatibility complex molecules MHC-I/HLA-I, recognize biological features involved in protein–protein interactions, determine anchor positions for peptide binding to MHC-I/HLA-I, construct more representative sequence characterization methods for MHC-I/HLA-I are constructed, and then introduce the influence of central tolerance mechanisms on TCR recognition to further promote the identification of tumor neoantigens.

BRIEF DESCRIPTION OF THE DRAWINGS

[0048]FIG. 1 is a flowchart of a method for identifying peptides capable of binding to MHC-I/HLA-I and being recognized by TCR according to the present invention.

[0049]FIG. 2 is a flowchart of an algorithm for identifying peptides capable of binding to MHC-I/HLA-I and being recognized by TCR according to the present invention.

[0050]FIG. 3 is a flowchart of an algorithm with additional features for identifying peptides capable of binding to MHC-I/HLA-I and being recognized by TCR according to the present invention.

DETAILED DESCRIPTION

[0051]It should be understood that the specific embodiments described herein are only for illustrating the present invention and are not intended to limit the present invention. In the following description, suffixes such as “module”, “component”, or “unit” used to denote elements are merely for convenience of description and do not have specific meanings. Therefore, the terms “module”, “component”, and “unit” may be used interchangeably.

[0052]As used herein, the term “concatenate/concatenation/concatenating” refers to a process of concatenating multiple features converted into numerical representations, thereby obtaining a concatenated sequence that can be used to fine-tune a protein representation pre-trained model.

[0053]As used herein, the term “pMHC” refers to a complex formed by binding a peptide to MHC-I/HLA-I.

[0054]As used herein, the terms “other features” and “more features” can be used interchangeably, and refer to features related to the binding of peptides and MHC-I/HLA-I calculated by using protein sequences, MHC-I/HLA-I sequences and peptide sequences. Exemplary features include a probability of cleavage of protein sequences, the binding stability of pMHC, and the presentation capability of pMHC.

[0055]The present invention provides a deep learning algorithm for identifying whether wild-type/mutant peptides are capable of binding to major histocompatibility complex class I (MHC-I/HLA-I), being transported and presented to the cell surface, and are recognized by TCR, thereby further promoting the identification of peptides capable of binding to MHC-I/HLA-I and the selection of tumor-specific neoantigens. Specifically, the method applies natural language processing techniques to protein sequence feature discovery and extraction for constructing a protein representation pre-trained model. The method utilizes protein sequences and protein-protein interaction datasets to extract amino acid residue combination motifs and key site features that are positively correlated with peptide binding, transport, and presentation, on the basis of the protein representation pre-trained model, the method fine-tunes the protein representation pre-trained model by collecting a positive peptide sequence dataset and constructing a negative (peptide sequence) dataset, thereby obtaining a model capable of

accurately inferring the binding potential of wild-type/mutant peptides to MHC-I/HLA-I. The model obtained by the above method can be applied to binding prediction for peptides of diverse types or for MHC-I/HLA-I types, provides an interpretable reflection of the biological characteristics of peptide presentation, and can be further applied to model development to obtain a prediction model with improved interpretability and accuracy compared with conventional algorithms. At the same time, the model further identifies recognition features of pMHC and TCR (T-cell receptor) following peptide binding to MHC-I/HLA-I, thereby further predicting whether a peptide has the capability to activate TCR (T-cell receptor) rather than merely binding to MHC, and ultimately enabling the selection of candidate neoantigens with higher potential.

[0056]FIG. 1 is a flowchart of a method for identifying peptides capable of binding to MHC-I and being recognized by TCR provided by the present invention. As shown in FIG. 1, the method includes following operations.

[0057]At S101, protein sequences, MHC-I/HLA-I sequences, and peptide sequences are acquired.

[0058]At S102, a protein sequence tokenizer is trained using the protein sequences.

[0059]At S103, a protein representation model is trained using the protein sequences tokenized by the protein sequence tokenizer, to obtain a trained protein representation pre-trained model;

[0060]At S104, the trained protein representation pre-trained model is fine-tuned using the MHC-I/HLA-I sequences and the peptide sequences tokenized by the protein sequence tokenizer, to obtain a binary classification model.

[0061]At S105, by using the binary classification model, the target wild-type/mutant peptides are identified, and the mutant peptides are classified, so as to obtain candidate neoantigen mutant peptides that are positively correlated with TCR recognition.

[0062]Specifically, the peptide sequences contain positive peptide sequences capable of binding to MHC-I/HLA-I and are labeled as 1 and negative peptide sequences randomly generated from the protein sequences based on features of the positive peptide sequences and are labeled as 0.

[0063]The protein sequences contain protein sequences collected from a protein database. Accordingly, the collected protein sequences are used to train a tokenizer for high-frequency repeated amino acid residue combination motifs, to obtain the protein sequence tokenizer.

[0064]Further, training the protein representation model using the protein sequences tokenized by the protein sequence tokenizer to obtain the trained protein representation pre-trained model includes: inputting unlabeled protein sequences into the protein sequence tokenizer for tokenization, to obtain protein sequences segmented into functional amino acid residue combination motifs; converting the protein sequences segmented into the functional amino acid residue combination motifs into numerically represented protein sequences; and training the protein representation model using the numerically represented protein sequences to obtain the trained protein representation pre-trained model.

[0065]More specifically, fine-tuning the trained protein representation pre-trained model using the MHC-I/HLA-I sequences and the peptide sequences that have been tokenized by the protein sequence tokenizer to obtain the binary classification model includes: inputting labeled MHC-I/HLA-I sequences into the protein sequence tokenizer for tokenization, to obtain MHC-I/HLA-I sequences segmented into functional amino acid residue combination motifs, and converting the MHC-I/HLA-I sequences segmented into functional amino acid residue combination motifs into numerically represented MHC-I/HLA-I sequences; converting labeled peptide sequences into numerically represented peptide sequences; concatenating the MHC-I/HLA-I sequences and the numerically represented peptide sequences to obtain concatenated sequences; and fine-tuning the trained protein representation pre-trained model using the concatenated sequences to obtain the binary classification model. The concatenating may be performed by direct concatenation, or by concatenation with additional numerically represented features including one or more selected from a probability of cleavage of protein sequences, a binding stability between MHC-I/HLA-I and peptides, and a presentation likelihood of complexes formed by MHC-I/HLA-I and peptides.

[0066]Further, identifying target wild-type/mutant peptides using the binary classification model includes: inputting the target wild-type/mutant peptides into the binary classification model, and outputting attention scores for each position of the mutant peptides predicted as positive, and attention scores of functional amino acid residue combination motifs of the

MHC-I/HLA-I sequences; determining anchor positions of binding between the MHC-I/HLA-I sequences and the positive peptides as well as key representative positions of the MHC-I/HLA-I sequences based on the attention scores for each position of the mutant peptides predicted as positive and the attention scores of the MHC-I/HLA-I sequences motifs; and classifying the positive mutant peptides based on the anchor positions of binding between the MHC-I/HLA-I sequences and the peptides and mutation positions of the peptides.

[0067]Further, classifying the positive mutant peptides based on the anchor positions of binding between the MHC-I/HLA-I sequences and the positive mutant peptides as well as mutation positions relative to wild-type peptides includes: classifying the positive mutant peptides based on the anchor positions of binding between the MHC-I/HLA-I sequences and the peptides in combination with positions of mutations, and determining, based on classification results, whether the positive mutant peptides are suitable for selection as candidate neoantigen mutant peptides that are positively correlated with TCR recognition.

[0068]More specifically, determining, based on the classification results, whether the positive mutant peptides are suitable for selection as candidate neoantigen mutant peptides that are positively correlated with TCR recognition includes: determining that the peptides are selectable as neoantigen mutant peptides when the wild-type peptides exhibit weak binding to MHC-I/HLA-I and weak TCR recognition, and the mutant peptides have mutations at anchor positions for MHC-I/HLA-I and are strong in MHC-I/HLA-I binding and TCR recognition; determining that the peptides are selectable as neoantigen mutant peptides when the wild-type peptides exhibit weak binding to MHC-I/HLA-I and weak TCR recognition, and the mutant peptides have mutations at non-anchor positions for MHC-I/HLA-I and are strong in MHC-I/HLA-I binding and TCR recognition; determining that the peptides are selectable as neoantigen mutant peptides when the wild-type peptides exhibits strong binding to MHC-I/HLA-I and weak TCR recognition, and the mutant peptides have mutations at non-anchor positions for MHC-I/HLA-I and are strong in MHC-I/HLA-I binding and TCR recognition; determining that the peptides are not selectable as neoantigen mutant peptides when the wild-type peptides exhibit strong binding to MHC-I/HLA-I and weak TCR recognition, and the mutant peptides have mutations at anchor positions for MHC-I/HLA-I and are strong in MHC-I/HLA-I binding and weak in TCR recognition.

[0069]An electronic device is provided by embodiments of the present application. The electronic device includes a memory, a processor, and a computer program stored in the memory and configured to be executed by the processor to implement a method, a device, and a storage medium for identifying peptides capable of binding to MHC-I/HLA-I and being recognized by TCR.

[0070]A computer-readable storage medium on which a computer program is stored is provided by embodiments of the present application. The computer program is executed by a processor to implement a method for identifying peptides capable of binding to MHC-I/HLA-I and being recognized by TCR.

[0071]In summary, the present invention provides a method for inferring the binding potential of peptides to MHC-I/HLA-I, by obtaining biological features from protein sequences and applying them to peptides and MHC-I/HLA-I binding prediction models with interpretability and accuracy based on a pre-trained model. Specifically, the method includes followings operations. The protein sequence tokenizer and the protein representation pre-trained model are constructed using lots of unlabeled protein data. Labeled training, validation, and test datasets including positive and negative datasets are created. The positive dataset includes positive peptide sequences obtained from databases through mass spectrometry or in vitro binding experiments and corresponding MHC-I/HLA-I subtypes; and the negative sequences are randomly generated from protein sequences according to length features of the positive peptide sequences. Negative peptides do not include peptides that are present in the positive peptides. MHC-I/HLA-I subtypes are tokenized using the protein sequence tokenizer, and the MHC-I/HLA-I sequences and peptide sequences are converted into numerically represented sequences. A binary classification model is constructed by fine-tuning the protein representation pre-trained model, and supervised learning is performed on the generated numerically represented data, thereby further improving model performance for the task. The binary classification model is further applied to predict attention scores of positive peptides and MHC-I/HLA-I sequence motifs, and then based on the predicted attention scores, the manner for representing MHC-I/HLA-I is obtained and anchor positions for peptide and MHC-I/HLA-I binding is inferred. At a deeper sense, the present invention provides a method capable of inferring whether peptides are capable of binding to MHC-I/HLA-I and activate TCR responses to elicit immune responses. For such method, more accurate neoantigens are obtained by

utilizing anchor positions for peptide–MHC-I/HLA-I binding, classifying wild-type and mutant peptides according to their MHC-I/HLA-I binding and TCR recognition types based on mutation positions within peptides and the influence of central tolerance, and further screening neoantigen peptides according to classification results.

[0072]The present invention relates to a novel method for predicting binding, transport, presentation, and TCR activation of wild-type/mutant peptides with MHC-I/HLA-I. Specifically, a model predictor with improved accuracy and biological interpretability is constructed by extracting protein residue combinations and key sites that play important roles during the binding process, better characterizing MHC-I/HLA-I sequences, and applying transfer learning techniques. Using the constructed model, key binding sites of MHC-I/HLA-I are explored, and neoantigens are further classified according to positions of mutations within peptides, thereby enhancing the accuracy of neoantigens identification. Compared with conventional classifiers, the method is more efficient and accurate, and has significant application value in tumor immunotherapy.

[0073]As demonstrated in the following embodiments, compared with other existing analysis methods, the present invention has the following advantages.

[0074](1) Biological features inherent to protein sequences are applied in model construction. By learning the combination patterns of amino acid residues within protein sequences, high-frequency amino acid residue combinations and key sites with significant features are found, enabling construction of a tokenizer that provides a more rational representation of amino acid sequences.

[0075](2) Natural language processing algorithms are applied to biological information language, namely protein sequences. That is, protein sequences are treated as sentences, and predictive models are constructed using natural language processing methods to learn features and contextual relationships in sequences.

[0076](3) Pre-trained models are constructed using lots of existing unlabeled data and fine-tuned using labeled data. By constructing a masked language model, the combination of masked amino acid residues can be inferred to learn contextual information and expression rules of protein sequences, enabling the pre-trained model to acquire rich language knowledge.

Subsequently, fine-tuning the pre-trained model is performed by limited binding data of labeled peptides and MHC-I/HLA-I, thereby constructing a classification model more suitable for affinity prediction.

[0077](4) By means of natural language processing models, attention scores for each position/motif within protein sequences can be output, thereby enabling determination of key sites in the binding of MHC-I/HLA-I to peptide, determination of key binding sites of peptides, and derivation of more representative and significant characterization methods for MHC-I/HLA-I sequences.

[0078](5) In addition to predicting the binding between peptides and MHC-I/HLA-I, features involving binding stability between peptides and MHC-I/HLA-I, transport, and presentation are also covered via concatenating. From the perspective of actual biological processes, features that play important roles in biological pathways are converted into numerical representations and combined with other features for model training and prediction, thereby improving the practical applicability of the model.

[0079](6) By determining positions of tumor-specific mutations within peptides, peptides are classified based on different pMHC binding and TCR recognition patterns, thereby further screening out neoantigen peptides with more accuracy. According to literature reports, antigen sequences not only can bind to MHC, but also can be recognized by TCR. The positions of tumor-specific antigen mutation simultaneously affect MHC binding and TCR recognition. Therefore, key MHC binding sites are obtained based on attention scores at each antigen peptide position, and neoantigens are further classified according to mutation positions, thereby excluding peptides that exhibit strong MHC-I/HLA-I binding potential but fail to activate TCR, and improving the accuracy of antigen peptide screening. The improved algorithm enables more accurate prediction of whether wild-type/mutant peptides possess binding, transport, and presentation potential with major histocompatibility complex (MHC) and TCR recognition potential, thereby promoting identification of tumor neoantigens and their clinical application.

[0080]The following detailed description of the content of this application is provided with reference to FIG. 2 and FIG.3.

[0081]Embodiment 1

[0082] 1.1 Dataset required for model training and screening on them

[0083] 1.1.1 Obtaining protein sequences from databases such as UniProt;

[0084] 1.1.2 Obtaining positive peptide sequences from databases such as IEDB, where the positive peptides must meet the following conditions:

[0085] The length thereof is 8-15 bp;

[0086] The MHC/HLA type for recognizing positive peptides is type I;

[0087] The corresponding wild-type peptide has a specific UniProt number.

[0088] 1.1.3 Obtaining HLA sequences from databases such as IPD-IMGT/HLA, and obtaining HLA-I sequences after multiple sequence alignment for characterizing HLA sequences.

[0089] 1.2 Constructing positive and negative datasets

[0090] The positive dataset consists of protein sequences/positive peptide sequences, in which the protein sequences are used to construct a pre-trained model and learn the expression features and contextual relationships. The positive peptide sequences are labeled peptides capable of binding to MHC-I/HLA-I. The negative sequences are randomly construct based on the positive sequences, meaning they are peptides that do not bind to MHC-I/HLA-I, deduced from positive peptides. The construction method for negative peptides is as follows:

[0091] Negative peptides with a length of 8-15 bp;

[0092] Randomly changing the amino acids within the same protein sequence as the positive peptide;

[0093] Removing peptides appearing in the positive peptide.

[0094] 1.3 Constructing and applying a protein sequence tokenizer

[0095] 1.3.1 amino acid residues are combined on a large number of proteins By using machine learning algorithms or statistical inference models, so as to infer frequently occurring amino acid residue combinations in the protein sequence. These frequently occurring amino acid

residue combinations are likely to be functional motifs, or in other words, play important roles when proteins perform their functions. Simultaneously, the locations of important motifs can be inferred. Unlike simply focusing on individual amino acid residues, the tokenizer is more helpful in exploring key features that enable proteins to perform their functions.

[0096]1.3.2 Protein sequence tokenization. All protein sequences are tokenized using the tokenizer, and potentially functional amino acid residues are grouped together to facilitate subsequent model construction.

[0097]In this embodiment of the invention, the combination patterns of amino acid residues in protein sequences can be learned from a protein database, and high-frequency combinations of residues in the protein sequence are grouped into a group to construct a protein sequence tokenizer. Specifically, a protein dataset is downloaded from a protein database to extract protein sequences therefrom; the protein sequences are trained based on the frequency of amino acid residues/residue combinations, left and right connectivity, and other characteristics of known protein sequences to obtain a tokenizer with interpretable amino acid residues of protein sequences.

[0098]1.4 Construction of protein representation pre-trained model

[0099]A large number of tokenized protein sequences imply protein function-related features, protein expression patterns, and relationships between motifs within the sequence. Therefore, a masked language model from natural language processing is used to randomly mask motifs in the protein sequences and thus predict the protein sequences, learning the features implied in the protein sequences. The masked language model is a self-supervised learning task designed to allow the model to learn the contextual information and language rules of motifs during pre-training. The model receives the input sequences, then randomly selects a certain ratio of motifs from the input text sequences to mask, and requires the model to predict these masked motifs. For example, following sentences are given:

[00100] Input sequence: MAF SAE DVL KE YD RRRR ME ALLL

[00101] Masked sequence: MAF SAE [mask] KE YD RRRR ME ALLL

[00102] In the present invention, unsupervised learning is performed on a large amount

of unlabeled protein data, to learn rich motif-related information of proteins. Then, transfer learning techniques are applied to adapt these models to specific labeled classification tasks. Specifically, a large number of unlabeled protein sequences are input into a protein sequence tokenizer to tokenize the protein sequences. The protein sequences are represented in the form of amino acid residue combinations. A masked language pre-training model (MLM) is constructed to randomly mask amino acid residue combinations in the tokenized protein sequences. The contextual information and expression rules of the protein sequences are learnt by the model when inferring the masked amino acid residue combinations, helping the model to get key residue features.

[00103] 1.5 Fine-tuning the protein representation pre-trained model

[00104] 1.5.1 The generated and labeled positive peptides capable of binding to MHC-I/HLA-I, and the negative peptides randomly generated based on the positive peptides, are labeled as 1 or 0, respectively. It is worth noting that these positive data are obtained through experiments, which means that only limited data for labeled training are available. This is the reason why a large amount of unlabeled protein data are used to create the pre-trained model first.

[00105] 1.5.2 The labeled data in step 1.5.1 are divided into a training dataset/a validation dataset/a test dataset at a certain ratio.

[00106] 1.5.3 Fine-tuning the generated and pre-trained model using the data from step 1.5.2, and input them to construct a binary classification model.

[00107] 1.5.4 The model structure is shown in FIG. 2.

[00108] 1.6 Application of the binary classification model

[00109] 1.6.1 Identifying peptides with the potential to bind to MHC-I/HLA-I.

[00110] 1.6.2 Extracting key features including functional amino acid residue combinations motifs, important function sites/binding sites, etc from protein sequences.

[00111] Using the above methods, peptides with the potential to bind to MHC-I/HLA-I can be predicted based on biological characteristics, and their key features can be identified.

[00112] 1.7. Identifying the impact of mutation positions on the potential of peptide binding to MHC-I/HLA-I and TCR recognition

[00113] Based on the obtained peptides and MHC-I/HLA-I binding sites in combination with the location of mutations in wild-type/mutated peptides, and considering the influence of central tolerance mechanisms on neoantigen screening, pMHC binding and TCR recognition are categorized into four types in which the first three types have neoantigen potential, and the last type is required to be excluded upon neoantigen screening.

[00114] Category 1: The wild-type peptides exhibit weak binding to MHC-I/HLA-I and weak TCR recognition, while the mutant peptides have mutations at MHC-I/HLA-I binding sites and are strong in MHC-I/HLA-I binding and TCR recognition. Under the absence of central tolerance, mutations occurring at anchor positions are also capable of activating TCR.

[00115] Category 2: The wild-type peptides exhibit weak binding to MHC-I/HLA-I and weak TCR recognition, while the mutant peptides have mutations at a non-MHC-I/HLA-I binding site and are strong in MHC-I/HLA-I binding and TCR recognition. Under the absence of central tolerance, mutations occurring at non-anchor positions are capable of activating TCR.

[00116] Category 3: The wild-type peptides exhibit strong binding to MHC-I/HLA-I and weak TCR recognition, while the mutant peptides have mutations at a non-MHC-I/HLA-I binding site and are strong in MHC-I/HLA-I binding and TCR recognition. In the presence of central tolerance, mutations occurring at non-anchor positions are capable of activating TCR.

[00117] Category 4: The wild-type peptides exhibit strong binding to MHC-I/HLA-I and weak TCR recognition, while the mutant peptides have mutations at an MHC-I/HLA-I binding site and are strong in MHC-I/HLA-I binding and weak in TCR recognition. In the presence of central tolerance, mutations occurring at non-anchor positions are not capable of activating TCR. This can be further explained as follows: under this combination, the mutation sites are located at MHC-I/HLA-I binding sites, and the sequence exposed to the TCR recognition interface is not altered. As a result, such peptide is still affected by central immune tolerance and is not suitable as a neoantigen peptide.

[00118] Features that play important roles in peptide and MHC-I/HLA-I recognition are inferred in the present invention based on biological characteristics and attention scores,

through the above steps including, representing MHC-I/HLA-I sequences and identifying key functional amino acid residue combinations and their positions; predicting, based on the binary classification model, the roles of amino acid residue combination types in the MHC-I/HLA-I sequence and their different positions in peptide recognition, binding, and presentation; and identifying key amino acid residue combinations at key sites according to attention scores; key binding sites of peptides are inferred to identify key functional amino acids and their positions. Using the obtained peptide binding sites information, wild-type/mutant peptides are classified with respect to MHC-I/HLA-I binding and TCR recognition based on the positions of mutations within the peptides. The influence of central tolerance on neoantigen screening is introduced, and neoantigen peptides are further excluded based on the classification results, thereby obtaining neoantigens with more accuracy.

[00119] Embodiment 2

[00120] In the present invention, protein sequences are obtained from the UniProt protein database to construct a tokenizer. Positive peptide binding sequences to MHC-I/HLA-I are obtained from the IEDB, and negative peptide are generated based on the positive sequences. Protein sequences tokenized by the tokenizer are used to construct a protein representation pre-trained model. Tokenized MHC-I/HLA-I sequences and peptides are converted into numerically represented formats and concatenated, and are input into the protein representation pre-trained model for fine-tuning, thereby ultimately obtaining a binary classification model capable of predicting peptide binding potential to MHC-I/HLA-I. Based on the trained model and algorithmic characteristics, key functional sites and motifs in protein sequences are extracted to represent MHC-I/HLA-I sequences; using the positions of tumor-specific mutations within antigenic peptides and their MHC binding potential, antigenic peptides are further classified, antigenic peptides that cannot be recognized by TCR due to the influence of central immune tolerance are excluded, and more accurate and comprehensive neoantigens are ultimately obtained. The specific steps include:

[00121] 2.1 Construction and use of the tokenizer

[00122] A tokenizer is trained with a vocabulary size set to 10,000. The tokenization results are as follows:

[00133] 2.4 Application of the binary classification model

[00134] 2.4.1 By inputting data into the binary classification model in accordance with the data format described in Section 2.3.2, the binding potential between peptides and MHC-I/HLA-I can be predicted, and the prediction results are shown in Table 2.

[00135] Table 2

peptide	MHC	Label	prediction	probability
KAISKNVLF	MRVT.....QTA-----	1	1	[0.050, 0.950]
RPTRPDKARR	MAV.....CKV-----	0	0	[0.990,0.001]

[00136] where “Label” denotes a ground-truth label, “Prediction” denotes a label predicted by the model, and “probability” denotes a model-predicted result [A,B] in which A represents a probability that the prediction result is negative, and B represents a probability that the prediction result is positive.

[00137] 2.4.2 Evaluation of the binary classification model

[00138] The evaluation dataset is a validation dataset that is independent of the dataset used for model training. The composition of the validation dataset is shown in Table 3.

[00139] Table 3

Number of positive peptides	Number of negative peptides	Number of HLA types
3,725	108,113	40

[00140] Model performance is calculated using the following formulas:

[00141] $Precision = \frac{TP}{TP+FP}$

[00142] $Recall/TPR = \frac{TP}{TP+FN}$

[00143] $FPR = \frac{FP}{TN+FP}$

[00144] TP: true positives; TN: true negatives; FP: false positives; FN: false negatives;

[00145] AUC is the area under the curve of the true positive rate (TPR) versus the false positive rate (FPR).

[00146] A comparison with the commonly used NetMHCpan 4.1 software is performed to evaluate model performance, and the results are shown in Table 4.

[00147] Table 4

Type	NetMHCpan4.1	LKMHCpan
AUC	0.77	0.85
Precision	0.75	0.84
Recall	0.79	0.85

[00148] 2.4.3 A method is provided for extracting MHC-I/HLA-I binding sites and obtaining positions of key functional amino acid residues, so as to biologically interpret the prediction performance of the model. Heat maps are depicted based on attention scores at different positions, in which brighter colors indicate a higher likelihood of being key binding sites between peptides and MHC. This also indirectly demonstrates the favorable performance of the model, and thus it can be further used for extracting key site features of MHC-I/HLA-I sequences in subsequent steps.

[00149] 2.4.4 A method is provided for representing protein sequences and identifying key functional amino acid residue combinations and their positions, so as to biologically interpret the prediction performance of the model. Heat maps are depicted based attention scores at each position along HLA-A/B/C sequences, to infer overall key binding sites of MHC-I/HLA-I sequences, thereby facilitating improved representation of MHC-I/HLA-I sequences. Compared with hotspot positions provided by common LKMHCpan or MHCflurry, (31, 33, 48, 69, 83, 86, 87, 90, 91, 93, 94, 97, 98, 100, 101, 104, 105, 108, 115, 119, 121, 123, 126, 138, 140, 142, 167, 171, 174, 176, 180, 182, 183, 187, 191, 195, and 223) are more comprehensive and more representative. When focusing on functional motif sequences at hotspots, and further extracting additional key sites beyond the above positions, such as positions 39 and 235, it would be more consistent with the functional significance embedded in protein sequences.

[00150] 2.5 Adding classification information for peptide binding to MHC-I/HLA-I and TCR recognition to promote more detailed and comprehensive neoantigen sequence identification.

[00151] Based on the above analysis, after obtaining the anchor positions for peptide binding to MHC-I/HLA-I, peptide types can be classified into four groups according to the binding potential of wild-type peptides/mutant peptides and the positions of tumor-specific mutations, as shown in Table 5.

[00152] Table 5

Category	Mutation Position	Wild-Type Peptide	Mutant Peptide	Suitable for Selection as Neoantigen
1	MHC binding site	Weak MHC binding & weak TCR recognition	Strong MHC binding & strong TCR recognition	Yes
2	Non-MHC binding site	Weak MHC binding & weak TCR recognition	Strong MHC binding & strong TCR recognition	Yes
3	Non-MHC binding site	Strong MHC binding & weak TCR recognition	Strong MHC binding & strong TCR recognition	Yes
4	MHC binding site	Strong MHC binding & weak TCR recognition	Strong MHC binding & weak TCR recognition	No

[00153] Antigen peptides belonging to Category 4 are excluded during neoantigen screening, thereby improving the accuracy of neoantigen prediction.

[00154] Embodiment 3

[00155] Steps 3.1 to 3.4 are the same as steps 1.1 to 1.4 in Embodiment 1.

[00156] 3.5 Fine-tuning the protein representation pre-trained model

[00157] 3.5.1 The generated and labeled positive peptides that can bind to MHC-I/HLA-I, and the negative peptides randomly generated based on the positive peptides, are labeled as 1 or 0, respectively. It is worth noting that these positive data are obtained through experiments, which means that only a limited data for labeled training data are available. This is the reason why a large amount of unlabeled protein data are used to create the pre-trained model first.

[00158] 3.5.2 The labeled MHC-I/HLA-I sequences and peptide sequences tokenized by

the protein tokenizer in step 3.5.1 are converted into numerical representations and concatenated. The concatenation strategy is as follows.

[00159] The probability of cleavage of the protein sequence is calculated based on the peptide and its corresponding protein sequence.

[00160] The binding stability between the peptide and MHC-I/HLA-I, as well as the presentation capability after binding are calculated.

[00161] The numerically represented probability of cleavage of the protein sequence, the binding stability between the peptide and MHC-I/HLA-I, and the presentation capability following binding between the peptide and MHC-I/HLA-I are concatenated with the numerically represented MHC-I/HLA-I sequences and peptide sequences tokenized by the protein tokenizer.

[00162] 3.5.3 The labeled data in step 3.5.2 are divided into a training dataset/a validation dataset/a test dataset at a certain ratio.

[00163] 3.5.4 Fine-tuning the generated and pre-trained model using the data from step 3.5.3, and input them to construct a binary classification model.

[00164] 3.5.5 The model structure is shown in FIG. 3.

[00165] Steps 3.6 to 3.7 are the same as steps 1.6 to 1.7 in Embodiment 1.

[00166] Features that play important roles in peptide and MHC-I/HLA-I recognition are inferred in the present invention based on biological characteristics and attention scores through representing MHC-I/HLA-I sequences and identifying key functional amino acid residue combinations and their positions; predicting, based on the binary classification model, the roles of amino acid residue combination types in the MHC-I/HLA-I sequence and their different positions in peptide recognition, binding, and presentation; and identifying key amino acid residue combinations at key sites according to attention scores; key binding sites of peptides are inferred to identify key functional amino acids and their positions. Using the obtained peptide binding sites information, wild-type/mutant peptides are classified with respect to MHC-I/HLA-I binding and TCR recognition based on the positions of mutations within the

peptides. The influence of pMHC binding stability, pMHC presentation capability, and central tolerance on neoantigen screening are introduced, and neoantigen peptides are further excluded based on the classification results, thereby obtaining neoantigens with more accuracy.

[00167] Embodiment 4

[00168] In the present invention, protein sequences are obtained from the UniProt protein database to construct a tokenizer. Positive peptide binding sequences to MHC-I/HLA-I are obtained from the IEDB, and negative peptide are generated based on the positive sequences. Protein sequences tokenized by the tokenizer are used to construct a protein representation pre-trained model. Tokenized MHC-I/HLA-I sequences and peptides are converted into numerically represented formats. Probability of cleavage of protein sequences is calculated based on peptides and their corresponding protein sequences, binding stability between peptides and MHC-I/HLA-I, as well as presentation likelihood following binding, are calculated based on peptides and MHC-I/HLA-I. The numerically represented MHC-I/HLA-I sequences and peptide sequences are concatenated together with the corresponding probability of cleavage of peptide, pMHC binding stability, and pMHC presentation likelihoods by torch.cat() as the concatenating manner, and are input into the protein representation pre-trained model for fine-tuning. The original pre-trained model is inherited while adding a fully connected layer and a classification layer, thereby ultimately obtaining a binary classification model capable of predicting the binding potential between peptides and MHC-I/HLA-I. Based on the trained model and algorithmic characteristics, key functional sites and motifs in protein sequences are extracted to represent MHC-I/HLA-I sequences. using the positions of tumor-specific mutations within antigenic peptides and their MHC binding potential, antigenic peptides are further classified, antigenic peptides that cannot be recognized by TCR due to the influence of central immune tolerance are excluded, and more accurate and comprehensive neoantigens are ultimately obtained. The specific steps include:

[00169] Steps 4.1 to 4.2 are the same as steps 2.1 to 2.2 in Embodiment 2.

[00170] 4.3 Fine-tuning the protein representation pre-trained model

[00171] 4.3.1 MHC-I/HLA-I sequences are tokenized according to the corresponding negative peptide generated by the positive peptide binding to MHC-I/HLA-I in which the data

for the positive or negative peptides are labeled as 1 or 0, respectively. The positive dataset is distributed as shown in Table 6.

[00172] Table 6

Number of positive peptides	Number of negative peptides	Number of HLA types
502,735	12,505,198	121

[00173] 4.3.2 Features that influence binding between peptides and MHC-I/HLA-I are calculated, including probability of cleavage of peptide, pMHC stability, and presentation likelihood of pMHC after binding. For example, the calculation mentioned may be performed as follows:

[00174] By using software pepsickle-v0-2-2, peptide sequences are input to output probability of cleavage scores at each position. The probability of cleavage of peptide is calculated as the average of cleavage scores of the C-terminal/N-terminal amino acids of the peptide, with a value range of [0, 1], in which higher score indicates a higher probability of cleavage.

[00175] By using the software LKMHCstablepan-v1, peptide sequences and MHC-I/HLA-I sequences are input to calculate pMHC binding stability, and a pMHC binding stability score is output, in which higher score indicates greater stability;

[00176] By using the software DeepTAP-v1.0.1, peptide sequences are input to calculate the likelihood of presentation to the cell surface following pMHC binding, and a presentation likelihood score is output with a value range of [0, 1], in which a higher score indicates a higher presentation likelihood.

[00177] 4.3.3 Example of the raw data format input into the protein representation pre-trained model is as follows:

[00178] E I E I C D G F,M RVT AP RTLL LLL WG AV ALT ET WAG SH SM RY FH
TS VS RPG RG EP RFI TVG YVD DT.....AGL AV ----- L AVV - VIG AV VAA VM - CRR
KS SGG KGG SY SQ AAC SD S AQG SD VSL TA -----,0.057,0.20,0.12,0
(Corresponding in order to: peptide sequence, MHC-I/HLA-I sequence, probability of cleavage

of peptide, score of pMHC binding stability, score of presentation likelihood of pMHC after binding, positive or negative).

[00179] After the MHC-I/HLA-I sequences and peptide sequences tokenized by the tokenizer are converted into numerically represented sequences, they are concatenated, along a specified dimension, with the probability of cleavage score of a peptide, the pMHC binding stability score, and the pMHC presentation likelihood score, and then input into the protein representation pre-trained model for model training.

[00180] 4.4 Application of the binary classification model

[00181] 4.4.1 By inputting data into the binary classification model in accordance with the data format described in Section 4.3.3, the binding potential between peptides and MHC-I/HLA-I can be predicted, and the prediction results are shown in Table 7.

[00182] Table 7

Peptide	MHC	Label	Prediction	Probability
KAISKNVLF	MRVT.....QTA-	1	1	[0.0320, 0.9681]

RPTRPDKARR	MAV.....CKV---	0	0	[0.9995, 0.0005]

[00183] where “Label” denotes a ground-truth label, “Prediction” denotes a label predicted by the model, and “probability” denotes a model-predicted result [A,B] in which A represents a probability that the prediction result is negative, and B represents a probability that the prediction result is positive.

[00184] 4.4.2 Evaluation of the binary classification model

[00185] The evaluation dataset is the same as that in step 2.4.2 of Embodiment 2.

[00186] The formula for calculating the model performance is the same as that in step 2.4.2 of Embodiment 2.

[00187] Compared with the commonly used NetMHCpan 4.1 software and the binary classification model constructed in Embodiment 2, model performance evaluation is conducted, as shown in Table 8:

[00188] Table 8

Type	NetMHCpan4.1	LKMHCpan(Embodiment 2)	LKMHCpan(Embodiment 4)
AUC	0.77	0.85	0.89
Precision	0.75	0.84	0.91
Recall	0.79	0.85	0.86

[00189] Steps 4.4.3 to 4.4.4 are the same as steps 2.4.3 to 2.4.4 in Embodiment 2.

[00190] Step 4.5 is the same as step 2.5 in Embodiment 2.

[00191] The preferred embodiments of the present invention have been described above with reference to the accompanying drawings, and the present invention is not limited thereto. Any modifications, equivalent substitutions, and improvements made by those skilled in the art without departing from the scope and spirit of the present invention shall fall within the scope of protection of the present invention.

CLAIMS

1. A method for identifying peptides capable of binding to MHC-I/HLA-I and being recognized by T-cell receptor (TCR), comprising:

acquiring protein sequences, major histocompatibility complex class I (MHC I)/ major histocompatibility complex class I in humans (HLA-I) sequences, and peptide sequences;

training a protein sequence tokenizer using the protein sequences;

training a protein representation model using the protein sequences tokenized by the protein sequence tokenizer to obtain a trained protein representation pre-trained model;

fine-tuning the trained protein representation pre-trained model using the MHC-I/HLA-I sequences and the peptide sequences tokenized by the protein sequence tokenizer, to obtain a binary classification model; and

identifying target wild-type/mutant peptides classifying mutant peptides using the binary classification model, to obtain candidate neoantigen mutant peptides that are positively correlated with TCR recognition;

wherein, the MHC-I refers to major histocompatibility complex class I; the HLA-I refers to major histocompatibility complex class I in humans; and the TCR refers to a T cell receptor.

2. The method according to claim 1, wherein the peptide sequences comprise positive peptide sequences and negative peptide sequences, wherein the positive peptide sequences are peptides capable of binding to MHC-I/HLA-I and are labeled as 1, and the negative peptide sequences are peptide sequences randomly generated from protein sequences based on features of the positive peptide sequences and are labeled as 0.

3. The method according to claim 2, wherein the protein sequences comprise protein sequences collected from a protein database; accordingly, collected protein sequences are used to train a tokenizer with high-frequency repeated amino acid residue combination motifs, to obtain the protein sequence tokenizer.

4. The method according to claim 3, wherein training the protein representation model using the protein sequences tokenized by the protein sequence tokenizer to obtain the trained protein representation pre-trained model comprises:

inputting unlabeled protein sequences into the protein sequence tokenizer for tokenization, to obtain protein sequences segmented into functional amino acid residue combination motifs;

converting the protein sequences segmented into the functional amino acid residue combination motifs into numerically represented protein sequences; and

training the protein representation model using the numerically represented protein sequences to obtain the trained protein representation pre-trained model.

5. The method according to claim 4, wherein fine-tuning the trained protein representation pre-trained model using the MHC-I/HLA-I sequences and the peptide sequences that have been tokenized by the protein sequence tokenizer to obtain the binary classification model comprises:

inputting labeled MHC-I/HLA-I sequences into the protein sequence tokenizer for tokenization, to obtain MHC-I/HLA-I sequences segmented into functional amino acid residue combination motifs, and converting the MHC-I/HLA-I sequences segmented into functional amino acid residue combination motifs into numerically represented MHC-I/HLA-I sequences;

converting labeled peptide sequences into numerically represented peptide sequences; and

concatenating the numerically represented MHC-I/HLA-I sequences and the numerically represented peptide sequences to obtain concatenated sequences, and fine-tuning the trained protein representation pre-trained model using the concatenated sequences to obtain the binary classification model.

6. The method according to claim 5, wherein the concatenating further comprises concatenating the numerically represented MHC-I/HLA-I sequences, the numerically represented peptide sequences, and numerically represented additional features related to binding between the peptides and the MHC-I/HLA-I, wherein the additional features are one or more selected from a probability of cleavage of the protein sequences, a binding stability between the MHC-I/HLA-I and the peptides, and a presentation likelihood of complexes formed by the MHC-I/HLA-I and the peptides.

7. The method according to claim 6, wherein prior to the fine-tuning, the method further comprises calculating the additional features related to binding between the peptides and the

MHC-I/HLA-I using the protein sequences, the MHC-I/HLA-I sequences, and the peptide sequences.

8. The method according to claim 5, wherein identifying target wild-type/mutant peptides using the binary classification model comprises:

inputting the target wild-type/mutant peptides into the binary classification model, and outputting attention scores for each position of the mutant peptides predicted as positive, and attention scores of functional amino acid residue combination motifs of the MHC-I/HLA-I sequences; and

determining anchor positions of binding between the MHC-I/HLA-I sequences and the positive peptides, as well as key representative positions of the MHC-I/HLA-I sequences, based on the attention scores for each position of the mutant peptides predicted as positive and the attention scores of the MHC-I/HLA-I sequences motifs, and classifying positive mutant peptides based on the anchor positions of binding between the MHC-I/HLA-I sequences and the peptides and mutation positions of the peptides.

9. The method according to claim 8, wherein classifying based on the anchor positions of binding between the MHC-I/HLA-I sequences and the positive mutant peptides as well as mutation positions relative to wild-type peptides comprises:

classifying the positive mutant peptides based on the anchor positions of binding between the MHC-I/HLA-I sequences and the peptides in combination with the mutation positions, and determining, based on classification results, whether the positive mutant peptides are suitable for selection as candidate neoantigen mutant peptides that are positively correlated with TCR recognition.

10. The method according to claim 9, wherein determining, based on the classification results, whether the positive mutant peptides are suitable for selection as candidate neoantigen mutant peptides that are positively correlated with TCR recognition comprises:

determining that the peptides are selectable as neoantigen mutant peptides when the wild-type peptides exhibit weak binding to MHC-I/HLA-I and weak TCR recognition, and the mutant peptides have mutations at anchor positions for MHC-I/HLA-I and are strong in MHC-I/HLA-I binding and TCR recognition;

determining that the peptides are selectable as neoantigen mutant peptides when the wild-type peptides exhibit weak binding to MHC-I/HLA-I and weak TCR recognition, and the mutant peptides have mutations at non-anchor positions for MHC-I/HLA-I and are strong in MHC-I/HLA-I binding and TCR recognition;

determining that the peptides are selectable as neoantigen mutant peptides when the wild-type peptides exhibit strong binding to MHC-I/HLA-I and weak TCR recognition, and the mutant peptides have mutations at non-anchor positions for MHC-I/HLA-I and are strong in MHC-I/HLA-I binding and strong TCR recognition;

determining that the peptides are not selectable as neoantigen mutant peptides when the wild-type peptides exhibit strong binding to MHC-I/HLA-I and weak TCR recognition, and the mutant peptides have mutations at anchor positions for MHC-I/HLA-I and are strong in MHC-I/HLA-I binding and weak in TCR recognition.

11. An electronic device, comprising: a memory; a processor; and a computer program, wherein the computer program is stored in the memory and configured to be executed by the processor to implement the method according to any one of claims 1-10.

12. A computer-readable storage medium, having a computer program stored thereon, wherein the computer program is executed by a processor to implement the method according to any one of claims 1-10.