**Task 1**

**Coloums selection**

According to [Ido Springer](https://pubmed.ncbi.nlm.nih.gov/?term=Springer+I&cauthor_id=33981311), [Nili Tickotsky](https://pubmed.ncbi.nlm.nih.gov/?term=Tickotsky+N&cauthor_id=33981311), [Yoram Louzoun](https://pubmed.ncbi.nlm.nih.gov/?term=Louzoun+Y&cauthor_id=33981311) in 2021, who discussed ERGO-II, a model that assesses the contribution of T Cell Receptor (TCR) α and β CDR3 sequences to peptide binding prediction. It highlights that for most peptides, the β chain CDR3 sequence contributes most significantly to binding prediction accuracy, followed by the β chain V and J genes, and then the α chain. The MHC allele is noted as the least contributing component. This signifies the crucial role of the β chain CDR3 in the interaction, with additional contributions from other components enhancing prediction accuracy.

T lymphocytes (T cells) play a crucial role in cellular immune responses as they can recognize characteristic antigen peptides bound to the Major Histocompatibility Complex (MHC). This recognition is controlled by the T cell receptor (TCR) dimer. In the TCRβ chain, the Complementarity-Determining Region (CDR) 1 loop and CDR2 loop contact the MHC (α-helix), while the highly variable CDR3 region primarily interacts with the peptide. The CDR3 regions in both TCRα and TCRβ chains exhibit the highest sequence diversity, which is a key determinant of receptor binding specificity. CDR1 and CDR2 are determined by the V gene used in the TCR, whereas CDR3 is determined by the recombination of V (to D in TCRβ), J, and the addition or deletion of nucleotides at the V(D)J junction (in the α chain, V to J).

**Task 2**

**One-hot encoding**

One-hot encoding is a widely used approach that is applied to categorical variables. It turns every category into a binary vector representation. And only one element is “hot”, and the rest are “cold”. For example, if a feature contains categories such as “red”, “blue” and “green”. One-hot encoding would create 3 binary columns, each column representing one specific colour. If an instance corresponds to the colour “green”, its vector would be [0, 1, 0].

The limitations of using one-hot encoding is apparent:

**1, Dimensional Explosion**: CDR3 is known for its high diversity, leading to a huge potential sequence space. Using one-hot encoding means each amino acid position will become a separate feature. Moreover, one-hot coding may produce very sparse matrices, which may affect the performance of the model.

**2, Loss of Sequence Information**: One-hot encoding represents each amino acid as a separate, unrelated feature vector. This results in ignoring the interactions between amino acids and their positions in the sequence.

In order to better address these issues, we have found the following two approaches:

**K-mer counting**

The k-mer algorithm is a widely used method in bioinformatics, especially in genome sequence analysis. Simply put, a k-mer is just a sequence of k characters in a string (or nucleotides in a DNA sequence) (Clavijo, 2018).

Decomposing the sequence into its k-mer for analysis allows this set of fixed-size chunks to be analysed instead of the sequence, which can be more efficient. K-mers are very useful in sequence matching (having a rich history of string matching with n-grams). Set operations are faster and easier, and there are many off-the-shelf algorithms and techniques for using them (Clavijo, 2018).

Here is a simple example showing how the algorithm works:



By k-mer coding, biological sequences are decomposed into subsequences of length k. The frequency of occurrence of each k-mer is counted and these frequencies are converted into a vector of values. In this way, each sequence corresponds to a vector. This allows further machine learning and sequence analysis.

Experiments by Lv et al. show that compared with the model encoded by 1-mer one-hot, the model using 2-mer one-hot encoding shows better performance, stability and robustness, effectively improving the identification of rice genome. Ability of the DNA N6-methyladenine site (Lv, Z. et al., 2021).

**Embedding**

In contrast with one-hot and k-mer, embedding appears to be more complex, but it can better capture the hidden biological information of sequences and also reduce the dimensionality.

Embedding technique is a method of transforming high-dimensional data into a low-dimensional space, which makes it easier to handle machine learning tasks with large-scale sparse inputs such as word vectors. With this technique, semantically similar elements are represented close to each other in the low-dimensional space, thus effectively capturing the semantic features of the input data.

Let us explain how embedding techniques deal with biological sequences through a concrete example. Take the example of protein language models (pLMs), which treat amino acids in a protein sequence as tokens (equivalent to vocabulary in NLP) and consider the entire protein sequence as a sentence. These models are first trained without any annotation by self-supervised learning, with the aim of learning to predict the amino acids that are masked in the sequence. This process allows the models to capture the intrinsic features and patterns of the protein sequence. In this way, the models learn not just the individual amino acids in the sequence, but the deep structure and functional patterns of the entire sequence, thus providing an effective means of understanding the complexity of proteins (Elnaggar et al., 2022).

Clavijo, B. J. (2018). K-Mer counting, part I: Introduction. BioInfoLogics. <https://bioinfologics.github.io/post/2018/09/17/k-mer-counting-part-i-introduction/>

Sidhom, JW., Larman, H.B., Pardoll, D.M. *et al.* DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nat Commun* **12**, 1605 (2021). <https://doi.org/10.1038/s41467-021-21879-w>

Guangyuan Li, Balaji Iyer, V B Surya Prasath, Yizhao Ni, Nathan Salomonis, DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity, *Briefings in Bioinformatics*, Volume 22, Issue 6, November 2021, bbab160, <https://doi.org/10.1093/bib/bbab160>

Lv, Z., Ding, H., Wang, L., & Zou, Q. (2021). A convolutional neural network using dinucleotide one-hot encoder for identifying DNA N6-methyladenine sites in the rice genome. Neurocomputing, 422, 214-221.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2022). ProtTrans: Toward understanding the language of life through self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 7112–7127. https://doi.org/10.1109/tpami.2021.3095381

Springer I, Tickotsky N, Louzoun Y. Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. Front Immunol. 2021 Apr 26;12:664514. doi: 10.3389/fimmu.2021.664514. PMID: 33981311; PMCID: PMC8107833.

5 min

数据处理 task2 目前的进度/挑战 下一步计划

1. 为什么要选择这些colum

如何处理数据

2.onehot