OXFORD

BIOLOGY
Methods & Protocols

Biology Methods and Protocols, 2024, bpae055
https://doi.org/10.1093/biomethods/bpae055
Advance Access Publication Date: 30 July 2024
Methods Article

Harmonizing immune cell sequences for computational analysis with large language models

Areej Alsaafin[1] and Hamid R. Tizhoosh[1,*]

[1]Department of Artificial Intelligence & Informatics, KIMIA Lab, Mayo Clinic, Rochester, MN, 55905, United States

*Corresponding author. Laboratory for Knowledge Inference in Medical Image Analysis (KIMIA Lab), Department of Artificial Intelligence and Informatics, Mayo Clinic, 200 1st Street SW, Rochester, MN 55905, United States. E-mail: tizhoosh.hamid@mayo.edu

Abstract

We present SEQuence Weighted Alignment for Sorting and Harmonization (Seqwash), an algorithm designed to process sequencing profiles utilizing large language models. Seqwash *harmonizes* immune cell sequences into a unified representation, empowering LLMs to embed meaningful patterns while eliminating irrelevant information. Evaluations using immune cell sequencing data showcase Seqwash's efficacy in standardizing profiles, leading to improved feature quality and enhanced performance in both supervised and unsupervised downstream tasks for sequencing data.

Keywords: Immune cells, sequencing, large language models

Immune repertoires represent a critical modality affected by the challenge of processing biological sequences at the patient level [1]. These repertoires, comprising T-cell receptor (TCR) and B-cell receptor (BCR) sequences, serve as crucial indicators of a patient's immune response, reflecting the diverse repertoires of antigen-specific receptors generated by lymphocytes [2]. However, the ordering and length of these sequences within profiles exhibit significant variability among patients, posing challenges for accurate analysis and interpretation using artificial intelligence (AI) and machine learning techniques. While the sequence order itself does not directly contribute to understanding the immune response, Large Language Models (LLMs) still analyze these diverse orderings as part of extracting patterns from the sequences [3]. LLMs have revolutionized natural language processing (NLP) tasks as they are trained on vast amounts of textual data [4]. LLMs are powerful AI models that aim to capture the context from human language by analyzing the sequential order and relationships between words. While biological sequences, in general, can be treated similarly to textual data, there are unique challenges in processing sequences that do not arise in traditional text data. For instance, TCR and BCR sequences lack inherent order-related meaning across patients, posing a challenge that limits the utilization of LLMs for textual-like data [3]. Enabling LLMs to learn a low-dimensional representation of patients' sequencing data can lead to capturing the semantic relationships and similarities between patients [5, 6]. In addition, it enables embedding the patient's molecular signatures into a dense feature vector that facilitates building a multimodal framework [7, 8].

Deciphering patterns encoded in immune profiles presents a significant challenge, especially with large-scale data, due to various factors such as varying lengths and hidden patterns contributing to the complexity of processing these data with LLMs. The focus of existing approaches in the literature tends to be on single-sequence analysis [9, 10], overlooking the holistic signature that can be captured at the patient level, that is analysis of all sequences as one input. This single-sequence processing prevents the representation of patients through vectorized feature embeddings, hindering meaningful comparisons among patients. Moreover, single-sequence analysis is a major roadblock to the design of a multimodal approach that considers data from different sources. For example, a multimodal approach that processes histopathology whole slide images (WSIs) and corresponding molecular data requires matching the two modalities at the patient level.

In this article, we present SEQuence Weighted Alignment for Sorting and Harmonization (short Seqwash), an algorithm designed to facilitate the processing of the entirety of immune cell sequencing profiles for use by LLMs, models generally intended for NLP (see Fig. 1). Seqwash enables LLMs to extract meaningful patterns from a set of TCR/BCR sequences by eliminating non-related and noisy information. Seqwash aligns sequences within each profile into a unified representation before feature extraction using an LLM. The objective is to *harmonize* sequences across patients, thereby providing a more accurate latent representation. This approach ensures that each patient's immune signature is represented with a dense embedding, capturing essential information while mitigating the impact of varying sequence orders and lengths within each profile.

Materials and methods

Biological sequence data

Raw RNA-seq files were obtained from TCGA (The Cancer Genome Atlas) to reconstruct the immune repertoire of every patient. In raw RNA-seq data, the expressed TCR and BCR sequences can be identified by analyzing the reads that align to these specific genomic regions [11]. In this work, TRUST4 [11] was

Received: 26 April 2024. **Revised:** 07 July 2024. **Editorial decision:** 23 July 2024. **Accepted:** 29 July 2024

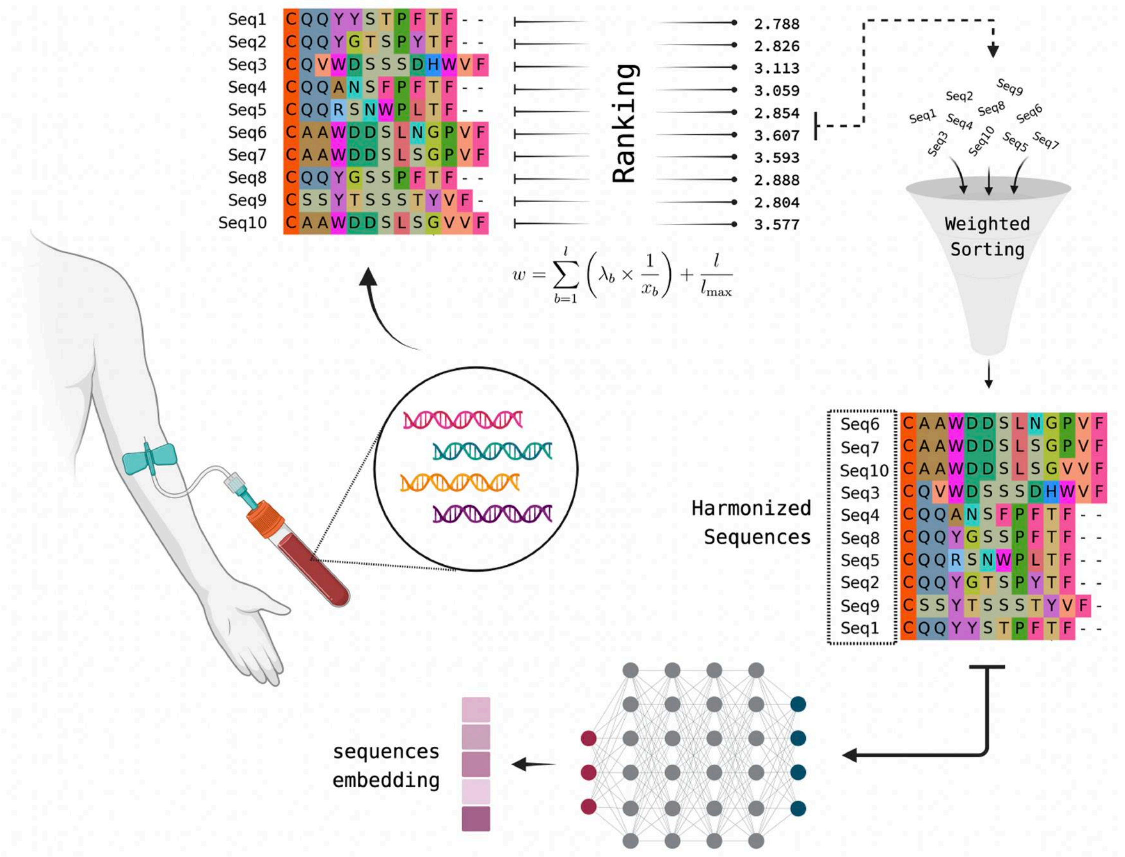

Figure 1 The mechanism of Seqwash in harmonizing the sequence profiles of a patient involves calculating a ranking weight for every sequence and then applying weighted sorting to align the sequences based on the calculated ranking weights before feeding them into an LLM.

employed to obtain the TCR and BCR sequences of each patient from their RNA-seq profiles.

TCGA Datasets for two primary sites, namely lung and kidney, were included in the evaluation as they have reasonable number of samples per subtype. Lung dataset includes 1181 lung adenocarcinoma (LUAD) cases and 1085 lung squamous cell carcinoma (LUSC) cases. Kidney dataset includes 1035 kidney renal clear cell carcinoma (KIRC), 394 kidney renal papillary cell carcinoma (KIRP), and 69 kidney chromophobes (KICH). Infrequent sequences were filtered out by excluding those that were not common to sufficient proportion of patients within each subtype class. A threshold of 15% was experimentally determined and applied to the patient profiles.

Sequence ranking weights

Seqwash was used to harmonize the immune repertoire of each patient. Seqwash hinges on two pivotal factors for calculating the ranking weight of each sequence within a given profile (see Algorithm 1): *sequence length* and *base position* within the sequence. Sequence length represents the number of bases within a sequence, while base position denotes the order of a base within a given sequence. The first step of Seqwash is to assign a weight to every base. Since we are dealing with TCR and BCR sequences here, which consist of 20 amino acids (bases), arbitrary but fixed weights were assigned to each amino acid (Algorithm 1, Line 1). These factors are used to calculate the ranking weight ω for each sequence through

Algorithm 1 ***Seqwash*** **Approach for Harmonization of biological Sequences**

```
1: Initialize $\lambda$ as a dictionary with arbitrary weights for each
   amino acid
2: procedure COMPUTE RANKING WEIGHT $(S, \lambda)$
3:     for each sequence $s_i \in S$ do
4:         Initialize $\omega_s \leftarrow 0$
5:         $l \leftarrow Length\ of\ s_i$
6:         for each base b in $s_i$ do
7:             $\lambda_b \leftarrow \lambda[b]$
8:             $x_b \leftarrow Position\ of\ b\ in\ s_i$
9:             $\omega_{s_i} \leftarrow \omega_{s_i} + \left(\lambda_b \times \frac{1}{x_b}\right)$
10:    $l_{max} \leftarrow Maximum\ sequence\ length\ in\ S$
11:    $\omega_{s_i} \leftarrow \omega_{s_i} + \frac{l}{l_{max}}$
12:    $S_h \leftarrow descending\_sort(S, key = \omega_{s_i})$
13:    return $S_h$
14: procedure FEATUREEXTRACTION $(S_h, \mathcal{G})$
15:    $g \leftarrow \mathcal{G}(S_h)$ % extract embedding g using the LLM model $\mathcal{G}$
16:    return g
```

$$\omega = \sum_{b=1}^{l} \left(\lambda_b \times \frac{1}{x_b} \right) + \frac{l}{l_{\max}} \tag{1}$$

In this context, λ_b represents the arbitrary weight assigned to each base b, while x_b indicates the position of base b within the sequence, starting from 1. The variable l stands for the length of the sequence, and $l_{\max}$ signifies the maximum sequence length within the set of sequences, that is the available population. Equation (1) encapsulates the process of computing the weighted sum of the base factors for every base within a sequence, considering their positions. This sum, combined with the ratio of sequence length to the maximum sequence length, ultimately determines the ranking weight. The ultimate goal of utilizing Equation (1) is to unify the sequence orders (indices) across patient profiles, primarily leveraging the shared foundation of bases constituting these sequences. By employing a consistent set of parameters for these bases, along with the sequential arrangements of bases and the length of the sequence that a given base is part of, we can effectively prioritize the sequences based on a standardized criterion. As illustrated in Algorithm 1, Lines 2–11, the ranking weight is computed for sequence s_i, where $s_i \in S$.

Weighted sorting

Once the ranking weights are calculated for the entire set of sequences of a patient, the sequences are sorted in descending order based on these weight values to generate the harmonized set S_h (Algorithm 1, Lines 12–13). This weighted sorting process is critical to unify the sequence representation across the patients. This is achieved by arranging the sequences within every profile in which the sequences with higher ranking weights occupy prominent positions in the resulting list.

Feature extraction

With the sequences harmonized, they are now ready for feature extraction using pre-trained LLMs (Algorithm 1, Lines 15–16). Five LLMs have been employed to extract features from both raw and harmonized immune cell profiles, namely DistilBERT, ALBERT, XLNet, XLM-RoBERTa, and DeBERTa. The rational for employing these LLMs for Seqwash evaluation is to leverage their pre-trained knowledge representations and utilize them to produce compact and expressive embeddings of fixed sizes across the patients to enable consistent comparisons. Each of these LLMs brings its own unique architecture and pre-training methodologies, which could potentially capture different aspects of the input data. DistilBERT, for instance, is a distilled version of BERT (Bidirectional Encoder Representations from Transformers), designed to be faster and more memory-efficient while maintaining much of BERT's performance. ALBERT, on the other hand, introduces parameter reduction techniques to improve efficiency without sacrificing accuracy. XLNet explores an autoregressive model that considers all permutations of the input sequence, enhancing its understanding of language coherence. XLM-RoBERTa extends BERT to multilingual settings, potentially beneficial for datasets with diverse linguistic characteristics. DeBERTa incorporates self-attention mechanisms with relative position representations, aiming to capture long-range dependencies effectively. All these models produce a 768-dimensional embedding for a given immune repertoire. For all models, we applied average pooling at the last hidden layer to generate a single feature vector for the complete profile from the sequence feature vectors. This method resulted in a better performance than extracting *CLS (classification)* layer.

Statistical significance analysis

The variance in the extracted feature vectors was analyzed using the Analysis of Variance (ANOVA) test to determine whether significant differences exist in the mean feature values across the primary diagnoses (subtypes) in each dataset. We calculated the *P*-value of each feature (768 features for each patient) extracted from every LLM for every patient in lung and kidney datasets. We set the significance level at 0.001 to determine the significance of each feature in terms of differentiating between different subtypes.

Patient-level supervised classification

The performance of Seqwash was assessed in a supervised classification task. Support Vector Machine (SVM) classifiers were trained on 70% of the datasets, with the remaining 30% used to test their performance. Two classifiers were trained and tested on each of the lung and kidney datasets: one using features of raw sequences and another using features of sequences harmonized by Seqwash. The hyperparameters of each SVM classifier were fine-tuned using the *grid search* method. Macro F1-score was calculated for each classifier, which averages the F1-scores of each class, treating all classes equally.

Patient-level unsupervised similarity search and retrieval

To assess the resulted representation on an unsupervised task, we performed search and retrieval using leave-one-out validation utilizing all the samples in each dataset as no training is needed for nearest neighbor search. Leave-one-out validation is the extreme scenario of cross-validation, where the fold size is $k = 1$. Hence, every patient is treated as a query once and therefore excluded from the archive when searching for the top-n matches. We calculated the Euclidean distance between the query sample and every other sample to identify the most similar samples with the minimum distances to the query. Majority vote among top-5 retrievals (MV5) was calculated to determine the subtypes. The majority vote criterion indicates that at least $n/2 + 1$ of the top-n samples should belong to the same class as the query. Macro F1-score was calculated using the primary diagnosis of the retrieved samples.

Results

Seqwash was employed to standardize immune repertoires of patients according to a unified criterion illustrated in the Materials and Methods section, facilitating different LLMs in capturing meaningful information and producing predictive embeddings for an immune repertoire. The datasets were acquired from the Genomic Data Commons repository of The Cancer Genome Atlas (TCGA), encompassing samples from lung and kidney as primary sites.

To visually assess the impact of Seqwash on immune repertoires, we generated heatmaps for randomly selected sequencing profiles from the TCGA lung cancer dataset, both before and after applying Seqwash, as shown in Fig. 2(a). Each amino acid within a TCR/BCR sequence is represented by a distinct color. Notably, both heatmaps show the profiles of the same patients. The first heatmap displays raw sequences pre-Seqwash application, revealing no consistent pattern across patient profiles. Conversely, in the second heatmap, noticeable patterns across the patient profiles enable the identification of the beginning of each profile.

We assessed the Seqwash features by analyzing embeddings produced by five LLMs: DistilBERT, ALBERT, XLNet, XLM-RoBERTa, and DeBERTa. We conducted ANOVA tests on the extracted embeddings of patients from lung and kidney cancer datasets from TCGA, aiming to identify significant features with a $P < 0.001$, as shown in Fig. 2(b). The application of Seqwash resulted in an increased number of statistically significant

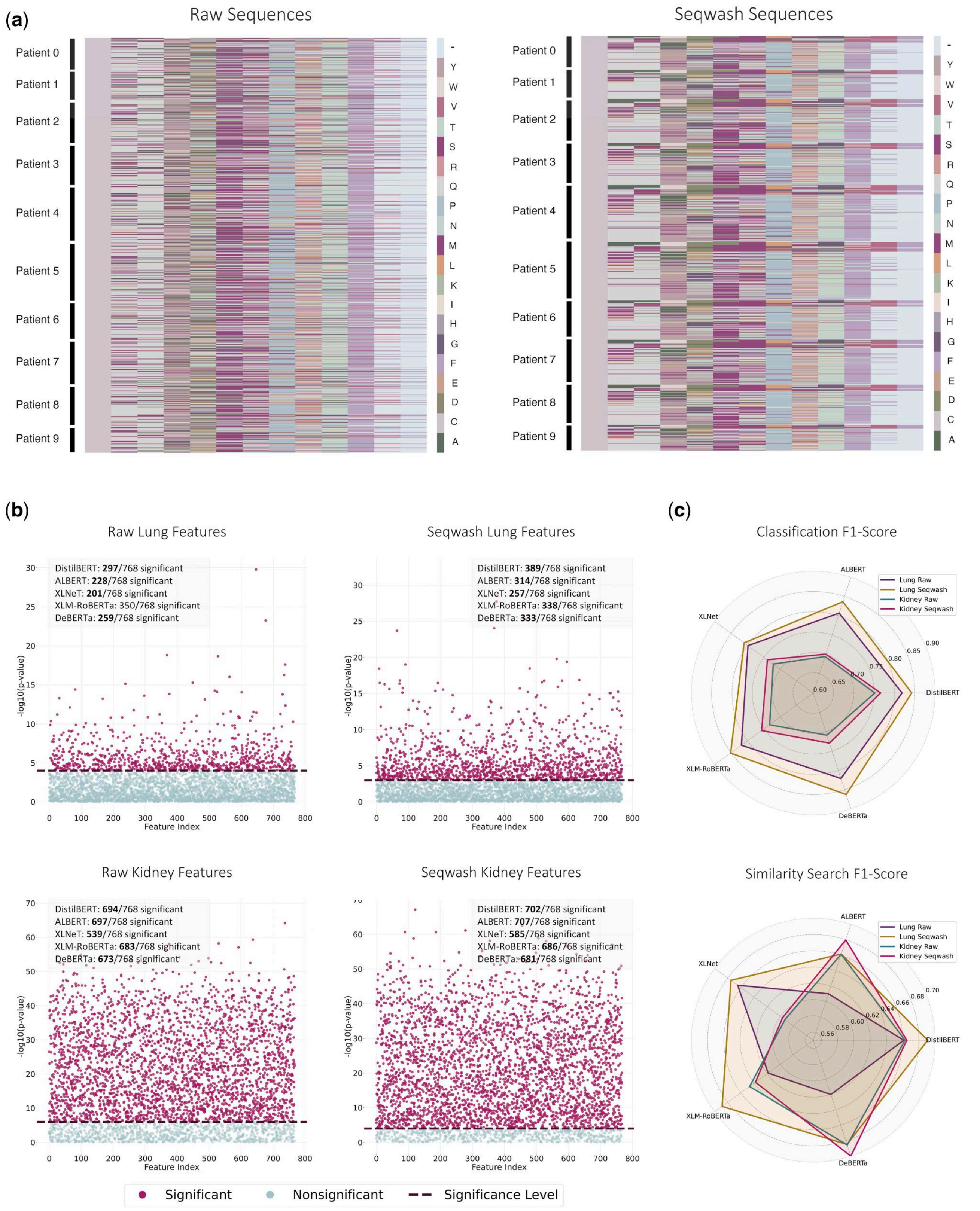

Figure 2 (**a**) Heatmaps of immune repertoires of 10 randomly selected patients before and after applying Seqwash. (**b**) Statistical significance analysis using ANOVA test on deep features extracted from five LLMs, with a significance level indicated by a $P < 0.001$. (**c**) Macro F1-score results for both SVM classification and majority-vote k-NN (k nearest neighbour) search using lung and kidney cancer data from TCGA.

features compared to raw features, indicating the enhancement of feature quality and relevance.

The efficacy of Seqwash features was also evaluated through both supervised classification and unsupervised similarity search applications. As illustrated in the first radar plot in Fig. 2(c), the macro F1-score of classification demonstrates an overall enhanced performance when employing Seqwash features compared to raw features. The second radar plot illustrates the macro F1-score results of applying similarity search. Across almost all the LLMs, an improvement in performance is observed when utilizing Seqwash features.

In summary, Seqwash contributes to overcoming the challenge of processing biological sequences at the patient level by harmonizing them into a unified representation. This enables LLMs to capture informative patterns while eliminating irrelevant information. Evaluation has demonstrated the effectiveness of Seqwash in standardizing profiles, resulting in enhanced feature quality and, consequently, improved performance on both supervised and unsupervised downstream tasks.

Discussion

The results of the validation demonstrate that Seqwash significantly enhances the harmonization and standardization of immune cell sequences at the patient level, a critical aspect for extracting meaningful patterns from complex datasets. By enabling LLMs to capture comprehensive and informative patterns while filtering out irrelevant information, Seqwash improves the quality and relevance of deep features. This results in enhanced performance in both supervised and unsupervised tasks. This advancement has implications for computational biology, particularly in the analysis of immune repertoires. Accurate and efficient processing of sequence data is crucial in unraveling and predicting immune signatures among patients, whether they exhibit similar or diverse immune responses. Moreover, this progress is essential for personalized medicine, where patient-specific insights can lead to the discovery of predictive features, facilitating more accurate classification and decision-making.

To the best of our knowledge, there are no existing harmonization methods in the literature that analyze immune cell sequences at the patient level, which is the gap that Seqwash addresses. Current approaches typically focus on single-sequence analysis, overlooking the holistic signature that can be captured at the patient level, that is analyzing all sequences in a patient's immune repertoire as a unified input. This single-sequence processing approach prevents the effective representation of patients through vectorized feature embeddings, hindering meaningful comparisons among patients. Furthermore, in general, there is a notable lack of approaches that harness AI and machine learning techniques to address challenges in other fields, particularly in medical applications and the analysis of complex data like genomics. This represents a significant obstacle to fully leveraging powerful AI tools that have the potential to drive advancements in these areas.

One common approach in biological sequence representation is one-hot encoding. While widely used [12, 13], this method has several limitations, including high dimensionality, high computational complexity, and significant memory requirements for categorical variables with many categories. Additionally, the sparse representation of vectors dominated by zeros is inefficient for storage and computation, especially in large datasets. Most importantly, one-hot encoding does not consider semantic relationships between categories, treating them as independent entities without acknowledging any underlying connections. Applying one-hot encoding at the patient level is impractical due to the large number of sequences each patient has, leading to exhausted memory requirements. Seqwash addresses these limitations by enabling the representation of patient-level data, thus facilitating more meaningful and efficient comparisons among patients and advancing the application of AI and machine learning in computational biology.

In this article, we effectively employed Seqwash to process immune cell sequencing data derived from TCGA datasets with two primary sites: lung and kidney. We considered immune repertoire in this work since the immune cell profiles have reasonable lengths, and the TCRs/BCRs are shorter sequences compared to other biological sequences such as DNA or RNA. The primary aim of our investigation is to clarify how sequence structure influences the quality of features extracted by LLMs, showcasing Seqwash's efficacy on both raw and processed profiles.

While Seqwash demonstrates promise for extension to diverse biological sequences such as DNA and RNA-seq data, these contexts may pose unique challenges due to their increased variability. In the complex landscape of cancer research, these datasets might contain only sparse relevant changes amidst a vast array of sequences. Hence, proactive preprocessing steps are essential to effectively navigate these challenges before harmonizing the profile structures.

Our future research endeavors are dedicated to broadening the applicability of Seqwash across various biological contexts, including DNA and RNA sequences. We aim to enhance Seqwash by incorporating additional preprocessing stages tailored to the specific characteristics of these datasets. This will involve implementing filtration mechanisms to mitigate noise before harmonizing the profiles.

Author contributions

Areej Alsaafin (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Software [equal]), and Hamid R. Tizhoosh (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Project administration [lead], Resources [lead], Software [equal], Supervision [lead], Validation [equal], Writing—original draft [equal], Writing—review & editing [lead]).

Data availability

Data is public at The Cancer Genome Atlas (TCGA).

References

1. Liu X, Wu J. History, applications, and challenges of immune repertoire research. *Cell Biol Toxicol* 2018;**34**:441–57.
2. Sanchez-Trincado JL, Gomez-Perosanz M, Reche PA. Fundamentals and methods for t-and b-cell epitope prediction. *J Immunol Res* 2017;**2017**:2680160.
3. Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. *Comput Struct Biotechnol J* 2021;**19**:1750–8.
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K *et al*. Large language models in medicine. *Nat Med* 2023;**29**:1930–40.

5. Croce G, Bobisse S, Moreno DL *et al.* Deep learning predictions of TCR-epitope interactions reveal epitope-specific chains in dual alpha t cells. *Nat Commun* 2024;**15**:3211.
6. Sidhom J-W, Oliveira G, Ross-MacDonald P *et al.* Deep learning reveals predictive sequence concepts within immune repertoires to immunotherapy. *Sci Adv* 2022;**8**:5089.
7. Flam-Shepherd D, Zhu K, Aspuru-Guzik A. Language models can learn complex molecular distributions. *Nat Commun* 2022; **13**:3293.
8. Toufiq M, Rinchai D, Bettacchioli E *et al.* Harnessing large language models (LLMs) for candidate gene prioritization and selection. *J Transl Med* 2023;**21**:728.
9. Akiyama M, Sakakibara Y. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR Genom Bioinform* 2022; **4**:lqac012.
10. Hudson D, Fernandes RA, Basham M *et al.* Can we predict T cell specificity with digital biology and machine learning? *Nat Rev Immunol* 2023;**23**:511–21.
11. Song L, Cohen D, Ouyang Z *et al.* Trust4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nat Methods* 2021;**18**:627–30.
12. Zhao M, Xu SX, Yang Y *et al.* GGNPTCR: a generative graph structure neural network for predicting immunogenic peptides for t-cell immune response. *J Chem Inf Model* 2023;**63**:7557–67.
13. Cui F, Zhang Z, Zou Q. Sequence representation approaches for sequence-based protein prediction tasks that use deep learning. *Brief Funct Genomics* 2021;**20**:61–73.

Biology Methods and Protocols, 2024, 00, 1–6
https://doi.org/10.1093/biomethods/bpae055
Methods Article