



The rise of taxon-specific epitope predictors

Felipe Campelo  and Francisco P. Lobo 

Corresponding author. Felipe Campelo, Aston Centre for Artificial Intelligence Research and Application, Aston University, Aston Triangle, B4 7ET, Birmingham, UK. Email: f.campelo@aston.ac.uk

Abstract

Computational predictors of immunogenic peptides, or epitopes, are traditionally built based on data from a broad range of pathogens without consideration for taxonomic information. While this approach may be reasonable if one aims to develop one-size-fits-all models, it may be counterproductive if the proteins for which the model is expected to generalize are known to come from a specific subset of phylogenetically related pathogens. There is mounting evidence that, for these cases, taxon-specific models can outperform generalist ones, even when trained with substantially smaller amounts of data. In this comment, we provide some perspective on the current state of taxon-specific modelling for the prediction of linear B-cell epitopes, and the challenges faced when building and deploying these predictors.

Keywords: epitope prediction; machine learning; data mining; phylogeny-aware modelling

INTRODUCTION

Computational identification of immunogenic peptides, or epitopes, represents an important step in the development of diagnostic tests, vaccines and immunotherapeutic approaches, detecting potential targets for downstream experimental investigation. In the last few years, *in silico* prediction of linear B-cell epitopes (LBCEs) has received considerable attention, with several groups proposing new tools based on a rapidly expanding volume of experimentally-validated data [1].

Several recent works [2–6] describe the building of LBCE predictors based on pre-selecting training data based on taxonomic criteria. Taken together, these works highlight an emerging trend in epitope prediction, namely the development of models optimized for predicting epitopes from specific subsets of organisms rather than from all possible pathogens. Here, we highlight the main hypotheses underlying the development of taxon-specific LBCE predictors, recent published work in this area, and the main challenges for the development of bespoke models for specific taxonomic groups.

TAXON-SPECIFIC EPIPOPE PREDICTION

In biology, a *taxon* refers to a group of one or more species inferred to be phylogenetically related due to shared common ancestry, presenting within-group characteristics that differentiate it from other such groups. A taxon encompasses all included taxa of lower rank, down to individual species [7]. Organisms that are phylogenetically close are expected to be more similar in both their phenotypes and genotypes, a fact that must be taken into account when modelling species-derived data [8].

The main assumption underlying the development of taxon-specific LBCE predictors is that different taxa may exhibit distinct epitope signatures due to, e.g. protein characteristics arising from their evolutionary histories. Under this assumption, it is expected that, once projected onto a feature space, epitopes from pathogens from different taxonomic groups will occupy distinct regions of that space, while phylogenetically close pathogens will occupy closer regions. This pattern can be exploited by predictive pipelines that are either optimised specifically for a given taxonomic group, or incorporate taxonomic information as an additional predictive feature.

Supporting evidence for this assumption was presented in [2], where epitopes from a number of phylogenetically distant pathogens were found to exhibit clearly distinct patterns in terms of location on a feature space, including the superposition of LBCEs from one pathogen with known non-immunogenic peptides from others. This would compromise the performance of models trained without the use of taxonomic information, and motivated the development of organism-specific models, which were shown to significantly improve performance over generalist models [2, 9].

Multiple groups have independently considered the taxon-specific assumption, although not always explicitly framing it as such. Bahai *et al.* [10] presented a virus (superkingdom) specific version of their EpiTopeVec tool, suggesting that '*properties distinguishing epitopic and non-epitopic peptides could be specific to the source of the antigen species*'. Another model tailored for viruses was presented by Yin *et al.* [6], while others presented predictors tailored for distinct taxonomic levels, either by including Class information as an encoded variable as part of the feature

Felipe Campelo is a senior lecturer at Aston University, and a senior member of the IEEE and of the ACM. His research focuses on the application of optimization and data mining to problems in biology and health.

Francisco P. Lobo is an assistant professor at UFMG, where he leads the the Laboratory of Algorithms in Biology. His research focuses on the development of phylogeny-aware methodologies in bioinformatics.

Received: December 15, 2023. **Revised:** February 4, 2024. **Accepted:** February 18, 2024.

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

For commercial re-use, please contact journals.permissions@oup.com

space [4] or by proposing models specific to the family of different viruses [5].

Liu *et al.* [5] also advanced a hypothesis on why the use of family-specific models may be relevant, suggesting that this may be due to shared conserved epitope motifs across closely-related viruses. Although this is a likely contributing factor, performance gains were observed for organism-specific predictors even when extreme care is taken to prevent sequence similarities from playing a role on the estimated performance of the models [2]. This suggests that other factors besides sequence similarity such as, e.g. optimization of model parameters to a target pathogen, may also play a role on the success of taxon-specific models when compared to generalist ones.

CHALLENGES AND OPPORTUNITIES

Although this trend of developing tailored pipelines for specific groups of pathogens represents a promising avenue for improving the performance of epitope predictors, there remains challenges that must be carefully considered by researchers working in this area. These issues relate to several stages in the development of predictive pipelines, from data retrieval to model assessment and deployment.

A primary obstacle to the development of bespoke models, particularly for emerging pathogens, is data availability. Even in the largest curated databases [1], most pathogens have few, if any, labelled epitope data. As an example, only five LBCEs are listed on IEDB for the MPX virus as of November 2023, with no negative examples, a common scenario for emerging zoonotic pathogens that would preclude the training of models using exclusively organism-specific data [2].

This issue has been partially addressed by training bespoke models for higher taxonomic levels such as family [5] or class [4]. Although this approach can be useful, it is limited by the fact that both data volume and the expected homogeneity of traits within these taxonomic levels are often highly variable. A more promising approach, which we outline in a previous report [3], is to have pipelines capable of automatically selecting the optimal taxonomic level to use when building models for a specific pathogen. This approach obviates the need for a pre-defined taxonomic level and enables automatic adaptation to pathogens from data-rich as well as data-scarce groups. Another potential challenge, particularly for under-studied taxa or emerging pathogens, could be the accuracy of the available taxonomic information.

An issue of relevance is the adequate splitting of data for performance assessment and the prevention of leakage of information across splits. Although this is a complex issue and a full discussion would not fit in this comment, we highlight a few critical aspects that appear to be sometimes overlooked. The first and easiest to address is leakage due to homology, i.e. proteins or peptides with high similarity due to common ancestry being placed into different splits of the data. The implicit assumption that examples coming from distinct datasets are necessarily independent (e.g. by training a model on examples from IEDB and validating it on data from other databases without careful verification of homologous entries across datasets) is also another point of potential information leakage which can bias performance estimation, since common peptides can often be found in datasets from different sources. These issues can be addressed through data redundancy reduction (as done, e.g. in EpitopeVec [10] and other approaches) or the incorporation of similarity measures into data splitting strategies. Data splitting at the peptide level, i.e. having peptides from either the same

or homologous proteins placed in distinct splits, may also lead information being accidentally leaked across splits if feature calculation uses protein-level information. Splitting the data based on protein clusters [2, 3] is a simple way to prevent the issue, as this strategy guarantees that peptides from the same protein or from highly-similar proteins are always kept together during, e.g. cross-validation.

Finally, deployment of these taxon-specific models for use by the wider community needs to be considered. While generalist models can be deployed as a single entity, taxon-specific ones require users to be able to train models on demand for their taxon of interest, or to have access to multiple pre-trained models. This presents its own challenges, both in terms of computational resources and of designing appropriate user interfaces or software packages to reduce the set-up burden of building and using taxon-specific models for a potential user base composed of non-coding experts. Designing those systems may require approaches that are not part of the standard data science toolkit, such as user-centered design of high-performance user interfaces. Despite these challenges, taxon-specific epitope predictors present the potential of providing a valuable addition to the existing toolkit for combating infectious diseases.

Key Points

- Computational prediction of linear B-cell epitopes is a critical step in the development of serodiagnostic tests, vaccines, and therapeutic antibodies.
- Increasingly, models are being proposed that incorporate some degree of phylogenetic or taxonomic information about target pathogens.
- Taxon-specific predictors assume that different groups of pathogens exhibit distinct epitope signatures in a feature space, which can be exploited by machine learning methods.
- Challenges remain in terms of the determination of optimal taxonomic levels for model training, methodologically sound performance assessment, and deployment of bespoke models to end users.

REFERENCES

1. Vita R, Mahajan S, Overton JA, *et al.* The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res* 2019;**47**(D1): D339–43.
2. Ashford J, Reis-Cunha J, Lobo I, *et al.* Organism-specific training improves performance of linear B-cell epitope prediction. *Bioinformatics* 2021;**37**(24):4826–34.
3. Campelo F, Reis-Cunha J, Ashford J, *et al.* Phylogeny-aware linear B-cell epitope predictor detects candidate targets for specific immune responses to Monkeypox virus. *bioRxiv preprint*, 2022. <https://doi.org/10.1101/2022.09.08.507179>.
4. da Silva BM, Ascher DB, Pires DEV. epitope1D: accurate taxonomy-aware B-cell linear epitope prediction. *Brief Bioinformatics* 2023;**24**(3).
5. Liu R, Ye-Fan H, Jin D, *et al.* Family-specific training improves linear b cell epitope prediction for emerging viruses. *IEEE/ACM Trans Comput Biol Bioinform* 2023;**20**(6):3669–80.
6. Yin R, Zhu X, Zeng M, *et al.* A framework for predicting variable-length epitopes of human-adapted viruses using machine learning methods. *Brief Bioinform* 2022;**23**(5).

7. International Commission on Zoological Nomenclature. *International Code of Zoological Nomenclature - Glossary*. London: The International Trust for Zoological Nomenclature, 2000.
8. Hongo JA, de Castro GM, Menezes APA, et al. CALANGO: a phylogeny-aware comparative genomics tool for discovering quantitative genotype-phenotype associations across species. *Patterns* 2023;**4**(6):100728.
9. Ashford J, Ekárt A., Campelo F. Estimated limits of organism-specific training for epitope prediction. In: *Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine, Istanbul, Turkey, 2023*. IEEE, Computer Society, Los Alamitos, CA, USA.
10. Bahai A, Asgari E, Mofrad MRK, et al. EpitopeVec: linear epitope prediction using deep protein sequence embeddings. *Bioinformatics* 2021;**37**(23):4517–25.