The dynamic landscape of peptide activity prediction

Bárcenas, Oriol; Pintado-Grima, Carlos; Sidorczuk, Katarzyna; Teufel, Felix; Nielsen, Henrik; Ventura, Salvador; Burdukiewicz, Micha

Published in:
Computational and Structural Biotechnology Journal

DOI:
10.1016/j.csbj.2022.11.043

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Mini review

The dynamic landscape of peptide activity prediction

Oriol Bárcenas a,1, Carlos Pintado-Grima a,1, Katarzyna Sidorczuk b, Felix Teufel c,d, Henrik Nielsen e, Salvador Ventura a,*1, Michał Burdukiewicz a,f,*

a Autonomous University of Barcelona, Institute of Biotechnology and Biomedicine, Spain
b University of Wrocław, Faculty of Biotechnology, Poland
c University of Copenhagen, Copenhagen, Denmark
d Novo Nordisk A/S, Digital Science and Innovation, Denmark
e Technical University of Denmark, Denmark
f Medical University of Białystok, Clinical Research Centre, Poland

A R T I C L E  I N F O

Article history:
Received 19 October 2022
Received in revised form 21 November 2022
Accepted 21 November 2022
Available online 24 November 2022

Keywords:
Peptides
Activity
Prediction
Functional peptides
Machine learning
Deep learning
Reproducibility

A B S T R A C T

Peptides are known to possess a plethora of beneficial properties and activities: antimicrobial, anticancer, anti-inflammatory or the ability to cross the blood–brain barrier are only a few examples of their functional diversity. For this reason, bioinformaticians are constantly developing and upgrading models to predict their activity in silico, generating a steadily increasing number of available tools. Although these efforts have provided fruitful outcomes in the field, the vast and diverse amount of resources for peptide prediction can turn a simple prediction into an overwhelming searching process to find the optimal tool. This minireview aims at providing a systematic and accessible analysis of the complex ecosystem of peptide activity prediction, showcasing the variability of existing models for peptide assessment, their domain specialization and popularity. Moreover, we also assess the reproducibility of such bioinformatics tools and describe tendencies observed in their development. The list of tools is available under https://biogenies.info/peptide-prediction-list/.

C211 2022 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Contents

1. Introduction ................................................................. 6527
2. Methods for data acquisition ........................................... 6527
3. List of existing models ..................................................... 6528
4. Trends and associations in ML-based peptide predictors over time ........................................... 6528
4.1. The vast diversity of functional activities derived from peptide predictors .................................. 6528
4.2. Emergence of deep architectures .................................. 6528
4.3. ML algorithm: what is the most common choice? .......................................................... 6528
4.4. Bioinformatics resources: web servers, code repositories and reproducibility issues .................. 6529
5. Discussion .................................................................. 6530

Funding .......................................................................... 6531
CRediT authorship contribution statement ............................................. 6532
Declaration of Competing Interest ......................................................... 6532
Acknowledgements .................................................................................. 6532
Appendix A. Supplementary data .......................................................... 6532
References ............................................................................. 6532
1. Introduction

In living organisms, peptides are short amino acid sequences that accomplish a wide variety of biological functions [1–4]. Peptides are expressed either as mature products [5] or cleaved from natural precursor proteins. Considering their potential, researchers have been long interested in the identification and functional characterization of peptides with relevant activities.

Several machine learning (ML) models have been developed to this end, providing new and relevant insights into the field [6]. However, considerable discrepancies in theoretical assumptions can arise. For example, peptides are generally defined as short amino acid segments, but their minimum and maximum lengths is a matter of discussion. While some researchers have provided evidence of dipeptide self-assembly into higher-order structures [7,8], certain activities require longer peptide lengths [9]. A similar case applies to maximum lengths, with some studies considering the intended activities, e.g., by searching UniProt for entries not annotated with specific keywords and restricting sequence length.

Different strategies can be used to extract the information from unstructured data, such as peptide primary sequences. The usage of shallow models requires the conversion of unstructured sequential data into the structured, tabular format by employing some heuristic algorithm for feature engineering. Thanks to their architectural flexibility, deep models can learn from structured and unstructured data [31]. Irrespective of the method, predictive models are created to provide useful tools to the scientific community to solve specific problems. In this sense, users are interested in obtaining a wide range of high-quality predictions as quickly and efficiently as possible, providing valuable insight into their data. Furthermore, the method should be robust and stable so that users can accurately re-run computations. Web servers are usually the preferable option to fulfill these requirements, but researchers and private companies with privacy concerns may consider them unsuitable when dealing with sensitive data. Consequently, developers are also encouraged to provide standalone applications in code repositories that can be employed independently by users in local environments. Additionally, standalone tools can be easily integrated with the local analytical pipelines and thus be more appropriate for high-throughput studies.

Considering all the benefits and general appeal of predictive models, an overwhelming number of bioinformatics tools have been developed in the last years to systematically predict and classify peptides. Such tools are based on different features and ML algorithms that aim at identifying specific activities. This seemingly positive fact sometimes turns against users as they may struggle to find the best suited predictor for their precise needs. The problem is exacerbated when functional published tools drown among broken web servers that are no longer available [32] or unreported code repositories [33] that lead to serious reproducibility issues.

Although some overviews of these ML models have been published [34], there is still a need for a systematic review of the current state of the peptide prediction field. In this work, we inspect the characteristics of 140 existing ML tools for peptide activity identification and describe observed associations and tendencies. With this minireview, we expect readers to find an exhaustive and integrated resource for conducting their peptide activity prediction tasks successfully, as well as raising awareness about the reproducibility crisis of many ML models for the scientific community.

2. Methods for data acquisition

Current software tools appearing in the tool list of the online supplementary website were screened using PubMed's API and Google Scholar looking for the keywords antimicrobial, anticancer, antifungal, antiviral, antiparasitic, antitoxin, antiangiogenic, antibiofilm, antihypertensive, antiinflammatory, cell-penetrating peptides, blood–brain barrier peptides, chemotactic, quorum sensing, surface-binding or neuropeptides and peptide prediction in the title or the abstract of the publication until July 1st, 2022. The code provided in repositories was not tested, as it is considered to be the same that was provided during the review process. However, we did test web servers both for the availability of the hosting site and the functionality of the tool on October 14th, 2022. If web server links are no longer working, they are considered non-active. On the other hand, tools that provide an active web server but do not provide results when a model input is given are deemed non-functional. It is important to note that this assessment of functionality of the web server depends very highly on the date and hour of the accession and it could change dynamically.

To measure the popularity of tools, we obtained their citations from CrossRef on October 14th, 2022. Due to their nature, citations are noisier measures of popularity for more recent tools. However, they still at least partially reflect the general interest of the scientific community.
The list of tools was manually curated from original publications to assess their reproducibility into golden, silver, or bronze categories as proposed by Heil et al. in 2021 [35]. We included an additional category (“below bronze”) for tools that did not fulfill the minimum information required for bronze reproducibility: data, models, and source code published and downloadable. Besides these minimum requirements, if the model repository is well documented, key analysis details are suitably recorded and dependencies can be set up in a single command (e.g., via docker, conda, snakemake or renv) the tool is granted a silver category. The gold category is only given for those methods that allow the reproduction of the entire analysis with a single command (e.g., via makefile, snakemake, drake or targets). This workflow is summarized in the Supplementary Fig. 1.

3. List of existing models

The PubMed and Google Scholar screening rendered a total of 140 tools for predicting different peptide functions published from 2009 to July 2022, which are freely available for all users as an online supplementary website at https://biogenies.info/peptide-prediction-list/.

In this peptide prediction list, tools can be easily sorted according to their functional prediction as detailed in the previous section by keyword search. The overall availability of each method can be assessed by the presence of an active and functional web server, code and training repositories and general reproducibility. Therefore, each tool can be searched by name and publication DOI so that users can identify all available bioinformatics resources linked to the tool according to their needs.

To increase the accessibility of the online supplementary website, tools with available web servers or repositories can be easily accessed through the link provided in the corresponding columns. Additionally, all the information can be copied or downloaded as csv, Excel or PDF files.

4. Trends and associations in ML-based peptide predictors over time

4.1. The vast diversity of functional activities derived from peptide predictors

One of the main questions that developers ask themselves when developing a new predictor is what are the key points that differentiate average from successful and well-cited tools, beyond the journal of publication. This is important, as bioinformaticians are constantly aiming to provide the best possible resources that would help researchers in their respective work fields, offering solutions to major concerns. For example, pathologies affecting the nervous system, such as Parkinson’s or Alzheimer’s disease, require dedicated software to screen for peptides with the ability to cross the blood–brain barrier. Other cases include AMPs or anticancer peptides, which have been widely studied for their implication in human health and disease and the development of novel therapeutics in the last decade [36,37]. This is exemplified by the high number of tools predicting these activities (Fig. 1A). Anticancer predictors are found to be the most prevalent with 42 tools, closely followed by the antimicrobial activity with 40, which accounts for 20.5% and 19.5% of all annotated activities (205), respectively. Nonetheless, activities with an available tools do not have the highest median number of citations per year (Fig. 1B). This could be a consequence of the high level of competition between predictors in the most popular activities. In this sense, anticancer and AMPs have a median of around 5.5 citations per year, whereas, for example, toxic, antifungal or antiviral peptides surpass them despite having much less available tools (3, 10 and 20, respectively). Thus, the available activities highlight the variability of the functional landscape of peptide prediction, which is experiencing an evident rise in popularity. It is translated into a sixfold increase of available models developed in the last five years, from 34 in 2017 to 205 in 2022 (Fig. 1C). This year-to-year model increase is also appreciated when considering non-cumulative values of tools deployed per year (Supplementary Fig. 2). The overwhelming amount of peptide tools is becoming a source of confusion and misunderstanding for many researchers, as they struggle to find the optimal resource for their specific needs.

4.2. Emergence of deep architectures

Given the revolution of deep learning (DL) in the proteomics field, it is also relevant to consider the differences between deep and shallow ML algorithms in the peptide activity prediction area. The benefit of the deep frameworks in the field of peptide property predictions is not obvious as the peptide datasets are usually more limited than protein ones, especially considering the number of available sequences [38]. Still, DL models are rapidly gaining popularity, accounting for almost half of the published models in 2021, whereas no deep models were published before 2018 (Fig. 2A). When considering the number of citations of these models, some differences in the relative amount of citations for deep models in comparison with non-deep models can be observed (Fig. 2B). To quantify this observation, we introduce the citation score (C score) and plot it against the year of publication (Fig. 2C). The C score provides an overview of the appeal of deep models, representing the relative number of citations of deep models vs non-deep models. That is, a C score of 1 means that deep models have the same relative amount of citations as non-deep models. As an example, if this value were 1.2, it would indicate that deep models are 20% more cited than shallow models. With this score, it is clear that, except for those tools published in 2021, deep models are more appealing than non-deep models according to their relative number of citations. It highlights the popularity of deep models and provides a plausible explanation for the explosion of deep models created in 2021.

4.3. ML algorithm: what is the most common choice?

A key question when developing an ML algorithm relies on the final classifier that will be selected for the prediction. Among all ML classifiers used by the tools, shallow models represented by support vector machines (SVM) and random forests (RF) are the most frequent, with 64 and 59 independent architectures, respectively (Supplementary Fig. 3). They are closely followed by DL algorithms (51) that have emerged as a common method to build ML models either alone or in a combination of an ensemble. For the sake of completeness, all independent ML classifiers from ensembles were included. Both RF and SVM, like all other shallow ML algorithms, work with structured data. They are also quite popular in the ML field for being easy to train and tune, but they require converting unstructured sequential data to a tabular format using feature engineering. SVMs can perform both linear and non-linear classification and regression but are often difficult to scale to large datasets. In contrast, RF methods are less appropriate for regression but are able to learn how important each feature is to the prediction [39].

Although individual ML algorithms have been described to be highly accurate and precise in their predictions, some tools use the combination of different models to create an ensemble that can potentially improve the prediction capabilities of individual models. This includes deep and non-deep combinations that claim
to be a viable strategy for obtaining more reliable classifications, for example by using features obtained by a DL algorithm to train a shallow model [40].

4.4. Bioinformatics resources: web servers, code repositories and reproducibility issues

Once the purpose of the tool is defined and the corresponding method has been validated, developers need to think of suitable platforms to allow users to run predictions. Web servers are usually the priority option because they provide an easy framework to run calculations without the necessity of technical expertise in programming or command-line management.

Aside from practical usability aspects, the tool should be replicable given the source code and the data are provided. Interestingly, it is only recently that the authors have put more emphasis on the reproducibility of their tools. Indeed, the first tool with at least a bronze level of reproducibility was published in 2018 (Supplementary Fig. 4). In an effort to balance this bias, which might be driven by the recent assessment of reproducibility standards for machine learning in the life sciences [35], here we only analyze the reproducibility status of published tools from 2018 onwards (111 out of the 140 tools). We observed that tools that do not incorporate open web servers (57 out of 111) tend to have a lower number of citations than those that include them (Fig. 3A).

Nonetheless, having an active web server at the time of publication is not enough. Published web servers tend to go offline due to the associated maintenance costs and technical management required. Among all 57 peptide tools with associated web servers, 24 were no longer accessible or functional (42.1%) on October 15th 2022. It means that these tools have lost potential interactions with users that could be interested in running predictions using their algorithms. This is also reflected in the total number of citations, where tools with non-active web servers obtain fewer citations than active predictors.

The most concerning observation, however, is related to the reproducibility (or lack thereof) of ML methods (Fig. 3B). Of the 111 tools, only 38 (34.2%) contained the minimum information...
required to reproduce the ML algorithm independently: source code, data, and models published and downloadable. Among them, nine gold and four silver tools presented the highest possible standard of reproducibility (Table 1). The rest fell into the “below bronze” category where, in the best-case scenario, only web servers were provided to allow users to run predictions. Something remarkable, however, is that web server availability is a much better indicator of the average number of citations per year than the reproducibility standard of these tools (Fig. 3C). This situation may be the root cause of the ML reproducibility crisis, as making a reproducible model is time-consuming and does not render an apparent citation-wise benefit. Given the existing lag between the date of publication and when tools start receiving their first citations, it is also possible that the average number of citations per year for very recent tools is underestimated compared to longer-lived tools.

5. Discussion

Peptide property prediction has positioned itself as a key field of activity within bioinformatics in recent years. The ability of peptides to mediate different types of functional activities and interactions with other biomolecules offers innovative candidates to target specific pathologies. At the molecular level, peptide therapy emerges as a promising strategy to specifically target the pathological agents underlying these diseases [28,41,42]. Early predictors used architectures different from ML, but the difficulty in identifying the contribution weight of each feature to the final decision limited their performance. Thanks to recent advances in ML and bioinformatics, an increasing number of tools based on ML have been released, as it is becoming the standardized method to generate predictive algorithms.

The success of ML methods in predicting peptide properties has led to the development of a vast amount of tools that intend to be useful resources for the scientific community. In this sense, researchers generally prefer the development of user-friendly web servers that are easy to use rather than complex code repositories that only people with some bioinformatics and programming background can run. Based on our analysis, it seems that providing open-source code does not render any benefit citation-wise, and as a consequence, researchers may opt to keep their original code for future research without losing significant impact. We believe this is the origin of the reproducibility crisis we have described in this minireview since only 27% of all tools (38 out of 140) present the minimum information required to reproduce the ML algorithm from scratch.

A similar tendency applies to DL algorithms, which have become the standard “advanced” models for peptide property prediction for some researchers. The ability of these algorithms to learn relevant features directly from the data led many experts in data science and related disciplines to translate this knowledge into the peptide prediction field. Although DL was successfully applied to identify AMPs in metagenomics data [43], it is not clear whether the success DL has seen in other fields of science has yet
been translated to peptide property prediction. Indeed, some studies suggest that DL models so far have failed to improve performance over shallow models for AMP identification [38].

An influx of new tools for developing deep and shallow models significantly reduced the entry barrier for peptide activity prediction. Still, these tools can not replace researchers in assessing the correctness of their models. Here, biological domain knowledge helps in the design of robust benchmarks for a fair evaluation of the performance of proposed models. For example, it was reported that some AMP predictors yield higher probabilities of antimicrobial activity for long peptides which makes little sense from the biological point of view [10]. As the majority of the effort in the field is directed toward proposing new tools and not evaluating existing ones, we are still unaware of the similar shortcomings of available solutions.

Altogether, we argue that there is a need for reorganization and clarification in the peptide property prediction field. The redundancy of specific tools and the difficulty of finding optimal bioinformatics resources keep adding complexity to the standard use of these methods. To at least partially alleviate this issue, we enhance our minireview with an online supplementary web resource that collects all available peptidic tools to scrutinize accessible models and help users find the best for their needs. We aim to raise awareness among researchers and developers about the reproducibility crisis observed in peptide property prediction and encourage them to work together to promote open science and cooperation between researchers.

### Funding

C.P.-G was supported by the Secretariat of Universities and Research of the Catalan Government and the European Social Fund (2021 FI_B 00087). F.T. was funded in part by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606). K.S. was partially supported by the National Science Centre grant (2018/31/N/
Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors want to thank Dominik Rafacz (Warsaw University of Technology) for his help in the online supplementary website development and the anonymous reviewers for their valuable comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.11.043.

References


