Literature Review: Machine Learning for Processing Biological Sequence Data, by Rachel Ratajczak

November 1, 2025

INTRODUCTION

The evolving emergence of DNA, RNA, and protein sequence data has made the need for modern computational methods essential for biological research. In bioinformatics and computational biology, language models and machine learning techniques have become central tools for analyzing and processing biological sequences. Foundational computational biology methods include sequence alignment, statistical analysis, and structure modeling of biological data. These methods struggle when there are large volumes of data, if the data is incredibly diverse, or sparse. By incorporating machine learning, these methods can provide more meaningful information while addressing scalability and complexity issues of sequence data. Incorporation of machine learning supports a wide range of applications in computational biology such as protein classification, molecule interaction predictions, and disease identification. Prior work has explored different methods for processing sequence information, from traditional classification models to neural network architectures and embedding representation techniques. These contributions show how computational models have improved to handle complex biological data and produce more accurate results. The following papers highlight important ideas in this progression, including machine learning for disease prediction, transformer mechanisms for sequence modeling, vector embedding such as ProtVec, and combined approaches that integrate embeddings with deep learning for protein prediction, as well as broader applications of language models in public health.

IMPORTANT IDEAS

As a significant example of applying computational techniques to biological sequence data, Hamelin et al. [1] discusses how machine learning techniques are used to predict viral evolution. This challenge is critical because pathogens, such as SARS-CoV-2, can quickly mutate and evolve. The authors explore the use of machine learning methods on large scale viral genomic data sequences to predict what mutations are likely to occur and become prevalent. Their review has three main methods: traditional phylogenetic models to track lineage divergence over time, statistical models to estimate likelihood of mutations, and deep learning architectures that treat amino acid sequences like biological “text” to determine which variants are likely to emerge [1]. Their key findings show how subfields of machine learning like deep learning and language based models are effective at predicting future mutations before they occur. This is essential for public health sectors to be more proactive in pandemic preparedness. Overall, this work shows how machine learning techniques can transform raw sequence data into meaningful predictions. It showcases that the processing of sequence data extends beyond classification, and can enhance disease response planning.

Modern biological sequence models are inspired by natural language processing (NLP) techniques. In order to understand how transformers dominate the field for sequence modeling, it is necessary to briefly outline the obstacles older NLP models faced. Before the transformer model was created, the recurrent neural networks (RNN) and convolutional neural networks (CNN) were the leading architectures in NLP. RNNs process sequences step-by-step but it struggles with long-range context, while CNNs only filters for local patterns so it struggles with global context. These limitations make it difficult for models to fully understand the semantic meaning of biological sequences.

Vaswani et al. [2] introduces a new architecture called the Transformer model, that uses a self attention mechanism, eliminating recurrence and convolution.The self attention mechanism compares every token in the sequence working in parallel, which enables long-range context understanding. The model extends this concept by using multi-head attention to focus on different areas of the input simultaneously to increase its ability to find complex patterns [2]. The researchers put the transformer model to the test with the WMT 2014 English-German and English-French translation tasks, where the model achieved state-of-the-art performance and was significantly faster to train [2]. Since biological sequence data displays long-range dependencies like human language, the transformer lays the foundation for later sequence data and language models by using parallel processing, pattern extraction, and scalable context handling.

The success of transformer models highlighted the significance of quality sequence representations for efficient analysis. Before these leading NLP architectures were applied to biological sequences, researchers developed vector representation techniques for sequence data. Asgari and Mofrad [3] introduced one of the most influential methods, BioVec (named bio-vectors), which generate embeddings in the sequences much like word embeddings in NLP. Their work focuses on two versions of this named bio-vector, ProtVec for proteins and GeneVec for genomic sequences. ProtVec uses the overlapping k-mers from a protein sequence and treats them as “biological words”, then maps them into n-dimensional vectors to identify structural and functional information. These embedding techniques provide positional embedding to provide meaningful context to subsets of the sequence. This is essential for future applications like the transformer, which processes in parallel, so giving the sequence order enables the model to understand global context. Using the ProtVec embeddings, the authors successfully classified over 324,000 proteins into 7,027 families with an average accuracy of 93% [3] Their embedding method outperforms all other existing family classification methods and serves as a pre-trained mechanism for feature extraction. Their efforts demonstrate that meaningful insights can be found in the raw sequence data rather than hand written. These embedding techniques significantly impacted the bioinformatics field by showcasing that biological “text” could be modeled using NLP techniques. This becomes a crucial step in modern sequence models, enabling more accurate and scalable protein classification and prediction tasks.

Although powerful, the novel embedding techniques such as ProtVec, are still limited in their ability to capture long-range context in sequence data. The researchers in Nambiar et al. [4] build on this concept and introduce the Protein RoBERTa model (PRoBERTa), a transformer-based language model with embedding representations of amino acid sequences designed for protein prediction. PRoBERTa treats protein sequences as a structured language where symbols represent amino acids. The model uses the self attention mechanism which enables the full sequence length to be analyzed to identify complex patterns that traditional methods fail to detect. Unlike these earlier models that relied on manual feature extraction or fixed length embeddings, PRoBERTa learns meaningful context directly from the sequence data. By combining pretrained-embedding techniques with the transformer model, the authors demonstrate significant improvements in protein family classification and interaction predictions. PRoBERTa outperforms all other existing models and showcases the power of treating biological sequence data as structured, learnable text as in natural language processing [4]. These contributions show the significance of combining embedding techniques with transformer models on biological sequence data.

Building on the success of transformer protein language models, researchers apply these techniques to other biological sequences, including RNA. Traditional methods fail to derive accurate patterns in RNA sequence data due to its complex structure. The authors from Penić et al. [5] introduce the RiboNucleic Acid Language Model (RiNALMo), the largest RNA language model to date designed to learn patterns directly from the raw RNA sequence. The model uses an encoder-only transformer, which uses bidirectional embeddings unlike encoder-decoder transformers that map sequences together. The RiNALMo model was trained on roughly 36 million non-coding RNA sequences, and achieved an outstanding performance in prediction tasks [5]. Inspiration from the embedding techniques from raw sequence data eliminates the needs for older methods, and demonstrates the expanding advancements of the transformer models.Their work lays the foundation for scalable RNA analysis by opening new possibilities for RNA therapeutic design, synthetic biology, and public health applications.

CONCLUSION

In order to understand the complexities of life, develop new tools for analysis, and enhance modern medicine, scientists continue to look for ways to better process large amounts of biological data. Machine learning has significantly reshaped how biological sequence data is processed. Earlier approaches in bioinformatics relied on manual feature extraction and statistical models, which limited scalability and the ability to capture hidden biological meaning. The literature review shows the progression towards treating biological data as structured text and using natural language processing architectures to uncover meaningful patterns in sequence data [1]. Vaswani et al.'s introduction to the transformer model showed the power of self-attention for capturing long-range dependencies in text, making the transformer the dominant architecture for language processing [2]. Early embedding techniques such as ProtVec [3] demonstrated that semantic meaning can be discovered from raw sequence data, paving the way for transformer models like ProBERTa [4]. These efforts extend beyond proteins, such as the development of Penić et al.'s RiNALMo model to improve RNA analysis [5].

Together, the work of these researchers highlight the trend that large-scale, sequence models outperform older computational methods and support the advances in disease understanding and biomolecular research. While these machine learning techniques have made a huge impact in the field, there are still some challenges that remain. In particular, there are limitations on the amount of data available, the interpretation of the models logic, and overfitting of complex systems. Nevertheless, the field is continuing to evolve towards more powerful models to reveal the hidden secrets within biological data and expand modern computational biology.

REFERENCES

[1] Hamelin, D., Scicluna, M., Saadie, I., Mostefai, F., Grenier, J., Baron, C., Caron, E. and Hussin J. 2025. Predicting pathogen evolution and immune evasion in the age of artificial intelligence. Computational and Structural Biotechnology Journal 23 (2025), Article 1044. https://doi.org/10.1016/j.csbj.2025.03.044

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. 2017. Attention is all you need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17).* Curran Associates Inc., Red Hook, NY, USA, 6000-6010. https://doi.org/10.48550/arXiv.1706.03762

[3] Asgari, E. and Mofrad, M. R. 2015. Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS ONE 10, 11 (2015), e0141287. https://doi.org/10.1371/journal.pone.0141287

[4] Nambiar, A., Heflin, M., Liu, S., Maslov, S., Hopkins, M., and Ritz, A. 2020. Transforming the language of life: Transformer neural networks for protein prediction tasks. *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '20).* Association for Computing Machinery, New York, NY, USA, Article 5, 1–8. https://doi.org/10.1145/3388440.3412467

[5] Penić, R., Vlašić, T., Huber, R., Wan, Y., and Sikic, M. 2025. RiNALMo: general-purpose RNA language models can generalize well on structure prediction tasks. Nature Communications 16, 1 (2025), Article 60872 https://doi.org/10.1038/s41467-025-60872-5