"Heavy-Light Chain Pair Identification in Antibodies using BERT (work title)"

MASTER THESIS
FACULTY OF SCIENCE, UNIVERSITY OF BERN

HANDED IN BY

Lea Brönnimann

SEPTEMBER 2024

SUPERVISORS:

Prof. Dr. Thomas Lemmin
Chiara Rodella

Abstract

Contents

Glossary

Tokenization Splitting the text into individual tokens, which are the atomic units of information in a chosen language representation. English NLP models typically use words as tokens. Since proteins do not have a well-defined vocabulary of words, word-level tokenization is not a well-defined option in the case of proteins, which is why subword segmentation is often used.

Word Embeddings Mapping of words into vectors with real numbers.

1. Introduction

1.1. Background

1.2. Objectives

1.3. Scope of the Topic

1.4. Research Question

1.5. Nature of the Thesis

1.6. Relevance of the Thesis

2. Sequential Transfer Learning in NLP for Antibody Research

The following section explains the theoretical background necessary for the thesis with regard to the biology of antibodies and the functionality of the NLP models used. A basic understanding of machine learning and protein structures is assumed.

2.1. Introduction to Antibodies

Antibodies are proteins generated in response to foreign pathogens as part of the immune defense mechanism (Graves et al., 2020).

2.1.1. Antibody Structure

Human immunoglobulins are Y-shaped proteins consisting of two identical light chains (LCs) and two identical heavy chains (HCs). In natural systems, one LC pairs with one HC to form a heterodimer that then combines with another identical heterodimer to create the complete immunoglobulin structure. Disulfide bonds connect the HC and LC of the heterodimer, while disulfide bridges link the two HCs of the heterotetramer. Human LCs can belong to either of two functionally similar classes, κ or λ, each containing a constant domain (CL) and a variable domain (VL). On the other hand, human antibody HCs can be classified into five isotypes: IgA, IgD, IgE, IgG, and IgM, each playing a distinct role in the adaptive immune system (for the structure of. IgAs, IgDs, and IgGs consist of three constant (C) and one variable (V) domains, while IgEs and IgMs have one variable and four constant domains. IgA and IgM isotypes possess a J-chain, enabling the formation of dimers and pentamers, respectively, whereas the other isotypes exist as monomers (defined as a HC-LC pair) (Chiu, Goulet, Teplyakov, & Gilliland, 2019).

2.1.2. V domain complementarity-determining regions (CDRs)

The variable domains of antibodies house the binding surface known as the "paratope." This paratope is primarily composed of six distinct variable loopsâĂŤthree on the light chain (L1, L2, and L3) and three on the heavy chain (H1, H2, and H3). Referred to as the

complementarity-determining region (CDR), this region is crucial for enabling antibodies to bind to specific targets with high precision. The size of the CDR allows for numerous unique contacts, contributing to the antibody's exceptional specificity compared to small molecules, which have fewer contact points and are more prone to off-target interactions. The significant variability among the CDR loops is essential for the diverse binding capabilities of antibodies across a wide range of targets (Graves et al., 2020). While five of the six loops typically conform to established canonical structures, the CDR-H3 loop exhibits notable variability in both sequence and structure, making it challenging to characterize using a canonical model. In comparison to other protein loop structures, the CDR-H3 loop is distinguished by its remarkable structural diversity (Melnyk et al., 2023).

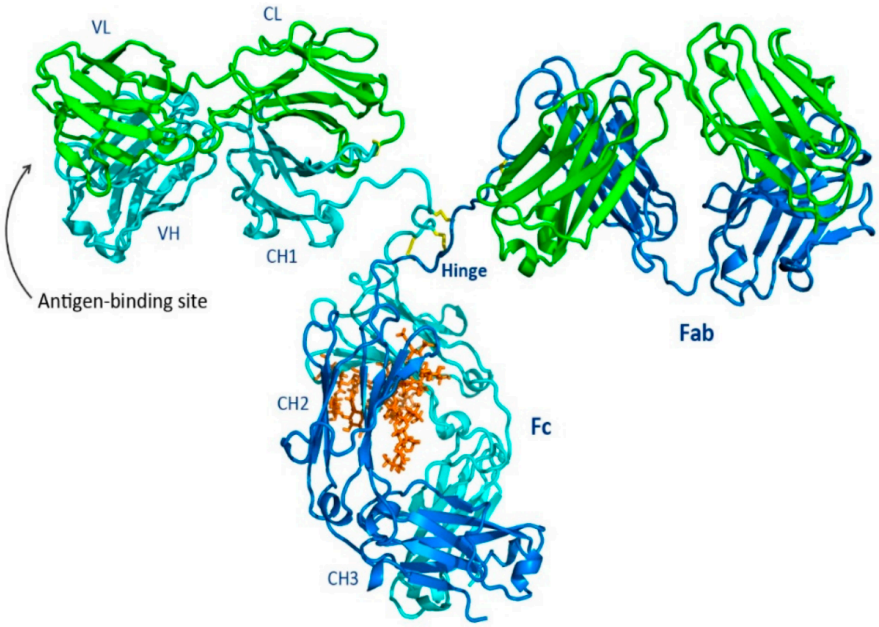

Figure 2.1.: Ribbon representation of a complete IgG molecule, representing a mouse IgG2a isotype. In the image, the light chains are depicted in green, the heavy chains in cyan and blue, the glycan in orange sticks, and the interchain disulfides in yellow (Chiu et al., 2019).

2.1.3. Antibody Engineering and Therapeutic Applications

Antibodies have become indispensable in treating cancer, autoimmune conditions, infectious diseases, and metabolic disorders. Since 1985, the FDA has approved around 100 monoclonal antibodies (mAbs) as therapeutic drugs. The use of antibody proteins

as therapeutics offers a significant advantage over small molecule drugs due to their high specificity, leading to fewer adverse effects. A critical aspect of antibody design is customizing their binding specificity, primarily governed by the CDR (Melnyk et al., 2023).

2.2. Deep Learning Methods for Antibody Research

Deep learning is a branch of machine learning that focuses on algorithms capable of identifying complex patterns in data by transforming low-level inputs (like pixels in an image) into high-level features (such as object shapes). It utilizes artificial neural networks (ANNs) with multiple layers between the input and output, making them "deep". These networks consist of nodes, or neurons, that process inputs and pass the outputs to subsequent layers, gradually extracting more abstract features. In the context of biochemistry, deep learning can start from basic data, like amino acid sequences, and learn to recognize complex biological structures or functions (Graves et al., 2020). NLP models can be effectively used for analyzing amino acid sequences due to the conceptual similarities between proteins and language. Proteins can be represented as strings of 20 amino acid letters, making them a natural fit for many NLP methods. This similarity in representation allows for the application of NLP algorithms to the study of proteins, leveraging the success and promise of NLP methods in other domains (Ofer, Brandes, & Linial, 2021). NLP methods have been successfully applied to protein sequences for tasks such as predicting protein families or properties. Word embedding models in NLP have been used to extract features of protein sequences and have demonstrated successful applications in protein family classification (Xu et al., 2020). This can be explained by the following similarities between natural language and protein sequences: Like natural language, natural proteins generally consist of reused modular elements that exhibit slight variations and can be rearranged and reassembled in a hierarchical fashion. In this analogy, common protein motifs and domains, which are the basic functional building blocks of proteins, are comparable to words, phrases and sentences in human language. Another similarity between proteins and human language is the completeness of the information. A protein is more than just a sequence of amino acids, it is also a three-dimensional machine with a specific structure and associated function, which is largely determined by the amino acid sequence. From an information-theory perspective, this means that the information of the protein is contained in the protein sequence. However, the analogies between proteins and human language only go so far. We can read and understand human language,

but not in the same way as the protein sequence. In addition, most human languages have uniform punctuation and stop words that clearly delineate structures such as words or sentences. There is no clear analogy between the building blocks of languages and those of proteins (Ofer et al., 2021).

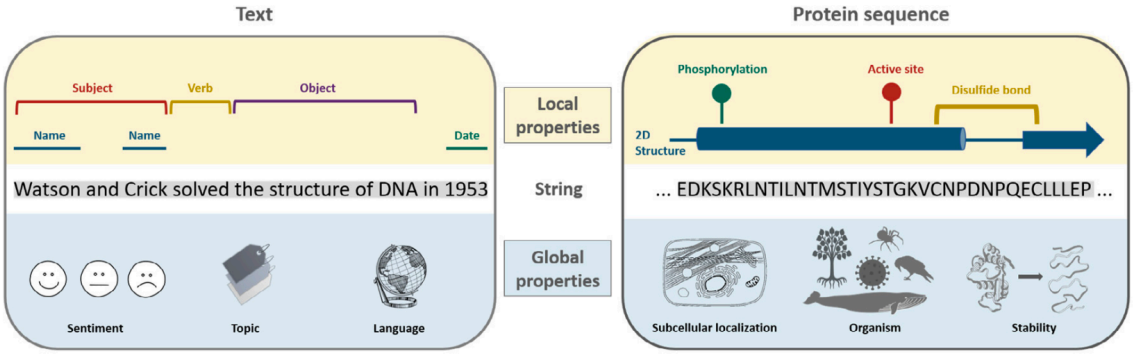

Figure 2.2.: (Ofer et al., 2021)

2.3. BERT and Transformers in Bioinformatics

BERT from Devlin, Chang, Lee, and Toutanova (2019) stands for "Bidirectional Encoder Representations from Transformers" and is based on a bidirectional language model. In bidirectional language modelling, the model considers the entire environmental context of a masked token instead of just the tokens preceding it (Ofer et al., 2021). BERT uses the transformer according to Vaswani et al. (2017) as its architecture. At the time of publication, BERT was able to establish the state of the art in 11 natural language processing tasks (Devlin et al., 2019). The first forms of language modelling in connection with machine learning can be found in Mikolov, Chen, Corrado, and Dean (2013) in the form of the "skip-gram" model. In the skip-gram model, text or unlabelled data is used to train the probability distribution of the next word based on the previous words in the sentence. This process can then be used to calculate static word vectors, which serve as a starting point for other NLP tasks (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). The idea of this unidirectional language model was subsequently used by various other publications and transferred to other architectures such as the Transformer according to Vaswani et al. (2017) (Radford, Narasimhan, Salimans, & Sutskever, 2018). In contrast to Radford et al. (2018), however, BERT uses a bidirectional language model. Figure 2.3 shows the structure of BERT with pretraining and fine-tuning in graphical form. The

language model, or the step known as "pre-training", is trained using two tasks:

Masked language modelling: Since the words of the sentence are processed in parallel in the transformer architecture Vaswani et al. (2017), individual words must be masked in bidirectional prediction. In the case of BERT, these are replaced with the token "[MASK]". The model is then trained to correctly predict these masked words.

Next Sentence Prediction: In order to encode connections between whole sentences in the language model, "Next Sentence Prediction" is used in addition to "Masked Language Modelling". For two sentences A and B, in 50% of the sentences, a sentence B is actually used, which occurs in the text as a directly following sentence after A, and in 50% of the cases a random other sentence is used, which is taken from the corpus. The model must then predict whether these sentences belong together or not.

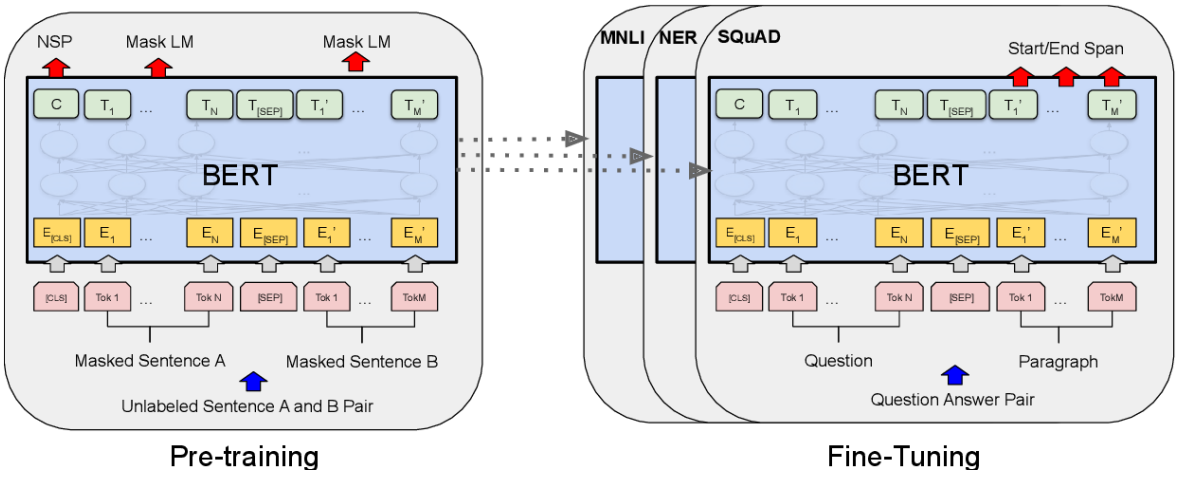

Figure 2.3.: BERT Overview (Devlin et al., 2019).

2.4. Heavy-Light Chain Pair Identification

2.4.1. Gap in the Literature

2.4.2. Conclusion

3. Materials & Methods

4. Results

5. Discussion

6. Conclusion

References

Chiu, M. L., Goulet, D. R., Teplyakov, A., & Gilliland, G. L. (2019). Antibody structure and function: The basis for engineering therapeutics. *Antibodies*, *8*(4). doi: 10.3390/antib8040055

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Graves, J., Byerly, J., Priego, E., Makkapati, N., Parish, S. V., Medellin, B., & Berrondo, M. (2020). A Review of Deep Learning Methods for Antibodies. *Antibodies*, *9*(2), 12. Retrieved from www.mdpi.com/journal/antibodies doi: 10.3390/antib9020012

Melnyk, I., Chenthamarakshan, V., Chen, P. Y., Das, P., Dhurandhar, A., Padhi, I., & Das, D. (2023). Reprogramming Pretrained Language Models for Antibody Sequence Infilling. *Proceedings of Machine Learning Research*, *202*, 24398–24419.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1–12.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 1–9.

Ofer, D., Brandes, N., & Linial, M. (2021, jan). The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, *19*, 1750–1758. doi: 10.1016/J.CSBJ.2021.03.022

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. In *Openai.* Retrieved from https://api.semanticscholar.org/CorpusID:49313245

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is All you Need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper{_}files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Xu, Y., Verma, D., Sheridan, R. P., Liaw, A., Ma, J., Marshall, N. M., ... Johnston, J. M. (2020, jun). Deep Dive into Machine Learning Models for Protein Engineering. *Journal of Chemical Information and Modeling*, *60*(6), 2773–2790. Retrieved from

https://pubs.acs.org/doi/full/10.1021/acs.jcim.0c00073 doi: 10.1021/ACS.JCIM.0C00073/ASSET/IMAGES/LARGE/CI0C00073_0008.JPEG

List of Figures

List of Tables

A. Appendix

A.1. Link to the Code

The link to the entire code of this thesis can be found at:
`https://github.com/ibmm-unibe-ch/OAS_paired_sequences_cls.git`

A.2. Declaration of Independence