AI approaches to unravel B cell evolution

Daphne van Ginneken

A review presented for the
Writing Assignment of the Master
Bioinformatics and Biocomplexity

Utrecht University

UMC Utrecht

Center of Translational Immunology UMC Utrecht
Utrecht University
September 15, 2023

Examiner:	prof. dr. Alexander Yermanos
Second Examiner:	prof. dr. José Borghans

Abstract

The increasing number of sequenced proteins has fueled the development of novel computational methods for their analysis. The emergence of deep learning models, such as Transformers, has led to a breakthrough in the ability to predict features of proteins, such as their function and structure. Protein language models (PLM) can be trained on diverse corpuses of protein sequences to learn generic patterns of protein evolution, or on antibody sequences to improve receptor-specific predictions. While PLMs have shown promise in antibody research, their suitability to help guide affinity maturation or predict somatic hypermutations (SHM) in the context of B cell evolution remains uncertain. This is largely due to the fundamental differences between general protein evolution and antibody generation, selection and evolution. Understanding the antibody-specific process of SHM and clonal selection is important for designing broadly neutralizing antibodies, or to get insights in immune-related diseases. In this review, we describe various phylogenetic and artificial intelligence frameworks for analyzing B cell evolution with their advantages and limitations. We describe how these approaches are able to capture patterns of SHM and selective pressure during B cell evolution, and aim to identify room for improvement.

Layman Summary

The immune system is made up of different cells with different functions. B cells are important to protect against infectious pathogens such as viruses or bacteria. Antibody molecules are created by B cells and play a major role in this protection. Antibodies are either attached to the B cells or float freely in the lymphoid system. These antibodies can recognise a pathogen when their structure fits perfectly to the antigen of the pathogen. The collection of antibodies is extremely diverse due to a process called V(D)J-recombination. During the development of B cells in the bone marrow, three gene segments (V, D and J) are randomly joined together. Each combination of these genes creates an antibody structure to recognise antigens. When a B cell encounters a matching antigen, the B cell will start to replicate and divide. Many errors occur during this replication process, named somatic hypermutation (SHM). This increases the diversity of the antibodies even more. B cells with mutations that improve the binding between the antibody and antigen are then selected to produce more antibodies. These antibodies can neutralize the pathogen, and other immune cells can destroy it. This process results in a strong memory of the antigen, which is stored in memory B cells. When the pathogen infects again, memory B cells with the matching antibody can make more antibodies quickly. Learning how the evolution of B cells work can help by creating antibodies that are able to recognise many different pathogens or could be useful in immune therapies. We could also use this knowledge to predict the course of an infection.

A common approach to learn B cell evolution is to create phylogenetic trees of antibody sequences. This way one can visualize how the different antibodies are related to each other and what possible events of SHM have occurred. These trees are very useful in studying the history of the infection, but are not able to make predictions about the future. As an alternative, people are now looking into artificial intelligence approaches to study B cell evolution. Neural networks such as protein language models can be used to learn patterns of protein evolution and make predictions about structure and function. These models study millions of protein sequences to recognise patterns and try to predict parts of the sequence, its 3D structure, or the function. There are a lot of different architectures of protein language models (PLM) available, all trained on different datasets and applied to specific tasks. Some models are not trained on general protein sequences, but only on antibody sequences. These antibody specific models can be better in making predictions about B cell evolution. This is because the V(D)J recombination and SHM are unique to antibodies, and therefore it is important to incorporate this process in the training of a PLM. Besides training in antibody data, it is also possible to first train the model on general protein sequences and later fine-tune it by adding antibody sequences. This method might be more appropriate because there are more general protein sequences available than antibody sequences, and this way the model could also learn some rules about natural protein evolution.

In this review we analyze a lot of different approaches to study B cell evolution. We look at the various methods to build phylogenetic trees, and the different methods to build a PLM. We also look at the different ways the output of the models can be useful, and how the output could be used to make predictions about B cell evolution.

Contents

Chapter 1

Introduction

In recent years, Transformer-based language models have gained popularity in analyzing structure and function of proteins. Transformers are a type of neural network that is usually applied to natural language or protein sequences. The architecture of Transformers is based on a parallel multi-head attention mechanism to capture long-range sub-sequence dependencies. The success of AlphaFold[1] to predict protein structure has attracted the attention of scientists, resulting in the development of more protein language models (PLM) for other applications. PLMs can be trained on general protein sequences (foundational PLM) and the resulting embedding represents the context and relationship between amino acids in the sequence. For example, various ESM models showed how the embedding of foundational PLMs can be applied beyond structure prediction, such as predicting the effects of a mutation[2]. The promising results of these foundational PLMs inspired antibody research and mutations that increased binding affinity between antibody and antigen were predicted with PLMs[3]. However, the nature of general protein evolution and antibody-specific evolution are intrinsically different. Evolution of proteins is usually focussed on conserving function across generations. As a contrast, evolution of B cell receptors is focused on searching a wide landscape of mutations and is classically assumed to select for variants that increase binding affinity to an antigen. For this reason, foundational PLMs might not be the best approach for B cell evolution and a more receptor-specific model could perform better. In this review, we give a clear description about the methods behind foundational PLMs and provide examples of modifications to create receptor-specific PLMs for B cell evolution. We outline their advantages and limitations, and compare them to traditional phylogenetic methods to study B cell evolution.

During the evolution of B cells, the B cell receptor (BCR) repertoire changes rapidly to be able to recognize new pathogens, but also maintains populations of memory B cells to remain protective against earlier encountered antigens. On a repertoire-wide level, a huge diversity of naive B cell receptors is first created by V(D)J-recombination during early development in the bone marrow. After this, the diversity increases on a lineage-specific level by iterative rounds of somatic hypermutation (SHM) and clonal selection in the germinal center (GC). The sequence diversity can potentially increase binding affinity to an encountered antigen in order to recognize and neutralize its pathogen. But when confronted with a highly virulent pathogen, this natural process can be too slow. Vaccines that induce antibodies of high affinity speed up the viral neutralization and decrease the death rate of infectious diseases dramatically[4]. But antibodies are not the only protein sequences that are rapidly evolving. The antigenic escape of viral proteins results in strains with increased fitness, which makes it harder for the B cell memory to keep recognizing them. Therefore, the antibody induced by a vaccine needs to be broadly neutralizing to be prepared for future strains.

While phylogenetic trees have been able to model SHM and clonal selection, and thereby reconstruct the evolutionary relationships between antibody sequences[5][6][7], it lacks the capacity to predict future evolutionary events. One important feature of PLMs is their competence to learn from patterns in the protein sequences, and use this insight to make predictions. This is predominantly due to their attention mechanisms, which enables the model to focus on the important

parts of the sequence and make long distance connections between residues. Besides predictions, the temporal change of PLM embeddings within a certain B cell lineages can provide a quantitative description of clonal selection during B cell evolution. However, it remains unclear how exactly these models capture signatures of antibody evolution. Especially since many contributing factors such as regularly targeted positions for SHM, paired heavy and light chains and clonal selection mechanisms are not considered in these models. To shed light on this uncertainty, we start with a recap to adaptive immunity and give an overview of antibody structure, development, activation and evolution. Next, we review various phylogenetic methods and PLMs to analyze aspects of B cell evolution. And finally, we give a concrete example of a PLM application for antibody evolution and synthetic antibody design.

Chapter 2

Recap to adaptive immunity

2.1 Antibody structure

Antibodies, also referred to as immunoglobulins (Ig), are proteins produced by the immune system to protect against foreign substances, such as viruses and bacteria. These antibodies can be soluble or membrane bound to B cells (BCR). Each B cell contains approximately 100.000 membrane bound antibodies that are unique per cell[8]. An Ig consists of variable (V), diverse (D) and joining (J) gene segments. The V domain has three variable parts (CDRs), of which CDR3 is the most variable. These CDRs are divided by four stable frameworks (FRs). When folded into 3D structure, Ig have two identical heavy chains and two identical light chains, forming an Y-shaped molecule. Antigens can bind with their epitopes between the heavy and light chains on both sides of the Y molecule and these antigen binding sites are called paratopes.

2.2 B cell development

A high diversity of all the BCRs in the body (BCR repertoire) is necessary to allow recognition of a wide range of possible antigens. This initial diversity is achieved by V(D)J recombination during early B cell development in the bone marrow and fetal liver[9]. Here, hematopoietic stem cells differentiate into immature B cells and selection takes place to prevent binding to self-antigens. The rearrangement of the V, D and J gene segments in the heavy chain, and the V and J segments in the light chain induces mutations. These mutations result in a diverse BCR repertoire with at least 10^{12} variations of antibodies[10]. Recombination activating genes (RAG-1/2) are responsible for the initial DNA double-strand break and receptor rearrangement[11], after which the DNA is repaired with nonhomologous end-joining (NHEJ). After early development, the immature B cells move from the bone marrow to the secondary lymphoid tissues (lymph nodes and spleen) for functional maturation. B cells are considered mature, also referred to as naive, after two transition stages in the spleen.

2.3 B cell activation

A mature B cell becomes activated when it binds to an antigen and usually receives help from an activated helper T cell. An antigen-presenting cell, such as a dendritic cell (DC), can engulf a pathogen and display its antigen on the cell surface with a MHC class II molecule. When a helper T cell binds to this antigen, together with cytokine stimulation, the helper T cell becomes activated and starts to proliferate. Simultaneously, this antigen can bind to a BCR together with co-receptor such as toll-like receptors (TLRs) and complement receptors, which initiates B cell activation. Membrane bound antigens on the B cell are internalized and peptides are displayed on the cell surface, likewise with MHC class II. When a helper T cell with a compatible antigen receptor binds to this BCR complex, together with cytokine stimulation or a coreceptor reaction, the B cell becomes activated. Some antigens are T-independent and do not require a helper T cell to activate a B cell. These antigens are usually repetitive bacterial sequences.

The activated B cell will proliferate and differentiate in short-lived plasma cells, germinal center (GC) B cells, and GC-independent memory B cells. Short-lived plasma cells will secrete the first wave of antibodies to quickly neutralize the pathogen, and then undergo apoptosis. The GC has various functions for adaptive immunity, which we will discuss in the next part. Class switch recombination (CSR) occurs for the most part prior to GC formation[12]. During CSR, the variable region of the Ig stays the same, but the constant region of the heavy chain (C) can switch from IgM and IgD to IgA, IgE, or IgG. Through CSR, B cells can express antibodies with different effector functions, but keep specificity for the encountered antigen. B cells exit the GC as either memory B cells in the secondary lymphoid tissues, or antibody-secreting long-lived plasma cells residing in the bone marrow. During a second encounter with the antigen, memory B cells can be reactivated and will again differentiate into GC B cells or plasma cells.[8]

The ability of B cells to keep memory and be reactivated makes them crucial for long-term protection against pathogens. After each encounter with an antigen, a new generation of B cells with specific receptors are produced. To understand how this repertoire keeps its memory and its ability to respond to new antigens, it is necessary to know the different processes that guide this evolving system.

Chapter 3

B cell evolution

3.1 Germinal center

The humoral immune response is an evolutionary system due to the selective pressure during B cell proliferation. While the antibodies secreted by short-lived plasma cells control the infection early on, a specialized microenvironment is formed in the B follicles of the secondary lymphoid tissues. This microenvironment, the GC, is divided into a dark and light zone. Rapidly dividing B cells undergo somatic hypermutations (SHM) in the dark zone to search the mutational space for affinity improving mutations[13]. The enzyme activation-induced cytidine deaminase (AID) mediates CSR and SHM[14]. AID has a preference of converting cytosines to uracils[15], which triggers an error-prone DNA repair mechanism and introduces additional mutations. This mechanism causes mostly substitutions, but insertions and deletions also happen occasionally[16]. Despite the bias towards cytosine mutation, AID can mutate all nucleotides in the Ig genes. Some motifs are more or less frequently mutated, referred to as hotspots and coldspots (respectively). Hotspots tend to lie in the CDRs and coldspots in the FRs[17]. After this, B cells move to the light zone where affinity selection and clonal expansion takes place. Here, cells with a higher antigen affinity are selected over cells with low affinity towards the antigen. When the SHM increases the strength of the antibody-antigen binding in comparison to the germline progenitor, the cell will receive more survival signals such as ICOS, CD40, IL-4, and IL-12[18][19].

3.2 Repertoire-wide evolution

The number of clonal families in the BCR repertoire, their relative sizes and their protein sequences can provide relevant information about factors such as infection history[20], age[21] and genotype[22] of the host, but also genotypic and phenotypic convergence[23]. Even though the central role of SHM is to increase binding affinity of a certain B cell lineage to foreign antigens, a diverse BCR repertoire is important to protect against rapidly evolving pathogens. There are various hypotheses about the mechanisms behind the maintenance of this diversity, despite clonal selection. A first explanation could be that B cells evolve separately due to the GC "islands" in different lymphoid follicles[24]. This confinement causes the evolution of distinct clones, which together provide a higher diversity in the BCR repertoire. Another hypothesis is that the timing of the GC output induces diversity of B cell clones. Early in the GC reaction, when diversity is still high, more B memory cells are generated. And later, when there has been more selection, plasma cells are generated[25]. A third hypothesis is the possibility of restricted clonal bursting, where not all GC have one dominant clone, but rather support a more balanced clonal diversity[26].

In order to anticipate the course of infections and to improve therapeutic strategies, it is essential to comprehend B cell evolution. This can be done by learning from previous evolutionary events, and applying this knowledge to predict possible immune responses in the future.

Chapter 4

Phylogenetic methods for B cell evolution

Similarities between SHM and evolution by natural selection led to the use of phylogenetic methods to study B cell evolution. High-throughput sequencing of the variable regions of the Ig (Ig-seq) enables quantification of the BCR repertoire. Reconstructing the various clonal lineages over time and studying their temporal relationships can provide valuable information about the mechanisms behind B cell evolution. Each phylogenetic tree from an Ig-seq experiment represents a clonal lineage from an independent V(D)J-recombination event.

4.1 Lineage assignment and clonotyping

In contrast to classic phylogenetic analysis, B cell clonal lineages do not share a common ancestry, but each have their own root. Instead of constructing one tree, antibody sequences must be grouped into clonal families and multiple lineage trees can be inferred. When a particular lineage is of interest, a seeded lineage inference can be performed. Here, the algorithm will search the entire repertoire for clonally related cells containing the "seed" sequence[27][28]. Contrary, unseeded lineage inference will attempt to group all the sequences into their clonal families[28].

Correct lineage assignment is difficult due to the absence of a ground truth, but various methods have been proposed. A common approach is to align the reads to the germline sequence and identify mutations indicative of haplotypes. Another method is to group sequences on their germline genes. These genes should remain identical within a clonal lineage because SHM only targets the variable regions of the antibody. Many studies only consider the V and J segment for clustering because the D segment usually has a low alignment accuracy[6][27]. Often, a threshold is set for the number of sequencing in a tree to distinguish noise from actual lineages. A PCR error for example can appear as a mutation suggestive of a clonotype, but this false lineage will contain only a minor amount of sequences. Additionally, an upper threshold can be set to prevent a high computational demand. A potential rule to assign the correct clones to lineages is to use the substitution-prefered nature of AID to only assign clones to a lineage when they have equal CDR3 sequencing lengths. There are various tools available for seeded and unseeded lineage assignment, each having their own advantages and limitations regarding germline sequences, computational demands and possible parameters. Partis is a tool for clonal lineage inference based on a hidden Markov model (HMM) likelihood method. In this model, the hidden states are the bases of the VDJ germline sequence and the emissions are the possible substitutions[27]. SONAR is a tool for alignment-based lineage inference, and is capable of subsequent phylogenetic tree reconstruction[6]. Their tree inference is a wrapper around DNAML[29], also a HMM maximum likelihood method. The tool Clonify determines clonal lineages based on hierarchical clustering of the edit distance, and is specialized in unseeded inference[30].

During antibody clonotyping many potential complications need to be considered, but diverse

methods circumvent these problems by taking into account antibody-specific evolutionary features. However, most of these current tools and methods are directed towards bulk Ig-seq data and are not yet available for single cell data.

4.2 Substitution models

A substitution model is necessary to estimate the relationship and divergence time between sequences. There are different substitution models with their own assumptions and parameters. Most of these classical substitution models assume that mutations are site-independent, but this is not the case for B cell evolution. The enzymatic nature of AID is neighbor dependent, resulting in hot/cold spots. The absence of models dealing with this phenomenon creates a gap in the phylogenetic analysis of B cell evolution. An important step towards closing this gap is the HLP17 codon substitution model[31]. Even though this model takes into account hotspots, it has computational limitations as not all motifs cannot be accounted for simultaneously. This same group proposed a hierarchical framework for repertoire-wide B cell evolution[5]. They assume that lineages within a repertoire have similar substitution patterns, and therefore constrain all parameters to be identical between lineages. This approach can quantify SHM feature diversity through time within and among individuals. They introduce the HLP19 substitution model which has significantly less parameters than HLP17 and improved estimations. Yaari *et al.*[32] improved models of SHM targeting and substitution to be selection independent by only basing them on synonymous mutations. They do this to study SHM mechanisms without selection bias. They use a large high throughput Ig-seq dataset of human blood and lymph node heavy chain sequences to model the dependency of the four surrounding bases of a substitution (S5F). In later work[33], similar models were created based on human and mice light chains. They confirm the neighborhood-dependent nature of AID, and discover additional hot/cold spots.

These alternative substitution models are another discrepancy between classical evolution analysis and B cell specific evolution. Despite the efforts to create substitution models more directed toward antibody evolution, the full complexity of SHM targeting can never be captured with these motif-based models as targeting could be influenced by bases further up/down-stream. However, the 5-mer model (S5F) seems to be a suitable trade-off between motif length and computational demand.

4.3 Tree inference

There are different methods to go from an Ig-seq multiple sequence alignment (MSA) to a constructed tree. The distance-based method, based on pairwise similarities and neighbor joining, is the most simple and fast one and does not require a substitution model. The Levenshtein distance is commonly used as a distance metric and represents the minimum number of point mutations necessary to change one sequence into another, this eliminates the need for a MSA. Another tree inference method is maximum parsimony, which searches for the tree that explains evolutions with the least amount of mutations. Various tools such as GCtree[34], Phylip[35], IgTree[7] are based on maximum parsimony. In comparison to distance-based and parsimony methods, likelihood-based and Bayesian methods are dependent on a substitution model. The parametric method of maximum likelihood allows for different substitution rates between the V(D)J segments, but is not able to incorporate intermediate nodes and polytomie[5]. These polytomies and intermediate sequences serving as internal nodes are another aspect where antibody trees are distinct from traditional trees. Tools such as IgTree[7] and ImmuniTree incorporate these antibody-specific features. The Bayesian inference method can incorporate prior biological knowledge, such as mutation-, duplication-, and death rates. A widely used tool for this is BEAST[36]. This method has a huge computational demand, mostly due to the use of the Markov chain Monte Carlo (MCMC) algorithm, but can incorporate calendar time and the comparison of multiple trees.

Aside from Bayesian inference, these methods return unrooted trees. Most biologically meaningful would be to use the germline V(D)J sequence as an outgroup, as this is the starting point

of affinity maturation. However, the exact genomic composition of the germline sequence is not always known. There is a high diversity in germline sequence between humans, which makes it difficult to identify whether there are mutations. This diversity could also alter the inferred mutation rate. BEAST infers the root sequence without germline information the same way it infers sequences at internal nodes. Another difference between traditional phylogenetic trees and those tailored for B cell development is the frequency component. Due to expansion, multiple B cells will have identical sequences. This clonal frequency can be displayed in the trees by node size. This is important information as the precursor frequency directly impacts the GC reaction[37]. Including this clonal abundance information in the tree inference process increases accuracy in simulated trees, this feature is included in GCtrees[34].

4.4 Quantification of B cell evolution

To draw relevant conclusions about B cell evolution from antibody trees, it is important to correctly interpret each individual lineage tree and the collection of all trees in the repertoire. Qualitatively, an imbalanced tree repertoire can be interpreted as a single clone outcompeting other clones. Conversely, a balanced tree repertoire implies that selection occurs evenly (Figure 1). Quantitatively, a metric such as the Colless index can be used that accounts for the number of descendants at each node. Similarly, the Sackin index represents the variance in leaf depths. Both metrics generate a number per tree, which can be compared between lineages[28]. Since phylogenetic trees are essentially open networks, graph theory metrics could also be applied. Examples of metrics to analyze the structural properties of such graphs are betweenness centrality and clustering coefficients[38]. If a node has a high betweenness centrality, it means that the shortest path between other nodes frequently passes through this high betweenness node. This node can be considered a potential bottleneck and might represent an essential protein in the evolutionary pathway. The clustering coefficients metric quantifies local connectivity in the network, and represents highly connected proteins.

Phylogenetic approaches are able to retrospectively reconstruct B cell evolution of clonotypes within a repertoire. But this approach does not learn from these models in order to make predictions about future evolutionary events. Additionally, it remains a challenge to study how the repertoire-wide "forest" of antibody trees evolves as a system.

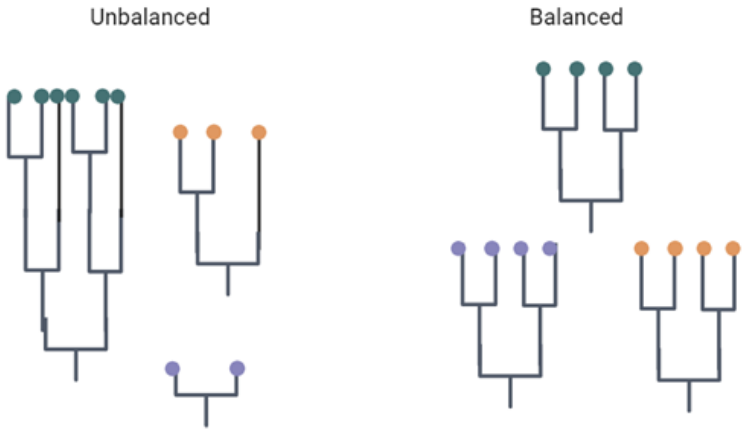

Figure 1: A schematic example of an unbalanced and balanced tree repertoire. With a balanced selection, all trees in the repertoire have approximately the same number of descendants.

Chapter 5

Protein language models for immune receptors

Protein language models (PLMs) can learn patterns of evolution, both general and antibody specific. This ability could be applied to learn representations of B cell receptors and potentially predict evolutionary trajectories. Over the last years, AI powered language models such as ChatGPT showed great progress in interpreting and generating text. Words in a text and amino acids in protein sequences are both context dependent. Words in large bodies of text can refer to each other. Similarly, amino acids can bind and interact over long sequence distances due to their dictated 3D structure. This led to the idea to apply language models to protein sequences.

5.1 Language models

Large language models are a type of neural network in the field of artificial intelligence. Neural networks are machine learning models composed of multiple layers to perform non-linear transformations on their input. Signals move from the input layer to the output layer through all hidden layers, and certain layers could be traversed multiple times (Figure 2). For each neural network there will be fixed hyperparameters and learned parameters. Hyperparameters are set before training and can for example be the number of layers and nodes, the learning rate, which activation function transforms the data, and a loss function. Learned parameters could be token embeddings, positional encoding, and architecture specific parameters.

Current successful language models use a Transformer[39] deep learning architecture. Transformers use parallel multi-head attention mechanisms to prioritize certain tokens over less important tokens (Figure 3). A Transformer needs a representation of the input sequences (input embedding). Here, the input sequence is split into tokens and a positional encoding is added to describe their position in the sequence. In the example of ChatGPT, full-length sentences can be tokenized into individual words. Multi-head self attention layers weigh the local and distant importance of each element to capture long range dependencies. After each self-attention layer comes a feed-forward layer and a normalization to enable parallelism[39]. The output of a language model that uses a Transformer is dependent on the task it was designed for. For sequence-to-sequence tasks, such as translations, the output is a transformed sequence. These sequence-to-sequence models usually work autoregressive, meaning it predicts one word at a time while considering the previously generated words. GPT-3[40] from the OpenAI GPT series is such an autoregressive language model that has approximately 175 billion parameters determining the model's behavior. Another example of a language model that uses a Transformer is BERT (Bidirectional Encoder Representations from Transformers)[41], which is designed for masked language modeling (MLM). BERT learns contextual relationships of tokens in both directions by randomly masking some tokens in a sequence during pre-training. The model predicts the original token that was masked, and the output is a probability distribution of the entire vocabulary for each hidden token. Some other examples of tasks for Transformer are classification, regression or summarization.

Transformer models are usually preferred over recurrent neural networks (RNN), because RNNs are not able to parallelise and struggle with long-range dependencies[42][43]. RNN processes input data one token at a time and uses a hidden state to provide information from the previous step to the next step (Figure 3). The information captured in the hidden state vanishes with each step, this makes it challenging for the model to capture relationships between tokens over a long distance. Common RNNs include the long-short term memory model (LSTM) and gated recurrent units (GRUs) because they mitigate some of RNNs limitations. Both LSTM and GRUs use gating to regulate the flow of information through the cell states and hidden states, and they update the information that should be transferred between steps.[44][45][46]

PLMs are usually Transformer models trained on protein sequences. These sequences can be extracted from databases such as UniProt[47]. On UniProt, sequences can be selected based on clusters of identity to control redundant sequences in the training data in order to reduce computational burden. Foundational PLMs are trained on all protein sequences and receptor-specific PLMs are usually trained on antibody sequences or T cell receptors. Most of these antibody sequences are derived from the Observed Antibody Space (OAS)[48][49][50][51][52]. The OAS database contains over a billion paired and unpaired human and mouse antibody sequences. An overview of some commonly used foundational and receptor-specific PLMs can be found in Table 1. Sequences in PLMs can be tokenized at amino acid level, but more complex tokenization schemes that include biochemical properties or structural information are also possible. The output embedding of a PLM gives a contextualized representation of the input sequences. These embeddings, but also the scores of the intermediate attention heads, can be used for various downstream analyses such as predictions about structure, function, interactions, homology, and evolutionary trajectory of proteins.

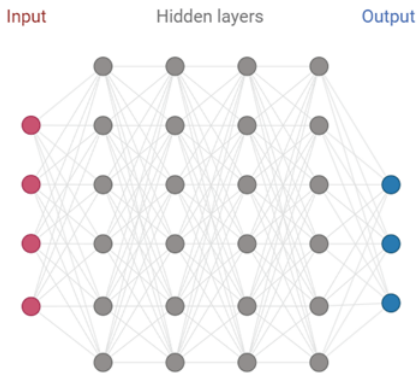

Figure 2: A classic artificial neural network.

5.2 Protein structure prediction

A central challenge in protein biology is understanding the relationship between sequence, structure and function. The success of AlphaFold2[1] protein structure prediction indicates how neural networks can be useful in this sequence-structure-function problem. AlphaFold identifies conserved regions by creating a multiple sequence alignment (MSA) of the input sequence with sequences from various databases. Next, the model uses a Transformer pre-trained with a MLM to capture long range dependencies. After AlphaFold, more language models for protein structure prediction were created. ESMFold[54] showed high similarity to AlphaFold2, but was an order of magnitude faster. Methods specific for predicting antibody structure were also proposed, such as IgFold[55] and DeepAb[56] and demonstrated comparable accuracy with AlphaFold.

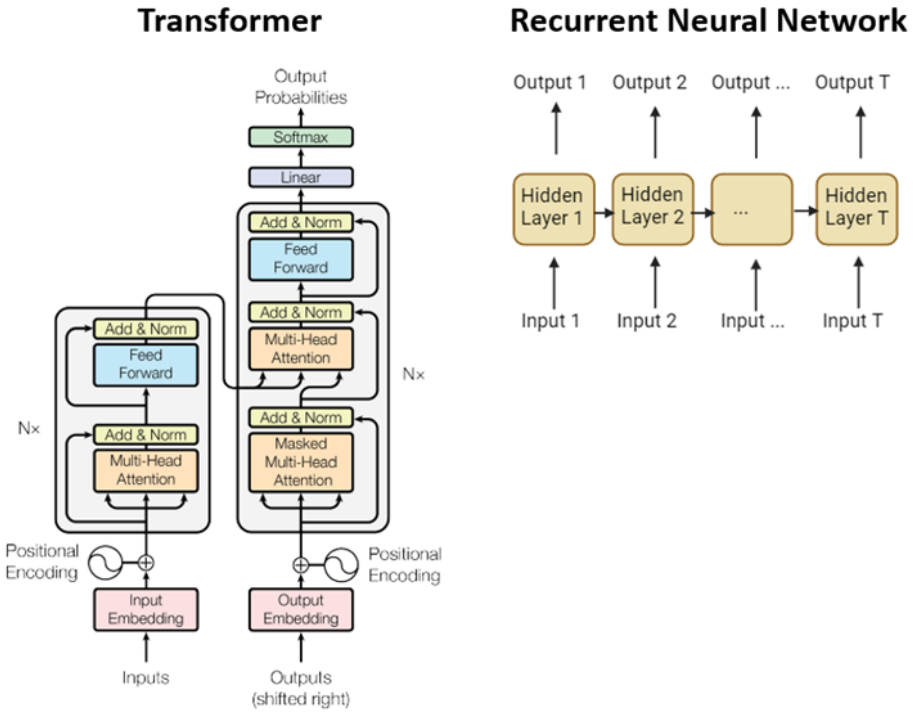

Figure 3: Algorithm example of a Transformer model (derived from Vaswani et al.[39]) and a Recurrent Neural Network (RNN). The transformer architecture repeats for each layer (N). The RNN repeats for each input token (T).

5.3 Protein functional prediction

Foundational PLMs showed great promise in observing patterns related to structure and function in general protein sequences. The output embedding of the Transformer ESM-1b[43] represents biochemical properties, remote homology, and protein family alignment. Even though ProtBERT[53] is trained on a substantially larger dataset, ESM-1b outperforms ProtBERT on residue prediction accuracy and performs with similar accuracy on predicting cellular location. Because some of these learned patterns from general protein sequences might differ from those of BCR sequences, receptor-specific PLMs such AntiBERTa[49], AbLang[51] and Sapiens[52] were proposed. While AntiBERTa is a single model trained on unpaired heavy and light chains sequences, AbLang and Sapiens are trained on the heavy and light chain individually. AbLang outperformed ESM-1b in residue prediction of antibody sequences. The final embedding of AntiBERTa is a combination of the self-attention scores of the 12 attention heads in the final layer. High self-attention scores tend to be between residues in the CDR3 region, reflecting how AntiBERTa pays attention to what's most variable between sequences. The AntiBERTa embedding showed a better representation of mutational load and B cell maturation state than ProtBERT and Sapiens, based on 1000 random BCR heavy chains[49]. Details about mentioned PLMs can be found in Table 1.

These comparisons show that the model's performance depends on the task at hand, and the model's architecture and training method. B cell specific evolution is significantly different from protein evolution. V(D)J recombination provides incredibly diverse naive B cell repertoires in a human body, which can further undergo somatic hypermutation to an even higher number of potential antibodies. Where antibody evolution is focused on gaining new functions, protein evolution is primed to conserving function. For this reason, it is hypothesized that receptor-specific PLMs might be better at learning and predicting B cell evolution than foundational PLMs.

This hypothesis was tested by Hie *et al.*[3] by guiding affinity maturation of various antibodies with the foundational ESM-1b[43] and ESM-1v[2] models. ESM-1v is an ensemble of 5 Trans-

Foundational PLMs	Model	Layers	Parameters	Training data
	ESM-1b[43]	33	650M	250M UniRef50
	ProtBERT[53]	30	420M	2122M BFD100 and 216M UniRef100
Receptor-specific PLMs				
	AntiBERTa[49]	12	86M	15M light chain and 42M heavy chain OAS
	AntiBERTy[50]	8	26M	558M OAS (heavy and light chain)
	AbLang[51]	12	110M	187,068 light chain and 14M heavy chain OAS
	Sapiens[52]	4	568,857	19M light chain and 20M heavy chain OAS

Table 1: Some foundational and receptor-specific Protein Language Models (PLMs). Training data is the number of sequences. M = million; BFD = Big Fantastic Database1; UniRef = UniProt Reference Cluster[47]; OAS = Observed Antibody Space[48]

former models based on transfer learning. With transfer learning, information learned from one task is used to solve another. In this work[3], they concluded that a foundational PLM without task-specific training data can better guide antibody evolution towards a higher affinity than the receptor-specific PLMs AbLang[51] and Sapiens[52]. They conclude that evolutionary likelihood is highly associated with higher fitness, because a foundational PLM can guide affinity maturation. But as mentioned earlier, the receptor-specific PLM AntiBERTa[49] provides a contextual understanding of BCR sequences. Despite only being trained on BCR sequences, the model is able to identify naive and memory B cells. The separation between these subsets was less obvious when processing the same set of BCR sequences with the foundational PLMs ProtBert and Sapiens. Besides B cell maturation state and and V gene family, the AntiBERTa embedding was also able to find a pattern in mutational load. This pattern was highly associated with maturation state, as memory B cells showed to have an overall higher number of mutations than naive B cells. This result is in line with the hypothesis that memory B cells have undergone more SHM than naive B cells. These findings indicate that AntiBERTa is capable of learning B cell evolution, as more mutations indicate a cell to be further along the evolutionary trajectory. So it appears that a receptor-specific model might be better in capturing information related to function[49], but a foundational PLM might be better in predicting mutations that increase fitness[3].

5.4 Fine-tuning

Language models are usually pre-trained on more general datasets in order to learn contextual patterns. After this, a pre-trained model can be fine-tuned for a specific task. During fine-tuning, some parameters of the pre-trained model are updated by further training on labeled data using a task-specific loss function. For example, AntiBERTa was fine-tuned for paratope prediction by adding a binary classifier head on top of the 12th layer[49]. Known paratope residues from the structural antibody database (SAbDab)5[57] were used for this. During fine-tuning, the self-attention scores changed and AntiBERTa outperformed other paratope prediction methods such as ProtBERT and Sapiens[49]. Fine-tuning of ESM-1b was performed to serve as a variant effect predictor, its accuracy was comparable with a state-of-the-art method[43].

Most receptor-specific PLMs use a separate model for the heavy and light chains, or use a single model with unpaired antibody sequences. Determining binding affinity and epitope position is dependent on the combination of the heavy and light chain. Therefore, receptor-specific PLMs

could be improved by training on paired data. Burbach *et al.*[58] recently tested this hypothesis by creating two BERT-based models, one with native paired and one with unpaired sequences. They concluded that cross-chain learning significantly improved the model's performance on predicting masked tokens. However, the high cost of generating a natively paired antibody dataset is a huge limitation and the shortage of data poses a drawback on this model. As an alternative, they also fine-tuned the ESM-2[54] model with the same dataset of paired antibody sequences. They observe that the cross-chain attention is now focussed on immunologically relevant regions, despite ESM-2 being a foundational PLM. This result supports the idea of using a PLM trained on a large dataset to capture general patterns of protein evolution, and fine-tuning it on a smaller dataset for task specific purposes.

Chapter 6

Protein language models for protein evolution

In nature, B cell evolution searches across a huge space of possible sequences by SHMs and V(D)J-recombination for sequences that improve fitness. Reproducing this approach for artificial evolution is experimentally unfeasible. However, PLMs are able to learn patterns of general evolution when trained on a variety of naturally occurring proteins.

6.1 Evo-velocity

Besides single residue mutations (local evolution), PLMs can be applied to study long evolutionary trajectories (global evolution). The term "evo-velocity" was introduced by Hie *et al.*[59] and refers to the dynamic trajectory of protein evolution. Insight in global evolution is gained by making predictions of local evolution. Here, the foundational PLM ESM-1b[43] is used to create an embedding and predict pseudolikelihood. First, they construct a sequence similarity graph based on Euclidean distance between the language model embeddings and connect the nodes to its k nearest neighbor. The pseudolikelihood for each sequence is predicted by the language model using a MLM and calculating conditional likelihoods for certain residues. The evo-velocity score, which is the pairwise difference in pseudolikelihood, is calculated for each edge in the graph and provides directionality. A score above zero moving from sequence A to sequence B indicates that sequence B evolves from sequence A, and vice versa. The vector field resulting from the graph can be used to predict the roots and the pseudotime order, which are hypothesized to represent evolutionary order. Additionally, mutations can be identified that correlate with the direction of evo-velocity. This evo-velocity was able to accurately reconstruct evolution of influenza A nucleoprotein and hemagglutinin, SARS-Cov-2 glycoprotein, eukaryotic globin protein family and cytochrome c, and the ancient evolution of serpin and glycolytic enzymes[59]. An important limitation of evo-velocity is the possible absence of intermediate sequences. The model relies on observed sequences in public databases with no ground truth about the actual evolutionary order. If there are generations between sequence A and sequence B that are not in the dataset, this approach might not be able to estimate the correct evolutionary order. Another limitation to consider is that this method does not consider insertions and deletions when predicting mutational effects.

6.2 Predicting protein evolution

Evo-velocity has some advantages over traditional phylogenetic methods. Evo-velocity can analyze larger amounts of sequences and include multiple roots. This landscape can model convergent evolution instead of just divergent evolution. Additionally, evo-velocity can provide a notion of uncertainty. Most importantly, where phylogenetic methods model retrospective evolution, evo-velocity is able to predict future evolution. This ability could have major implications in predicting

the progression of a viral outbreak and in designing broadly neutralizing antibodies. As an example, Hie *et al.* used a bidirectional LSTM language model and combined the ability of a sequence to keep infectivity but also increase fitness to predict escaping mutations[60]. In another case, Han *et al.* presented MLAEP, a multitask deep learning neural network to predict mutations in the spike receptor-binding domain (RBD) that could have high antigenic evolutionary potential[61]. They fine-tuned the ESM-1b model to predict the binding affinity of RBD variants towards ACE2, a crucial receptor for SARS-CoV-2 to enter cells.

Chapter 7

Protein language models for B cell evolution

B cell evolution can be learned retrospectively with phylogenetic methods, but is unable to provide a prospective outlook. PLMs showed capability to learn general rules of structure and function from sequences in order to pick up information about 3D structure, binding affinity, maturation state and viral escape, which could all be indicative of B cell evolution. This leads us to the main question of this review: Can PLMs predict B cell evolution? And if this is the case: Can these learned signatures of B cell evolution guide synthetic antibody evolution?

7.1 Pre-training mechanisms

Receptor-specific PLMs don't always outperform foundational PLMs in antibody specific tasks, but the models can be fine-tuned to improve performance. The AnTibody Understanding Evaluation (ATUE) benchmark further investigates the possible benefits of antibody-specific pre-training and fine-tuning by evaluating different models on supervised tasks with a range of specificity levels for antibody evolution[62]. They compare the pre-trained foundational PLM ESM-1b, the pre-trained receptor-specific PLM AntiBERTa, and another pre-trained receptor-specific PLM named EATLM where they included additional biological mechanisms. The biological mechanisms included during EATLM pre-training are ancestor germline prediction and mutation position prediction. These objectives could guide the model to take into account both global and local evolutionary patterns. This benchmarking states that foundational PLMs and receptor-specific PLMs perform similarly on antibody binding and paratope prediction. However, EATLM outperforms these models slightly. To challenge the prediction of B cell evolution, ATUE classifies B cell mature states during evolution. For this task, receptor-specific PLMs significantly outperform foundational PLMs. This is in agreement with the results from AntiBERTa[49]. EATLM performs even better than the receptor-specific PLMs in function prediction. This indicates that modeling the biological mechanisms of antibody evolution during pre-training could improve the prediction of B cell evolution.

7.2 Proof-of-principle

The receptor-specific PLM AntiBERTy[50] was created to understand the composition of an individual's immune-repertoire, see Table 1 for details. This model was pre-trained on the OAS database, and later applied to calculate embeddings of the immune repertoire of four donors that developed neutralizing antibodies (VRC01) against HIV-1 gp120. Evo-velocity scores based on the AntiBERTy embedding revealed evolutionary trajectories within the immune repertoire. The observed trajectories between germline sequences and highly mutated sequences within each repertoire showed consistency with affinity maturation[50].

7.3 Synthetic evolution

PLMs attempt to learn B cell evolution by predicting mutations that increase antigen binding affinity, predicting B cell maturation state, and by predicting the evolutionary trajectory. These predictions could potentially guide synthetic antibody production. This artificial evolution process could be faster than natural evolution and therefore be clinically relevant for various diseases. Hie *et al.*[3] showed that foundational PLMs were applicable in protein engineering for therapeutic purposes by producing antibodies with an increased affinity for SARS-Cov-2, Ebola and Influenza. Their synthetic antibodies had increased affinity and viral neutralization activity, but no significant changes in poly-specific binding or immunogenicity. Even though the improved affinity is lower than usually observed in *in vivo* evolutionary trajectories, the search space of possible mutations was significantly narrowed down in comparison to natural SHM. They used foundational PLMs with the hypothesis that high evolutionary plausibility would be translated into higher fitness. Selecting an antibody with high affinity towards an antigen is one of the most important parts of designing a broadly neutralizing vaccine. When the virus mutates frequently, the antibody must target an antigen that is conserved over multiple strains to prevent mutational escape. And by targeting the germline of this antibody, the maturation process can be initiated from an early stage.

Chapter 8

Discussion and conclusion

With its high diversity and rapid evolution, the B cell immune response is a unique evolutionary process. The evolving dynamics of individual B cell lineages and the entire BCR repertoire are hard to reconstruct and even harder to predict. Phylogenetic approaches with various models of SHM, V(D)J-recombination and selective pressure have been able to infer evolutionary relationships between BCR sequences. But these phylogenetic trees lack the ability to learn patterns from input sequences and apply this knowledge for prediction tasks. The introduction of Transformer models has brought protein language models (PLM) to a new level. The contextual embeddings and attention mechanisms of these models are now able to predict structure, function, binding, and evolution of proteins. This is promising for its application to antibody sequences and B cell evolution. However, there are limitations to these models to consider. Most prominent one being that these models need a fixed sequence length, and the memory and computational complexity grows with longer sequences. Second, pre-training can take a longer time when including more sequences and more layers. To deal with these shortcomings, it might be a good approach to combine PLMs with other methods such as fine-tuning, task-specific pre-training, or components of a phylogenetic analysis and network analysis.

Important work to show that PLMs can be used to study repertoire-wide B cell evolution is the introduction of AntiBERTy and its observation of trajectories in the immune repertoires of four donors where they observe key binding residues[50]. Another powerful result from a different study was the ability of PLM learned mutations to efficiently guide artificial evolution and produce synthetic antibodies with high binding affinity[3]. A solid application of these approaches would be to create more broadly neutralizing antibodies against viral infections. As many viruses mutate quickly, vaccines would be more effective if the induced antibodies can bind to potential upcoming variants. Employing artificial intelligence methods, next to experimental approaches, can accelerate the process of identifying key mutations for a higher binding affinity. This is especially important in situations where time is of the essence, such as a pandemic or a virus that is highly virulent. The production of high quality synthetic antibodies could also be useful for antibody-based immunotherapy for diseases such as cancer, auto-immune diseases or allergies.

Before PLM-guided monoclonal antibodies can become clinically relevant, various potential challenges need to be addressed. It is not yet clear what kind of pre-trained model is the most optimal to analyze B cell evolution. This is not only a trade-off between enough data and low computational burden, but the exact architecture and biological information are of big influence on the models performance. While Hie *et al.* showed that a foundational PLM without task-specific information can guide antibody evolution[3], the ATUE benchmark showed that a receptor-specific PLM with additional biological information during pre-training can predict B cell evolution with higher accuracy[62]. Some biological mechanisms to consider are the germline sequence, the mutation and its position, heavy and light chain pairing and the training data diversity. When assessing evolutionary order of sequences, an accurate germline sequence is necessary. In some cases, especially with lab mice or when working with *in vitro* data, the germline sequences can be characterized with high accuracy. But in many cases, such as *in vivo* human data, the germline needs to be

predicted. Some phylogenetic methods for lineage assignment are capable of reconstructing the germline from an IgSeq experiment. A second challenge are the unique mutational dynamics or antibody sequences due to SHM, clonal selection, and known hot/cold spots. Some substitution models for phylogenetic tree inference take these mechanisms into account, but it is not known how exactly the attention layers of receptor-specific PLMs consider these processes. An antibody needs both its heavy and its light chain to bind to an antigen. Despite the necessity of this combination, PLMs trained on unpaired antibody sequences are able to predict binding affinity and the mechanisms behind this phenomenon remain elusive. A recent study did show that fine-tuning a PLM with paired antibody sequences increased its accuracy[58]. This leads to a fundamental aspect of every model; what training data should be used? As stated before, general protein sequences can be used or only BCR sequences, these can be paired or unpaired, exclusively human, mice, or other species. Multiple identical sequences can be included, but the dataset can also be non-redundant. And what part of the data should be used for pre-training, fine-tuning, testing, or evaluation? The options are endless, and thus far have shown to be task-specific. Another challenge is to correctly use the PLM embeddings to infer evolutionary relationships. The creation of a vector field based on evo-velocity score is a huge step in the direction of identifying trajectories of B cell evolution. But like phylogenetic trees, the absence of a ground truth is a considerable limitation. Finally, when using synthetic antibodies to design vaccines, certain aspects of stability and safety are not considered by PLMs. The antigens must remain effective after storage, and they should not initiate unintended immune responses.

Despite all these uncertainties, the use of PLMs to study B cell evolution and engineer synthetic antibodies shows great promise. With the increasing amount of sequenced proteins and computational power, the full potential of PLMs has not been reached yet. We hope this review gave a clear overview of the different approaches to study B cell evolution, and a promising outlook on its downstream applications.

References

[1] John Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. en. In: *Nature* 596.7873 (Aug. 2021), pp. 583–589.

[2] Joshua Meier et al. “Language models enable zero-shot prediction of the effects of mutations on protein function”. en. Nov. 2021.

[3] Brian L Hie et al. “Efficient evolution of human antibodies from general protein language models”. en. In: *Nat. Biotechnol.* (Apr. 2023).

[4] Maria Ramunno and Ryan Savitz. “COVID-19 vaccination and decreased death rates: A county-level study in Pennsylvania”. en. In: *J. Med. Virol.* 95.7 (July 2023), e28883.

[5] Kenneth B Hoehn et al. “Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 116.45 (Nov. 2019), pp. 22664–22672.

[6] Chaim A Schramm et al. “SONAR: A High-Throughput Pipeline for Inferring Antibody Ontogenies from Longitudinal Sequencing of B Cell Transcripts”. en. In: *Front. Immunol.* 7 (Sept. 2016), p. 372.

[7] Michal Barak et al. “IgTree: creating Immunoglobulin variable region gene lineage trees”. en. In: *J. Immunol. Methods* 338.1-2 (Sept. 2008), pp. 67–74.

[8] Bruce Alberts et al. *B Cells and Antibodies*. Garland Science, 2002.

[9] S Tonegawa. “Somatic generation of antibody diversity”. en. In: *Nature* 302.5909 (Apr. 1983), pp. 575–581.

[10] Bruce Alberts et al. *The Generation of Antibody Diversity*. Garland Science, 2002.

[11] M A Oettinger et al. “RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination”. en. In: *Science* 248.4962 (June 1990), pp. 1517–1523.

[12] Jonathan A Roco et al. “Class-Switch Recombination Occurs Infrequently in Germinal Centers”. en. In: *Immunity* 51.2 (Aug. 2019), 337–350.e7.

[13] J Jacob et al. “Intraclonal generation of antibody mutants in germinal centres”. en. In: *Nature* 354.6352 (Dec. 1991), pp. 389–392.

[14] M Muramatsu et al. “Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme”. en. In: *Cell* 102.5 (Sept. 2000), pp. 553–563.

[15] Reuben S Harris, Svend K Petersen-Mahrt, and Michael S Neuberger. “RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators”. en. In: *Mol. Cell* 10.5 (Nov. 2002), pp. 1247–1253.

[16] T Goossens, U Klein, and R Küppers. “Frequent occurrence of deletions and duplications during somatic hypermutation: implications for oncogene translocations and heavy chain disease”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 95.5 (Mar. 1998), pp. 2463–2468.

[17] Nai-Ying Zheng et al. “Intricate targeting of immunoglobulin somatic hypermutation maximizes the efficiency of affinity maturation”. en. In: *J. Exp. Med.* 201.9 (May 2005), pp. 1467–1478.

[18] Dan Liu et al. “T-B-cell entanglement and ICOSL-driven feed-forward regulation of germinal centre reaction”. en. In: *Nature* 517.7533 (Jan. 2015), pp. 214–218.

[19] Ziv Shulman et al. “Dynamic signaling by T follicular helper cells during germinal center B cell selection”. en. In: *Science* 345.6200 (Aug. 2014), pp. 1058–1062.

[20] Yang Li et al. “Immune history shapes specificity of pandemic H1N1 influenza antibody responses”. en. In: *J. Exp. Med.* 210.8 (July 2013), pp. 1493–1500.

[21] Chen Wang et al. “Effects of aging, cytomegalovirus infection, and EBV infection on human B cell repertoires”. en. In: *J. Immunol.* 192.2 (Jan. 2014), pp. 603–611.

[22] Chen Wang et al. “B-cell repertoire responses to varicella-zoster vaccination in human identical twins”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 112.2 (Jan. 2015), pp. 500–505.

[23] Poornima Parameswaran et al. “Convergent antibody signatures in human dengue”. en. In: *Cell Host Microbe* 13.6 (June 2013), pp. 691–700.

[24] Jason G Cyster and Christopher D C Allen. “B Cell Responses: Cell Interaction Dynamics and Decisions”. en. In: *Cell* 177.3 (Apr. 2019), pp. 524–540.

[25] Florian J Weisel et al. “A Temporal Switch in the Germinal Center Determines Differential Output of Memory B and Plasma Cells”. en. In: *Immunity* 44.1 (Jan. 2016), pp. 116–130.

[26] Jeroen M J Tas et al. “Visualizing antibody affinity maturation in germinal centers”. en. In: *Science* 351.6277 (Mar. 2016), pp. 1048–1054.

[27] Duncan K Ralph and Frederick A Matsen 4th. “Likelihood-Based Inference of B Cell Clonal Families”. en. In: *PLoS Comput. Biol.* 12.10 (Oct. 2016), e1005086.

[28] Alexander Dimitri Yermanos et al. “Tracing Antibody Repertoire Evolution by Systems Phylogeny”. en. In: *Front. Immunol.* 9 (Oct. 2018), p. 2149.

[29] J Felsenstein and G A Churchill. “A Hidden Markov Model approach to variation among sites in rate of evolution”. en. In: *Mol. Biol. Evol.* 13.1 (Jan. 1996), pp. 93–104.

[30] Bryan Briney et al. “Clonify: unseeded antibody lineage assignment from next-generation sequencing data”. en. In: *Sci. Rep.* 6 (Apr. 2016), p. 23901.

[31] Kenneth B Hoehn, Gerton Lunter, and Oliver G Pybus. “A Phylogenetic Codon Substitution Model for Antibody Lineages”. en. In: *Genetics* 206.1 (May 2017), pp. 417–427.

[32] Gur Yaari et al. “Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data”. en. In: *Front. Immunol.* 4 (Nov. 2013), p. 358.

[33] Ang Cui et al. “A Model of Somatic Hypermutation Targeting in Mice Based on High-Throughput Ig Sequencing Data”. en. In: *J. Immunol.* 197.9 (Nov. 2016), pp. 3566–3574.

[34] William S DeWitt 3rd et al. “Using Genotype Abundance to Improve Phylogenetic Inference”. en. In: *Mol. Biol. Evol.* 35.5 (May 2018), pp. 1253–1265.

[35] J Felsenstein. “PHYLIP-Phylogeny inference package (Version 3.2)”. en. In: *Cladistics* 5 (Jan. 1989), pp. 164–166.

[36] Remco Bouckaert et al. “BEAST 2: a software platform for Bayesian evolutionary analysis”. en. In: *PLoS Comput. Biol.* 10.4 (Apr. 2014), e1003537.

[37] Robert K Abbott et al. “Precursor Frequency and Affinity Determine B Cell Competitive Fitness in Germinal Centers, Tested with Germline-Targeting HIV Vaccine Immunogens”. en. In: *Immunity* 48.1 (Jan. 2018), 133–146.e6.

[38] Haiyuan Yu et al. “The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics”. en. In: *PLoS Comput. Biol.* 3.4 (Apr. 2007), e59.

[39] Ashish Vaswani et al. “Attention Is All You Need”. In: (June 2017). arXiv: `1706.03762 [cs.CL]`.

[40] Tom Brown et al. “Language models are few-shot learners”. In: *Adv. Neural Inf. Process. Syst.* 33 (2020), pp. 1877–1901.

[41] J Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv* (2018).

[42] Abel Chandra et al. “Transformer-based deep learning for predicting protein properties in the life sciences”. en. In: *Elife* 12 (Jan. 2023).

[43] Alexander Rives et al. “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 118.15 (Apr. 2021).

[44] Kyunghyun Cho et al. “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation”. In: (June 2014). arXiv: 1406.1078 [cs.CL].

[45] S Hochreiter and J Schmidhuber. “Long short-term memory”. en. In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780.

[46] F A Gers. “Learning to forget: continual prediction with LSTM”. In: *9th International Conference on Artificial Neural Networks: ICANN ’99*. Edinburgh, UK: IEE, 1999.

[47] Baris E Suzek et al. “UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches”. en. In: *Bioinformatics* 31.6 (Mar. 2015), pp. 926–932.

[48] Tobias H Olsen, Fergus Boyles, and Charlotte M Deane. “Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences”. en. In: *Protein Sci.* 31.1 (Jan. 2022), pp. 141–146.

[49] Jinwoo Leem et al. “Deciphering the language of antibodies using self-supervised learning”. en. In: *Patterns (N Y)* 3.7 (July 2022), p. 100513.

[50] Jeffrey A Ruffolo, Jeffrey J Gray, and Jeremias Sulam. “Deciphering antibody affinity maturation with language models and weakly supervised learning”. In: (Dec. 2021). arXiv: 2112.07782 [q-bio.BM].

[51] Tobias H Olsen, Iain H Moal, and Charlotte M Deane. “AbLang: an antibody language model for completing antibody sequences”. en. In: *Bioinform Adv* 2.1 (June 2022), vbac046.

[52] David Prihoda et al. “BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning”. en. In: *MAbs* 14.1 (2022), p. 2020203.

[53] Ahmed Elnaggar et al. “ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning”. en. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 44.10 (Oct. 2022), pp. 7112–7127.

[54] Zeming Lin et al. “Language models of protein sequences at the scale of evolution enable accurate structure prediction”. en. July 2022.

[55] Jeffrey A Ruffolo et al. “Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies”. en. In: *Nat. Commun.* 14.1 (Apr. 2023), p. 2389.

[56] Jeffrey A Ruffolo, Jeremias Sulam, and Jeffrey J Gray. “Antibody structure prediction using interpretable deep learning”. en. In: *Patterns (N Y)* 3.2 (Feb. 2022), p. 100406.

[57] James Dunbar et al. “SAbDab: the structural antibody database”. en. In: *Nucleic Acids Res.* 42.Database issue (Jan. 2014), pp. D1140–6.

[58] Sarah M Burbach and Bryan Briney. “Improving antibody language models with native pairing”. In: (Aug. 2023). arXiv: 2308.14300 [q-bio.BM].

[59] Brian L Hie, Kevin K Yang, and Peter S Kim. “Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins”. en. In: *Cell Syst* 13.4 (Apr. 2022), 274–285.e6.

[60] Brian Hie et al. “Learning the language of viral evolution and escape”. en. In: *Science* 371.6526 (Jan. 2021), pp. 284–288.

[61] Wenkai Han et al. “Predicting the antigenic evolution of SARS-COV-2 with deep learning”. en. In: *Nat. Commun.* 14.1 (June 2023), p. 3478.

[62] Danqing Wang, Fei Ye, and Zhou Hao. “On Pre-trained Language Models for Antibody”. en. Jan. 2023.