Immunolinguistics

Applying linguistics to biological sequences

Mai Ha Vu

November 5, 2021

Outline

Overview of the talk

- Interdisciplinary project involving: immunology, informatics, statistics, and linguistics!
- Goal: use a combination of these methods to learn more about how immune cells work
- Today I will talk about:
  - What is adaptive immunity?
  - Challenges of the work
    - (talking across disciplines, e.g. explaining what linguistics is)
    - the specific challenges of applying linguistics to biological sequences
  - Our proposed framework to bridge the challenges, and hopefully to lay down a way to apply linguistics more generally to biological sequences
- Disclaimer: I am not an expert in immunology. I took a lot of figures/information from presentations/publications by people in the Immunology Department!

My background

PhD from University of Delaware

- Theoretical, formal syntax
- Some formal semantics
- Mathematical linguistics (using formal language theory) – studied how mathematically/computationally complex certain linguistic patterns are
- Had one side project with applying linguistics to "robot" language
- NO background in current artificial intelligence research/Natural Language Processing (NLP)
- NO real background in biology

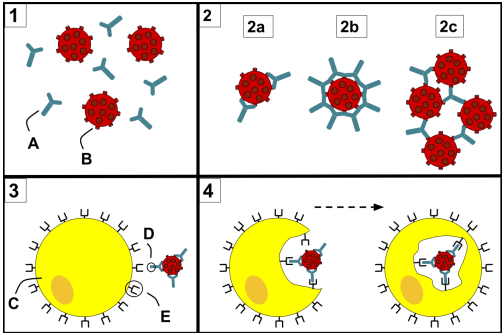
A simplified summary of the adaptive immune response

What happens when you get a pathogen (virus, bacteria, allergen) in your body?

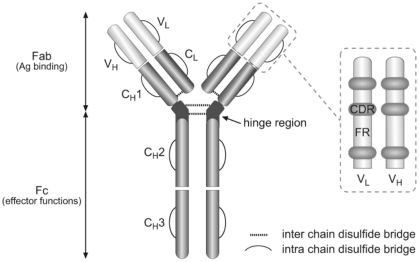
1. Pathogen or other undesirable substance gets into the body
2. For intracellular pathogens: T-cells kill infected cells
3. For extracellular pathogens: antibodies (which are produced by B-cells) bind to them, which
   - neutralizes them
   - marks them for elimination
4. Rapid proliferation of immune cells that recognized the antigen, some of these are memory cells that can live for decades

(we call the molecules recognized by immune cells *antigens*)
→ Immune cells have to be diverse and specific: recognize "bad" things, don't recognize the "self"

Antibodies

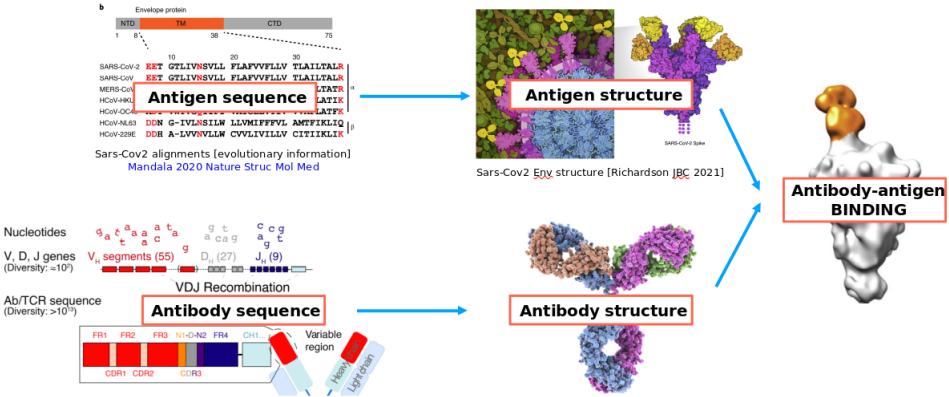


By Maher33 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=69535486

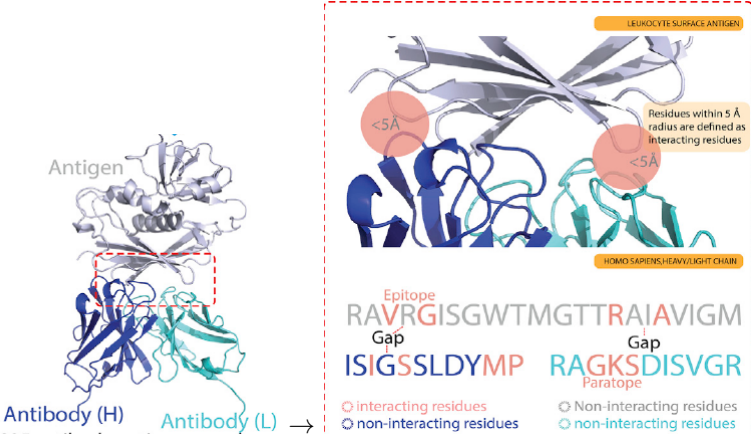


Greiff (2013). Exploring the genesis and specificity of serum antibody binding.

Antigen-antibody interaction, overview



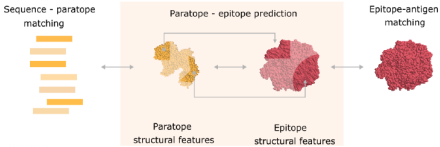
Antigen-antibody interaction, more detail



Our main research question

The antibody specifity problem:
given an antibody, which proteins would bind to it with high affinity and vice versa?

- Classification:
  - binary classificiation: does this antibody bind this antigen? (∼syntactic well-formedness)
  - multiclass classification: which antigens does this antibody bind? (∼semantics of antibody?)
- Regression: given an antibody and an antigen, what is the binding affinity? (∼ how well-formed is this antibody?)
- Paratope-epitope prediction (structure prediction):
  - what parts of the antibody form the paratope?
  - what parts of the antigen form the epitope?
  - what parts interact with each other?



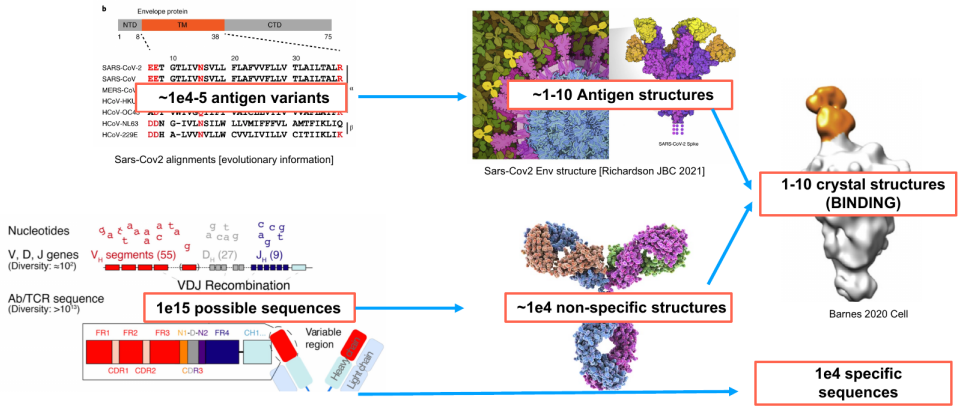
Medical applications

- **Generate therapeutic antibody drugs**: if we have a new virus (e.g. Covid) or cancerous cells, we can immediately generate a successful antibody that recognizes it, which helps having a fast immune attack on the virus/cancerous cell
- **Vaccine design**: design vaccines that contain proteins that share the important features with the real viruses without being the real ones
- **Diagnostics**: detect if somebody has a certain disease based on the antibodies in their body

What makes working with antibodies hard?

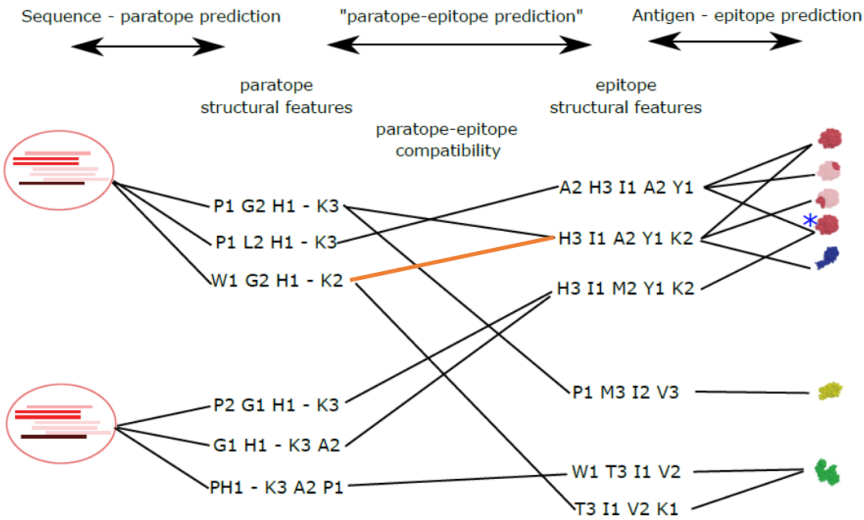
- Limited data
- Cross-reactivity
- 3D structure
- Too different from general proteins, so tools that work for proteins in general do not work well for antibody specificity

Limited data



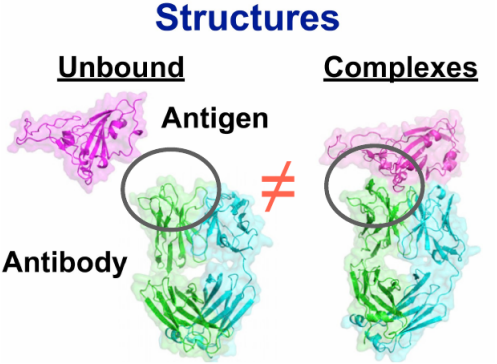
Also, the data is very noisy!

Cross-reactivity



3D structure

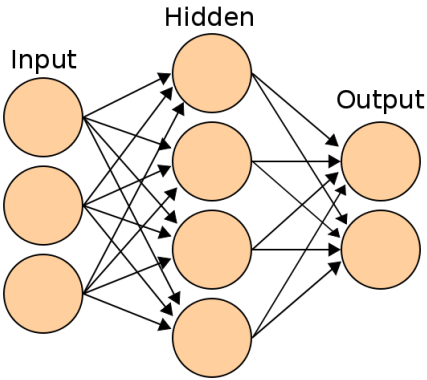
- actual structure of the sequences change when bound vs. unbound
- antibodies also take up different structure depending on what they bind
  $\rightarrow \sim$ structure determines meaning



Guest 2021 (Structure)

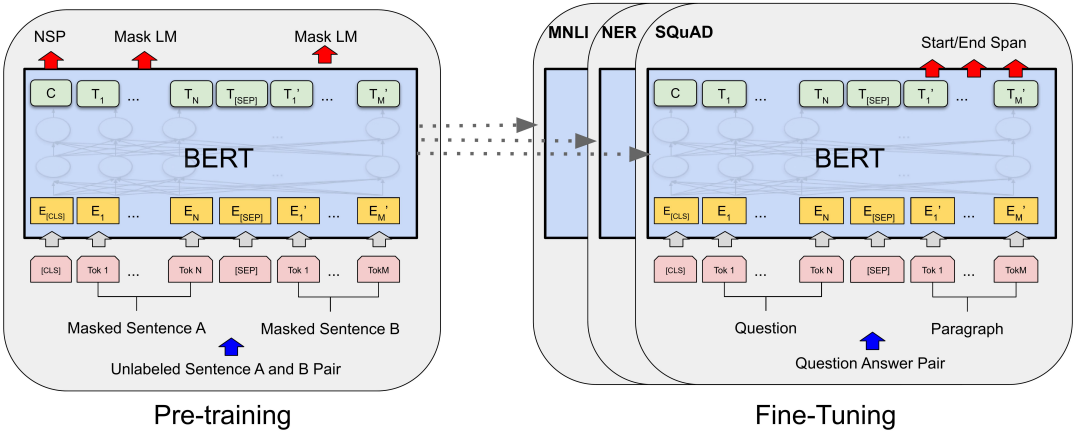
Current approach: Neural networks from ML

- powerful tool in artificial intelligence
- it can learn patterns from **large data** using statistics
- it can generate new things from the patterns it learned
- downside: "black box" learner, how much does it *really* understand?
- some applications:
  - image processing
  - natural language
  - biology: some good results for proteins in general, lot of difficulties for antibody specificity problem



Neural network-based language models

BERT:



Language models in biology

- Growing popularity of protein language models, because we can feed a lot of unlabeled sequence data
  - Some success with predicting a number of biological features, BUT
  - Many of them are rather coarse-grained and about large structures, none of them seems to be as fine grained as antibody specificity prediction
- Differences from applications to natural language:
  - Discrete units are usually assumed to be amino acids or 3-grams – not really a concept of lexical items

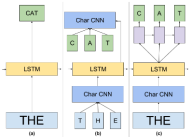


Figure 1. A high-level diagram of the models presented in this paper. (a) is a standard LSTM LM. (b) represents an LM where both input and Softmax embeddings have been replaced by a character CNN. In (c) we replace the Softmax by a next character prediction LSTM network.

Jozefowicz et al. 2016

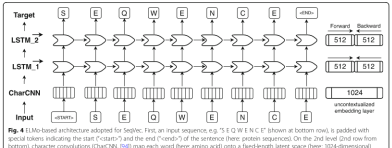


Fig. 4 ELMo-based architecture adopted for SeqVec. First, an input sequence, e.g. "S E Q W E N C E" (shown at bottom row), is padded with special tokens indicating the start ("<start>") and the end ("<end>") of the sentence (here: protein sequences). On the 2nd level (2nd row from bottom), character convolutions (CharCNN, [94]) map each word (here: amino acid) onto a fixed-length latent space (here: 1024-dimensional) without considering information from neighboring words. On the third level (3rd row from bottom), the output of the CharCNN-layer is used as input by a bidirectional Long Short Term Memory (LSTM, [93]) which introduces context-specific information by processing the sentence (protein sequence) sequentially. For simplicity, only the forward pass of the bi-directional LSTM-layer is shown (here: 512-dimensional). On the fourth level (4th row from bottom), the second LSTM-layer operates directly on the output of the first LSTM-layer and tries to predict the next word given all previous words in a sentence. The forward and backward pass are optimized independently during training in order to avoid information leakage between the two directions. During inference, the hidden states of the forward and backward pass of each LSTM-layer are concatenated to a 1024-dimensional embedding vector summarizing information from the left and the right context

Heinzinger et al. 2019

- We have very little ground truth knowledge about the biological sequences, the most we know is desired output labels → hard to study whether the model learned meaningful rules

Downsides for antibody specificity problem

- Does not give us interpretable rules
- Even if it works for proteins, it is unlikely to work for antibodies, due to
  - huge diversity in sequences
  - similar sequences might behave very differently, different sequences might behave similarly
  - lot of "orphan sequences", which are dissimilar to anything else we have, so cannot just learn from commonly seen patterns
- Requires a huge amount of structural, labeled data that we do not have
  $\rightarrow$ immunologists have created a huge database of synthetic data, just so we can test different methods

Can linguistics help?

Outline

The potential role of linguistics

- The core assumptions in (theoretical) linguistics:
  - language is a set of strings, built from a finite set of components (sounds, morphemes, words, phrases) with the use of some (finite set of) rules
  - the strings have semantic meaning which we can derive compositionally
- Hopeful assumption: this core concept is true for the biological strings too, but to what extent?
  - DNA: combination of 4 nucleic acids
  - Proteins: combination of 20 amino acids

  → Can we assume more, e.g. that there are interpretable syntactic and semantic rules? are the semantic rules compositional like in language?

What do we need to successfully apply linguistics?

- How do we define the 'language'?
- What counts as 'well-formed' in the language?
- What is the meaning of the strings in the language?
- What are the discrete units that the rules apply to?

All of these have many possible, equally good answers for biology!

Linguistics has it easy

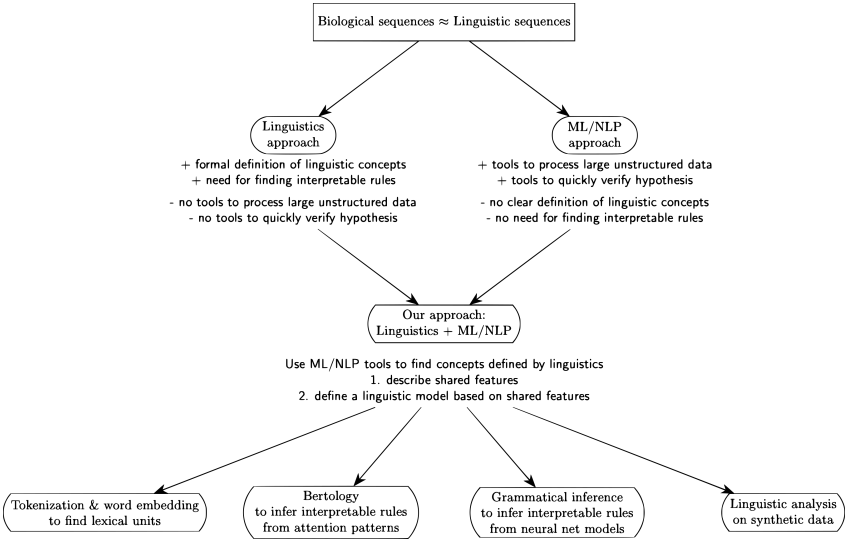
- Easy to access and query data: we *can* ask speakers to generate, judge, and interpret linguistic sequences for us
- Intuitions about the discrete parts: we can intuit, or at least test our ideas through elicitation, about the building blocks: phonemes, morphemes, words, etc.

→we lack all of these in biology (or at least it would take a lot of money and time), we need something that can process large, noisy data fast

We need to synthesize linguistics with NLP/ML

- Linguistics:
  - Provides a clear, formal definition of concepts
  - Strives to find interpretable, discrete rules
  - No tools to process large unannotated data for biology
  - No tools to quickly verify a hypothesis
- ML/NLP
  - Not really a clear, formal definition of concepts – tends to use whatever is most convenient
  - Emphasis on accurately modeling existing data, not finding interpretable, discrete rules
  - Has tools to process large unannotated data for biology
  - Has tools to quickly verify a hypothesis

Outline


Biological sequences ≈ Linguistic sequences
Linguistics approach
ML/NLP approach
+ formal definition of linguistic concepts
+ need for finding interpretable rules
- no tools to process large unstructured data
- no tools to quickly verify hypothesis
+ tools to process large unstructured data
+ tools to quickly verify hypothesis
- no clear definition of linguistic concepts
- no need for finding interpretable rules
Our approach: Linguistics + ML/NLP
Use ML/NLP tools to find concepts defined by linguistics
1. describe shared features
2. define a linguistic model based on shared features
Tokenization & word embedding to find lexical units
Bertology to infer interpretable rules from attention patterns
Grammatical inference to infer interpretable rules from neural net models
Linguistic analysis on synthetic data

Our proposed framework

1. Analogies: pinpoint the similarities and differences between linguistic and specific biological sequence, "in what ways can we talk about this like it is language?"
2. Conceptual models: develop linguistically-informed models of the biological sequence, "if this was natural language, what are our requirements for its parts?"
3. Practical integration with ML: use the conceptual models to really define what we look for when we use ML, "how can we use ML to look for the parts we defined in the models?"

Analogies between natural language and biological language

Some potentially relevant features:

- Discreteness:
  - linguistic sequences are built through the combination from smaller, discrete units
  - antibody sequences are built from a combination of amino acids, and we hope there are more meaningful discrete units (motifs) in-between
- Compositional semantics
  - discrete linguistic units have meaning, they can combine compositionally toward more complex meaning
  - antibody sequences: meaning are the antigens they bind, but compositionality is an open question
- Structure
  - linguistic sequences have structure (e.g. syllable structure, tree structure)
  - antibody sequences form 3D structure
- Ambiguity
  - a linguistic sequence can have multiple meaning, reflected in their structure and following the rules of compositional semantics
  - antibody sequences can be ambiguous too, reflected in their structure

Analogies help define the basics

- How do we define the 'language'?
- What counts as 'well-formed' in the language?
- What is the meaning of the strings in the language?
- What are the discrete units that the rules apply to?

Language	Well-formedness	Meaning
All antibody sequences	all antibody sequences	antigen(s) bound
Antibody sequences specific to one given antigen	only sequences that bind the same antigen	bound epitopes on antigen??

Conceptual models

These are hypothetical, sort of an ideal wish for how biology should work to be truly linguistic.

- Semantic model: requires a lot of annotated data
  - Well-formed sequences: all antibody sequences (ill-formed ones are those that never get generated)
  - Discrete units: some motifs with functional meaning
    e.g. CARICATURAL → CAR + I + CAT + URAL, CxxI + AR + CA + TxxxL + URA
    →Each discrete unit has some well-define 'meaning', e.g. CxxI signals something about the recognizable antigen's shape
  - Compositional semantic rules: e.g. CxxI + AR means it will recognize a spiky antigen, but CxxI + CA means it will recognize an antigen with a different shape
- Syntactic model: requires only data about one specific antigen
  - Well-formed sequences: all antibody sequences that bind one given antigen
  - Discrete units: motifs, but we don't necessarily care about their meanings
  - Syntactic well-formedness rules: define how motifs can combine so that the sequence would be 'well-formed'

Integration with ML

Our goal as linguist is to:

- Find meaningful units (lexical items)
- Extract rules about these items

Find meaningful units: tokenization

- Current status quo: protein language models assume amino acid or 3-gram level tokens
- What if we could find more meaningful tokens that resemble lexical items?
  - On the biology side, they have experimented with feeding neural networks with information about paratopes (which AAs on the antibody bind to the antigen), and that helped a lot
  - Maybe we could tokenize the sequence based on this information, and see if language models would perform better than when we just tokenize based on AAs or 3-grams → but this will only be useful in practice if we can automatically find these tokenizations

Extract rules

These require that we first of all have a well-working neural network model!

- BERTology: subfield in NLP that aims to figure out what exactly language models like BERT has learned about language
  - Have been used for protein language model, where there were more connections in the neural net model between amino acids that interacted with each other in a folded structure
- Grammatical inference: another subfield that aims to learn grammar from strings

Conclusions

- We have identified shortcomings of how ML/NLP is currently used when people talk about studying the "language of proteins" (or any other biological sequences)
- We have proposed a framework for how to actually define "the language of X", which should also make it clear in what ways biological sequences are *not* like natural language (e.g. the extent of compositionality, meaningful tokens)
- We are very much in the beginning of all this: our application of the framework is only to a very small (and still very big and complex) question in adaptive immunity
- Lot of work ahead of us!

Acknowledgements

- Prof. Dag Haug, Department of Linguistics and Scandinavian Studies, UiO
- Prof. Victor Greiff, Department of Immunology, UiO
- Prof. Geir Kjetil Sandve, Department of Informatics, UiO
- Prof. Ingrid Hobæk Haff, Department of Mathematics, UiO
- Prof. Bartlomiej Swiatczak, Department of History of Science, University of Science and Technology of China, China
- Dr. Philippe Robert, Department of Immunology, UiO
- Dr. Rahmad Akbar, Department of Immunology, UiO

Thank you for listening!