Immunolingo

Leveraging linguistic insights to answer immunological questions

Mai Ha Vu

March 28, 2022

Overview of the talk

- Interdisciplinary project involving: immunology, informatics, statistics, and linguistics!
- Goal: use a combination of these methods to learn more about the adaptive immune response
- Today I will talk about:
    - The core research question of the project
    - Reasons for applying linguistics methodology (and previous work)
    - Challenges of applying linguistics methodology
    - 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!

Outline

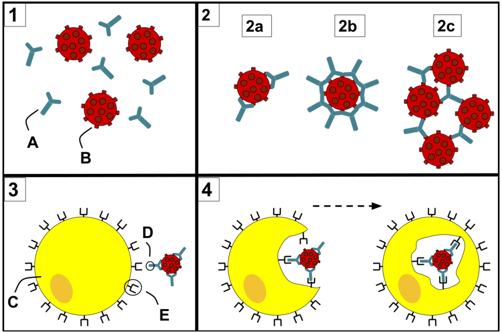
The research question

Why Linguistics?

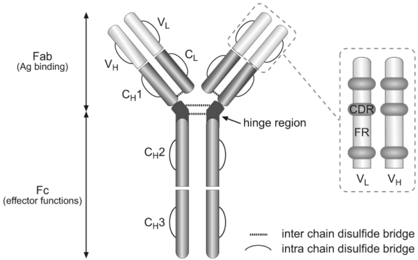
Challenges

Our framework

Adaptive immune response with antibodies

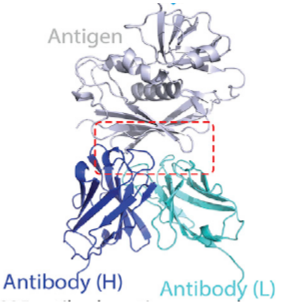


By Maher33 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=69535486

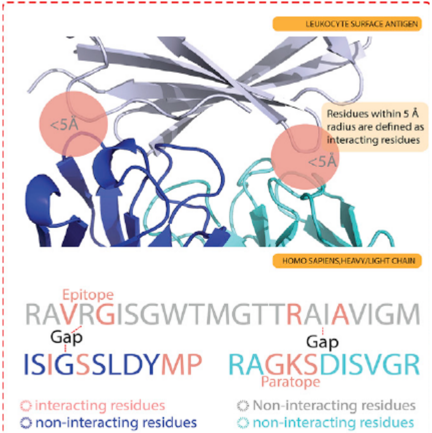


Greiff (2013). Exploring the genesis and specificity of serum antibody binding.

Antigen-antibody interaction, more detail

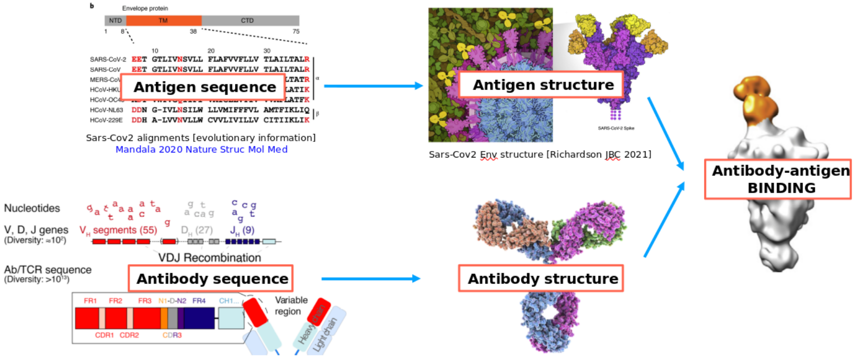


→



Our main research question

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



In other words, *What is the "grammar" of antibody receptors?*

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

Outline

The research question

Why Linguistics?

Challenges

Our framework

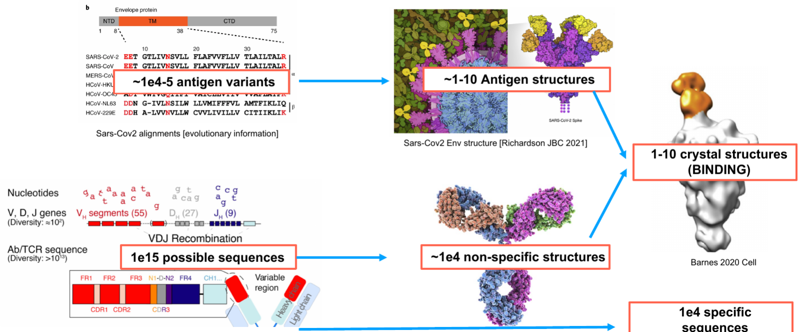
The main commonality: sequence rules

- 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, encoding some 'message'
  - Proteins: combination of 20 amino acids, encoding some function

$\rightarrow$ To what extent are biological sequence rules are like linguistic rules?

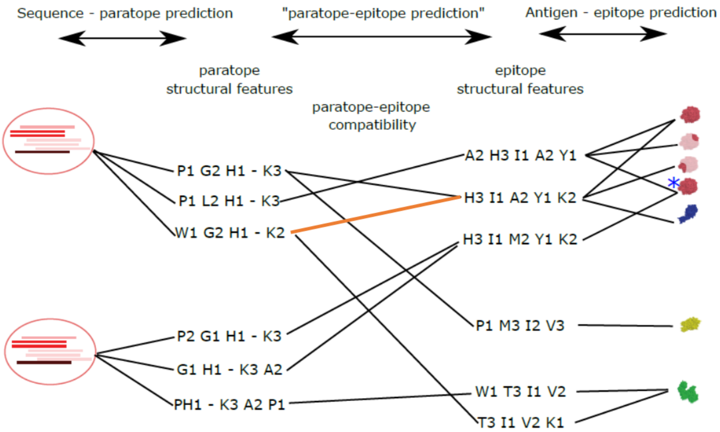
Linguistics for antibody-specific challenges

- Limited structural data, many more sequence data ∼ Inferring structure and rules from linguistic sequences



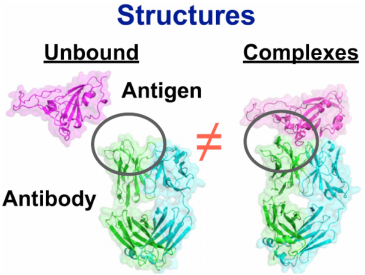
Linguistics for antibody-specific challenges

- Limited structural data, many more sequence data ∼ Inferring structure and rules from linguistic sequences
- Cross-reactivity ∼ Linguistic ambiguity



Linguistics for antibody-specific challenges

- Limited structural data, many more sequence data ∼ Inferring structure and rules from linguistic sequences
- Cross-reactivity ∼ Linguistic ambiguity
- Variable 3D structure ∼ Linguistic allomorphy/allophony



Guest 2021 (Structure)

Linguistics for antibody-specific challenges

- Limited structural data, many more sequence data ∼ Inferring structure and rules from linguistic sequences
- Cross-reactivity ∼ Linguistic ambiguity
- Variable 3D structure ∼ Linguistic allomorphy/allophony
- No connection between sequence similarity and function ∼ Linguistic arbitrariness

Previous work

Mostly two types:

- Many hypotheticals that allude to future possibilities
  - Jerne (1985): generative grammar of the immune system
- Analysis of already known biological patterns in formal grammar terms
  - Searl: formal grammar analysis of biological structures

Language	*Automaton*	*Grammar*	*Recognition*	*Dependencies*	*Biosequences*
Recursively Enumerable Languages	Turing Machine	Unrestricted Grammar $Baa \rightarrow a$	Undecidable	Arbitrary ?	Unknown ?
Context-Sensitive Languages	Linear-Bounded Automaton	Context-Sensitive Grammar $At \rightarrow aA$	NP-Complete	Crossing	Repeats Pseudoknots
Context-Free Languages	Pushdown Automaton	Context-Free Grammar $S \rightarrow gSc$	Polynomial	Nested	Orthodox Secondary Structure
Regular Languages	Finite-State Automaton	Regular Expression $((g\|a)(c\|t))^*$	Linear	Strictly Local	Central Dogma

But no *new rule* extraction from biological sequences → can linguistics do that?

Outline

The research question

Why Linguistics?

Challenges

Our framework

What are sequence rules?

The types of rules we seek are different:
In linguistics, syntactic rules need to be exhaustive:

- state all constraints that are needed to generate a 'grammatical' sentence (sufficient constraints)
- state all constraints that prevent 'ungrammatical' sentences (necessary constraints)

In biology, the interest currently is in sufficient, simple rules – but maybe to solve the main research question, we need more:

- 'If the antibody sequence has a "GKS" as subsequence, then it binds COVID'

→ Can we (and should we) seek more linguistic rules for biological data?
We are going to try!

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.
- It is somewhat easier to eliminate noisy data based on knowledge of the language

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

Synthesize linguistics with machine learning

- Linguistics:
  - ☺ Provides a clear, formal definition of concepts
  - ☺ Strives to find interpretable, discrete rules
  - ☺ Potential suitability for antibody-specific challenges
  - ☹ No tools to process large unannotated data for biology
  - ☹ No tools to quickly verify a hypothesis
- Machine learning:
  - ☹ Not really a clear, formal definition of concepts – tends to use whatever is most convenient
  - ☹ Emphasis on accurately modeling existing data, not on finding interpretable, discrete rules
  - ☹ Not well-suited for antibody-specific challenges
  - ☺ Has tools to process large unannotated data for biology
  - ☺ Has tools to quickly verify a hypothesis

Outline

The research question

Why Linguistics?

Challenges

Our framework

What do we need to successfully apply linguistics?

The core questions:

- 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?
- What is the nature of the rules?

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

Our proposed framework

1. Analogies: In what ways can we talk about this like it is language?
2. Conceptual models: If the biological system was natural language, what are our requirements for its parts?
3. Practical integration with ML: How can we use ML to look for the parts we defined in the conceptual model?

Analogies

Some potentially relevant shared features:

- Discreteness: sequences are built from discrete parts
- Structure: sequences form a structure
- Ambiguity: one sequence can have multiple meaning/function
- Compositional semantics(?): meaning is calculated from how discrete parts combine into a 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?

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??

→ This helps visualize the main components, discrete units and rules

Conceptual models

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

- ‘All antibody language’: requires a lot of annotated data
  - Well-formed sequences: all antibody sequences (ill-formed ones are those that never get generated)
  - Discrete units: Motifs with functional meaning, e.g. CxxI signals something about the recognizable antigen’s shape
  - 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
- ‘Specific antigen language’: requires only data about one specific antigen, but would only model one antigen-specific language
  - Well-formed sequences: all antibody sequences that bind one given antigen
  - Discrete units: motifs, but we don’t necessarily care about their meanings
  - Rules: define how motifs can combine so that the sequence would be ’well-formed’

Integration with ML

Conceptual model defines what we are looking for:

- Defining the language → What should be the data to model with ML?
- Defining well-formedness/sequence meaning → How to label the data?
- Defining discrete units → How to encode the sequences?
- Defining rules → What types of rules do we want to try to extract from the ML model?

Conclusions

- There are intuitive parallels between language and biological sequences, but not that much on using linguistic models to learn new things about biology
- To leverage linguistics to learn new biological rules, we need
  - a rigorous definition of language fit for the specific biological question
  - a way to handle large, noisy data (→ use ML)
- We have proposed a framework to synthesize linguistics and ML:
  - rigorously define the language based on workable analogies
  - build a linguistic model based on our definitions to guide ML application
- 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

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!