A gentle introduction to ML via antibody-engineering

NCBI: Building Transparent ML/AI Solutions to Advance Biological Research Virtual Codeathon Feb. 26 – Mar 1, 2024

Digital World Biology, LLC

Project Goals

Motivation

- ML/AI is hot
- Antibodies are important in therapeutics, diagnostics, and as reagents (spatial omics, flow cytometry, histology)
- Antibodies are used to teach protein purification, ELISAs, etc. in community college workforce education
- We get requests for ML modules

ML Education Challenges

- Vocabulary
- Methods, appropriateness
- Need infrastructure: data, tools, models
- Reproducing papers is hard/impossible
- Examples lack context (maps, apples, cars)
- Step by step examples are lacking
- Teaching: sysadmin >> coding

Can we?

- Identify a small number of key concepts (regression vs neural nets vs language models)
- Identify illustrative data sets, test cases
- Create infrastructure, libraries, install commands/scripts
- Documents steps and concepts taught
- Target different levels of experience

Many Computational Tools, Resources, Methods

AbLang: An antibody language model for completing antibody sequences

Tobias H. Olsen[1], Iain H. Moal[2] and Charlotte M. Deane[1]

[1] Department of Statistics, University of Oxford, Oxford OX1 3LB, United Kingdom
[2] GSK Medicines Research Centre, GSK, Stevenage SG1 2NY, United Kingdom

Antibody Watch: Text Mining Antibody Specificity from the Literature

Chun-Nan Hsu[1], Chia-Hui Chang[1,2], Thamolwan Poopradubsil[2], Amanda Lo[1], Karen A. William[1], Ko-Wei Lin[1], Anita Bandrowski[1,3], Ibrahim Burak Ozyurt[1], Jeffrey S. Grethe[1], and Maryann E. Martone[1,3*]

Antibody H3 Structure Prediction

C. Marks*, C.M. Deane

Department of Statistics, University of Oxford, 24-29 St Giles', Oxford OX1 3LB, United Kingdom

ARTICLE INFO

Article history:
Received 16 Decemb
Received in revised f
Accepted 27 January
Available online 1 Fe

Keywords:
Antibodies
H3
Loop modelling
Protein structure pre

ABSTRACT

Patterns

Deciphering the language of antibodies using self-supervised learning

EpiBERTope: a sequence-based pre-trained BERT model improves linear and structural

Generating Human Antibodies Using Language Models

Mira Partha mpartha@stanford.edu
Felipe Calero Forero fcalero@stanford.edu
Aaron Behr abehr@stanford.edu

Review

Linear B-Cell Epitope Prediction for In Silico Vaccine Design: A Performance Review of Methods Available via Command-Line Interface

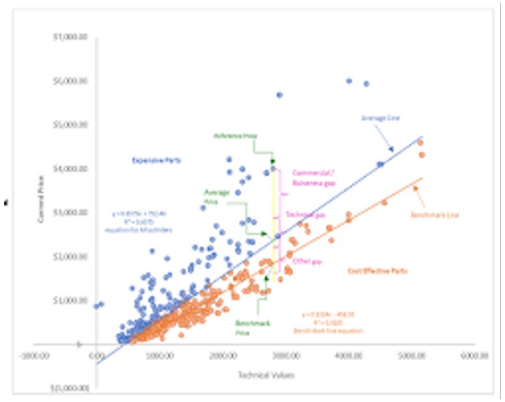
Kosmas A. Galanis, Katerina C. Nastou †, Nikos C. Papandreou, Georgios N. Petichakis, Diomidis G. Pigis and Vassiliki A. Iconomidou *

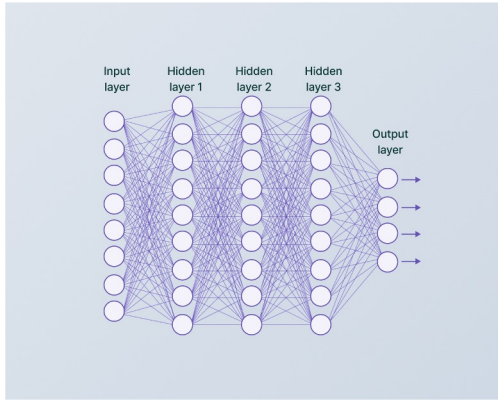
Emerged in the past 5-6 yrs

Machine Learning (ML) Concepts

- Methods
  - Regression
  - Neural nets
  - Language models
- Training data & models
  - Features (attributes)
  - Data structures
- Data resources & code

Jason's machine learning - https://docs.google.com/presentation/d/1kSuQyW5DTnkVaZEjGYCkfOxvzCqGEFzWBy4e9Uedd9k/edit#slide=id.g2397597de6_0_71





Previous work

DWB antibody engineering - https://antibody-engineers.org

- **2022 Hackathon**
  https://antibody-engineers.org/event/antibody-engineering-hackathon-august-2022
  https://github.com/AntibodyEngineers/covid-not-covid
  - ML – Ablang, covid not covid, a developer was successful
  - Others learned some programing, jupyter notebooks
- **2023 Hackathon**
  https://antibody-engineers.org/event/antibody-engineering-hackathon-august-2023
  https://github.com/AntibodyEngineers/immune-profiling
  - Immune profiling – data, pandas, graphing, some success with data set -> tools -> science
  - Still a lot of jupyter/basic python how tos
- **2024 Hackathon** (Aug 5-8: Zoom)
  - Build on this hackathon

Antibody Engineers

Project aims 1) develop modules to support course-based undergraduate research experiences. (CUREs) 2) investigate hackathons as a novel strategy for engaging participants in collaborative curriculum development.

Antibody Engineers is funded in part by the National Science Foundation DUE 2055036

ML and Antibodies

- A big goal:
  - Given an epitope sequence can we design an antibody *de novo*
- Other applications
  - Humanize mouse antibodies
  - Improve solubility, stability
  - Tune binding affinities
  - Convert Fab/scFab/svFv fragments into VHH (single domain, ”nanobody”) fragments
  - Multivalent antibodies
  - CAR-T / Immunotherapy
  - Antibodies to proteins in non-model organisms (one of ours)

Tools and Data

Tools

- AbLang – 2022 hackathon
  https://github.com/oxpig/AbLang,
  https://github.com/TobiasHeOl/AbLang2
- AntiBERT (AntiBERTy)
  https://github.com/dohlee/antiberty-pytorch
- SAM (Simple Antibody Model)
  https://github.com/Wang-lab-UCSD/AntPack
  https://github.com/jlparkI/humanness_evaluation
- IgFold – Ab structure prediction (IgFold -> iCn3d?)
  https://github.com/Graylab/IgFold
- More at:
  https://github.com/topics/antibody-design

Data

- OAS - Observed Antibody Space
  https://opig.stats.ox.ac.uk/webapps/oas
  - >1 billion sequences from 80 studies
- iReceptor http://ireceptor.irmacs.sfu.ca
  - Federated database
  - 5.2 Billion annotated sequences from 10,019 repositories
  - 67,000 clones from 64 repertoires
  - 133,000 sorted B/T cells from 142 repositories
- NCBI – SRA
  - Raw data from various experiments
  - Gotta dig

Schrödinger's Cat

- Trying to simplify ML requires learning it.
- Once you learn, your view of what's hard changes. You are no longer a novice.