UiO : Universitetet i Oslo

Machine Learning in Computational Biology: Overview

IN-BIOS5000/IN-BIOS9000

Milena Pavlović
Biomedical Informatics Research Group
Department of Informatics

milenpa@student.matnat.uio.no

Disclaimer

I am a machine learning researcher, not a biologist:
you are the experts there!

Learning aims

- Key points should be the intuition and high-level understanding of what machine learning is, types of problems it can help solving
- Machine learning is not a black box: every choice we make has a meaning
- Overall understanding that there is a data representation component and a machine learning algorithm
- High-level understanding of machine learning workflow, comparison and uncertainty related to it

Sequencing technologies provide data which can be examined for biological properties

CDR3	V gene	J gene	Species
CAAAERNTGELFF	TRBV28*01	TRBJ2-2*01	HomoSapiens
CAAGVENTGELFF	TRBV5-6*01	TRBJ2-2*01	HomoSapiens
CAAQATNTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAAQDSNTGELFF	TRBV5-1*01	TRBJ2-2*01	HomoSapiens
CAAQMTNTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAAQNLNTGELFF	TRBV15*01	TRBJ2-2*01	HomoSapiens
CAARDQRDLNTGELFF	TRBV2*01	TRBJ2-2*01	HomoSapiens
CAASDPNTGELFF	TRBV12-3*01	TRBJ2-2*01	HomoSapiens
CAASEMNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CACQELNTGELFF	TRBV30*01	TRBJ2-2*01	HomoSapiens
CAEGELNTGELFF	TRBV7-2*01	TRBJ2-2*01	HomoSapiens
CAGADSNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAGDYLNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAGGDPNTGELFF	TRBV7-9*01	TRBJ2-2*01	HomoSapiens
CAGGDSNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAGGRGNTGELFF	TRBV12-3*01	TRBJ2-2*01	HomoSapiens
CAGGVNPNTGELFF	TRBV5-1*01	TRBJ2-2*01	HomoSapiens
CAGQDLNTGELFF	TRBV7-2*01	TRBJ2-2*01	HomoSapiens
CAGQNLNTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAGQRANTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAIADANTGELFF	TRBV5-1*01	TRBJ2-2*01	HomoSapiens
CAIGDENTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAIGDRNTGELFF	TRBV5-5*01	TRBJ2-2*01	HomoSapiens
CAIGDRSSGEQYF	TRBV5-4*01	TRBJ2-7*01	HomoSapiens
CAIQDLNTGELFF	TRBV13*01	TRBJ2-2*01	HomoSapiens
CAIQESNTGELFF	TRBV10-3*01	TRBJ2-2*01	HomoSapiens
CAIQYANTGELFF	TRBV15*01	TRBJ2-2*01	HomoSapiens
CAIRTSGMLNTGELFF	TRBV2*01	TRBJ2-2*01	HomoSapiens

Motifs and data from VDJdb (Bagaev et al. 2020)

Sequencing technologies provide data which can be examined for biological properties

CDR3	V gene	J gene	Species
CAAAERNTGELFF	TRBV28*01	TRBJ2-2*01	HomoSapiens
CAAGVENTGELFF	TRBV5-6*01	TRBJ2-2*01	HomoSapiens
CAAQATNTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAAQDSNTGELFF	TRBV5-1*01	TRBJ2-2*01	HomoSapiens
CAAQMTNTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAAQNLNTGELFF	TRBV15*01	TRBJ2-2*01	HomoSapiens
CAARDQRDLNTGELFF	TRBV2*01	TRBJ2-2*01	HomoSapiens
CAASDPNTGELFF	TRBV12-3*01	TRBJ2-2*01	HomoSapiens
CAASEMNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CACQELNTGELFF	TRBV30*01	TRBJ2-2*01	HomoSapiens
CAEGELNTGELFF	TRBV7-2*01	TRBJ2-2*01	HomoSapiens
CAGADSNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAGDYLNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAGGDPNTGELFF	TRBV7-9*01	TRBJ2-2*01	HomoSapiens
CAGGDSNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAGGRGNTGELFF	TRBV12-3*01	TRBJ2-2*01	HomoSapiens
CAGGVNPNTGELFF	TRBV5-1*01	TRBJ2-2*01	HomoSapiens
CAGQDLNTGELFF	TRBV7-2*01	TRBJ2-2*01	HomoSapiens
CAGQNLNTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAGQRANTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAIADANTGELFF	TRBV5-1*01	TRBJ2-2*01	HomoSapiens
CAIGDENTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAIGDRNTGELFF	TRBV5-5*01	TRBJ2-2*01	HomoSapiens
CAIGDRSSGEQYF	TRBV5-4*01	TRBJ2-7*01	HomoSapiens
CAIQDLNTGELFF	TRBV13*01	TRBJ2-2*01	HomoSapiens
CAIQESNTGELFF	TRBV10-3*01	TRBJ2-2*01	HomoSapiens
CAIQYANTGELFF	TRBV15*01	TRBJ2-2*01	HomoSapiens
CAIRTSGMLNTGELFF	TRBV2*01	TRBJ2-2*01	HomoSapiens

discover motifs in the data →

Motifs and data from VDJdb (Bagaev et al. 2020)

Sequencing technologies provide data which can be examined for biological properties

CDR3	V gene	J gene	Species
CAAAERNTGELFF	TRBV28*01	TRBJ2-2*01	HomoSapiens
CAAGVENTGELFF	TRBV5-6*01	TRBJ2-2*01	HomoSapiens
CAAQATNTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAAQDSNTGELFF	TRBV5-1*01	TRBJ2-2*01	HomoSapiens
CAAQMTNTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAAQNLNTGELFF	TRBV15*01	TRBJ2-2*01	HomoSapiens
CAARDQRDLNTGELFF	TRBV2*01	TRBJ2-2*01	HomoSapiens
CAASDPNTGELFF	TRBV12-3*01	TRBJ2-2*01	HomoSapiens
CAASEMNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CACQELNTGELFF	TRBV30*01	TRBJ2-2*01	HomoSapiens
CAEGELNTGELFF	TRBV7-2*01	TRBJ2-2*01	HomoSapiens
CAGADSNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAGDYLNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAGGDPNTGELFF	TRBV7-9*01	TRBJ2-2*01	HomoSapiens
CAGGDSNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAGGRGNTGELFF	TRBV12-3*01	TRBJ2-2*01	HomoSapiens
CAGGVNPNTGELFF	TRBV5-1*01	TRBJ2-2*01	HomoSapiens
CAGQDLNTGELFF	TRBV7-2*01	TRBJ2-2*01	HomoSapiens
CAGQNLNTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAGQRANTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAIADANTGELFF	TRBV5-1*01	TRBJ2-2*01	HomoSapiens
CAIGDENTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAIGDRNTGELFF	TRBV5-5*01	TRBJ2-2*01	HomoSapiens
CAIGDRSSGEQYF	TRBV5-4*01	TRBJ2-7*01	HomoSapiens
CAIQDLNTGELFF	TRBV13*01	TRBJ2-2*01	HomoSapiens
CAIQESNTGELFF	TRBV10-3*01	TRBJ2-2*01	HomoSapiens
CAIQYANTGELFF	TRBV15*01	TRBJ2-2*01	HomoSapiens
CAIRTSGMLNTGELFF	TRBV2*01	TRBJ2-2*01	HomoSapiens

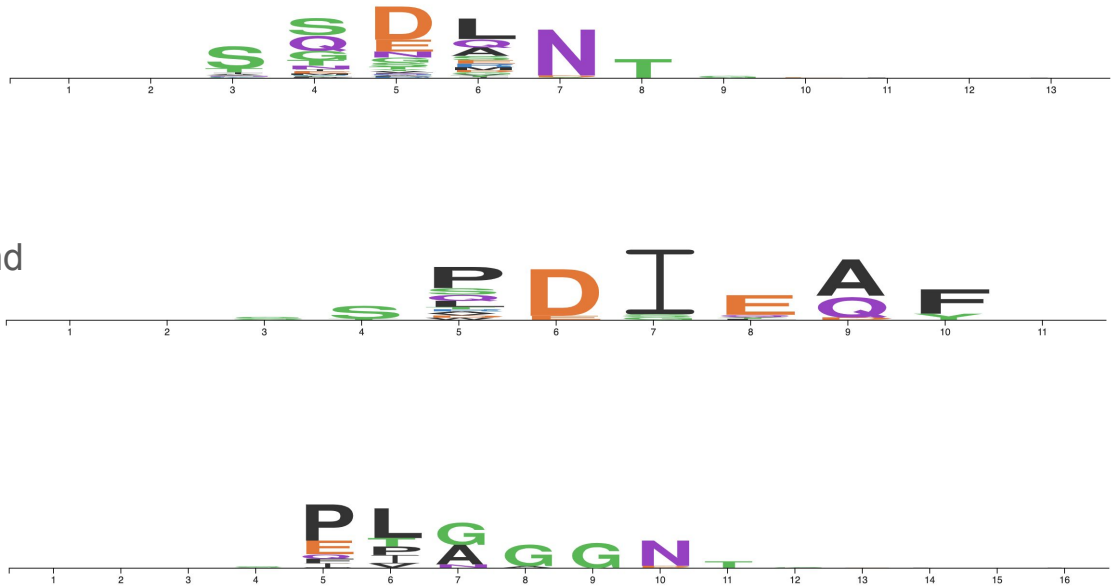
discover motifs in the data →

Motifs and data from VDJdb (Bagaev et al. 2020)

Sequencing technologies provide data which can be examined for biological properties

CDR3	V gene	J gene	Species
CAAAERNTGELFF	TRBV28*01	TRBJ2-2*01	HomoSapiens
CAAGVENTGELFF	TRBV5-6*01	TRBJ2-2*01	HomoSapiens
CAAQATNTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAAQDSNTGELFF	TRBV5-1*01	TRBJ2-2*01	HomoSapiens
CAAQMTNTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAAQNLNTGELFF	TRBV15*01	TRBJ2-2*01	HomoSapiens
CAARDQRDLNTGELFF	TRBV2*01	TRBJ2-2*01	HomoSapiens
CAASDPNTGELFF	TRBV12-3*01	TRBJ2-2*01	HomoSapiens
CAASEMNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CACQELNTGELFF	TRBV30*01	TRBJ2-2*01	HomoSapiens
CAEGELNTGELFF	TRBV7-2*01	TRBJ2-2*01	HomoSapiens
CAGADSNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAGDYLNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAGGDPNTGELFF	TRBV7-9*01	TRBJ2-2*01	HomoSapiens
CAGGDSNTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAGGRGNTGELFF	TRBV12-3*01	TRBJ2-2*01	HomoSapiens
CAGGVNPNTGELFF	TRBV5-1*01	TRBJ2-2*01	HomoSapiens
CAGQDLNTGELFF	TRBV7-2*01	TRBJ2-2*01	HomoSapiens
CAGQNLNTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAGQRANTGELFF	TRBV19*01	TRBJ2-2*01	HomoSapiens
CAIADANTGELFF	TRBV5-1*01	TRBJ2-2*01	HomoSapiens
CAIGDENTGELFF	TRBV7-8*01	TRBJ2-2*01	HomoSapiens
CAIGDRNTGELFF	TRBV5-5*01	TRBJ2-2*01	HomoSapiens
CAIGDRSSGEQYF	TRBV5-4*01	TRBJ2-7*01	HomoSapiens
CAIQDLNTGELFF	TRBV13*01	TRBJ2-2*01	HomoSapiens
CAIQESNTGELFF	TRBV10-3*01	TRBJ2-2*01	HomoSapiens
CAIQYANTGELFF	TRBV15*01	TRBJ2-2*01	HomoSapiens
CAIRTSGMLNTGELFF	TRBV2*01	TRBJ2-2*01	HomoSapiens

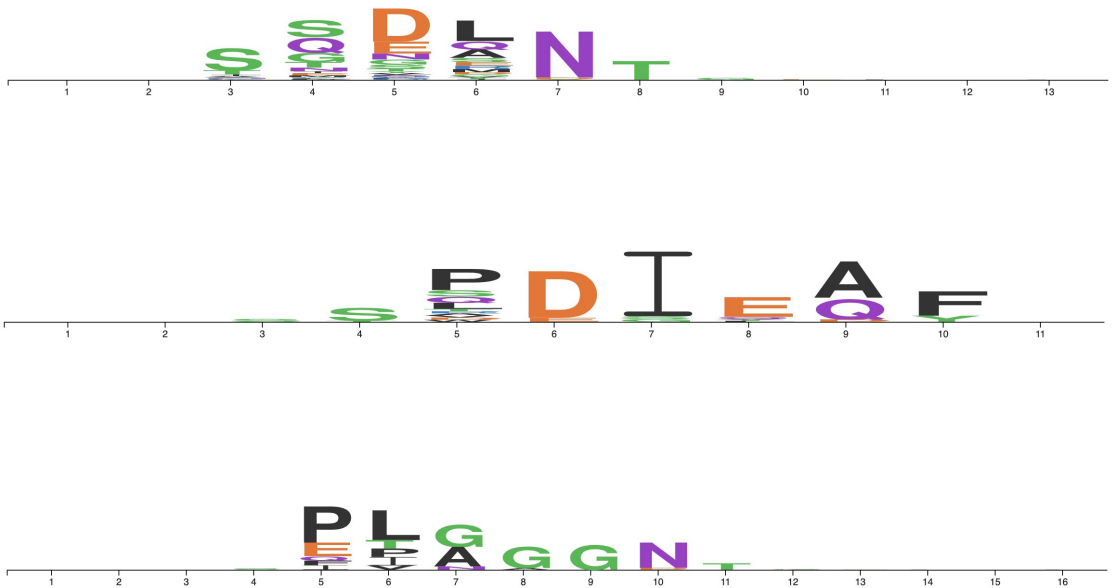
discover motifs and remove genetic background →



Motifs and data from VDJdb (Bagaev et al. 2020)

Sequencing technologies provide data which can be examined for biological properties

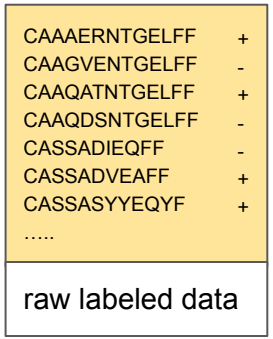
- One way to approach an analysis: make a position weight matrix showing product multinomial distribution of amino acids
- But what if we want to predict if a sequence is specific to a virus or not?



Motifs and data from VDJdb (Bagaev et al. 2020)

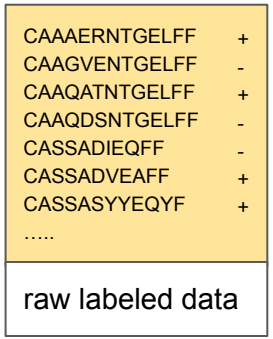
Machine learning is a powerful approach to discovering patterns in (biological) data

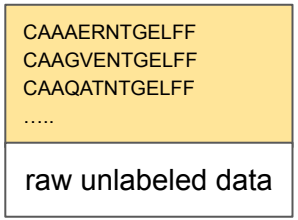
- A set of methods that allow for making inferences about the data
- Example: will the receptor bind to the virus or not? - we can fit a logistic regression model on receptor data and then predict binding for new receptors



Machine learning is a powerful approach to discovering patterns in (biological) data

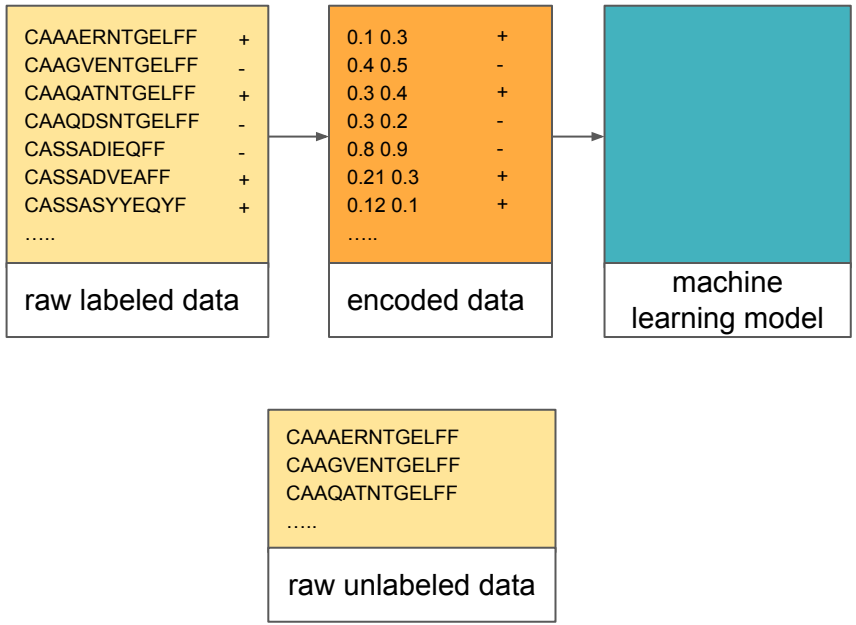
- A set of methods that allow for making inferences about the data
- Example: will the receptor bind to the virus or not? - we can fit a logistic regression model on receptor data and then predict binding for new receptors





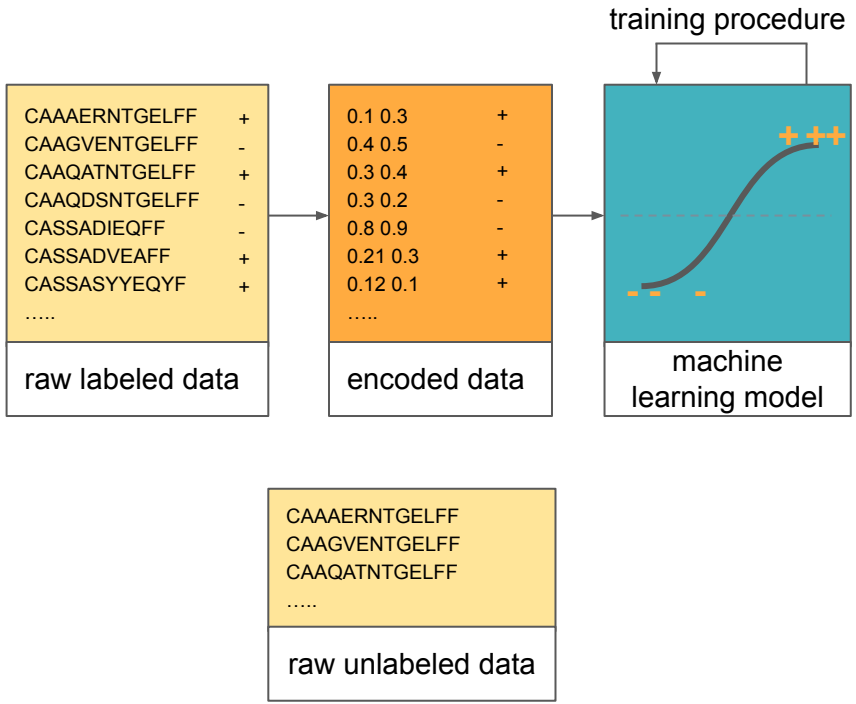
Machine learning is a powerful approach to discovering patterns in (biological) data

- A set of methods that allow for making inferences about the data
- Example: will the receptor bind to the virus or not? - we can fit a logistic regression model on receptor data and then predict binding for new receptors



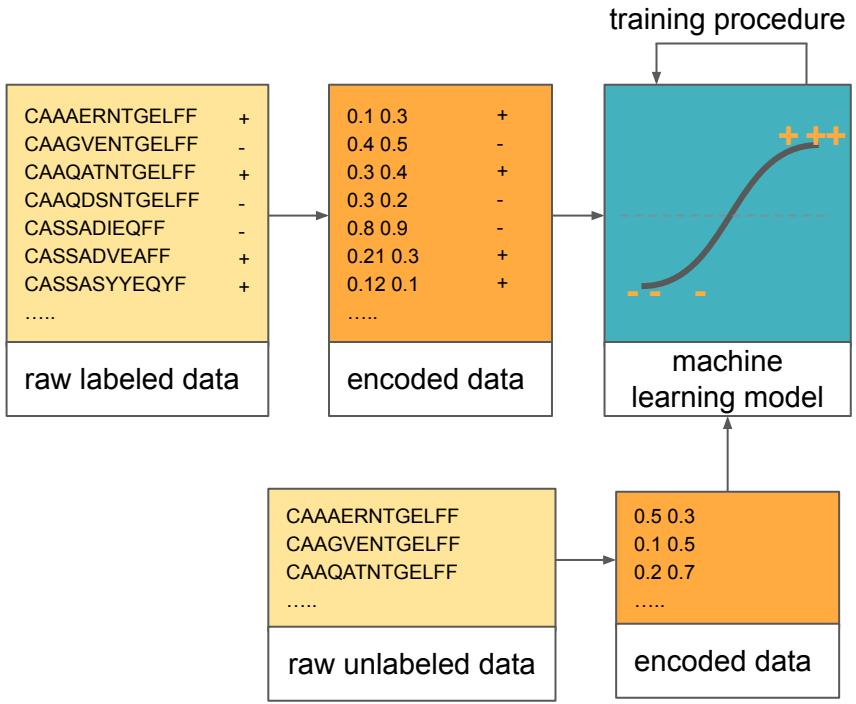
Machine learning is a powerful approach to discovering patterns in (biological) data

- A set of methods that allow for making inferences about the data
- Example: will the receptor bind to the virus or not? - we can fit a logistic regression model on receptor data and then predict binding for new receptors



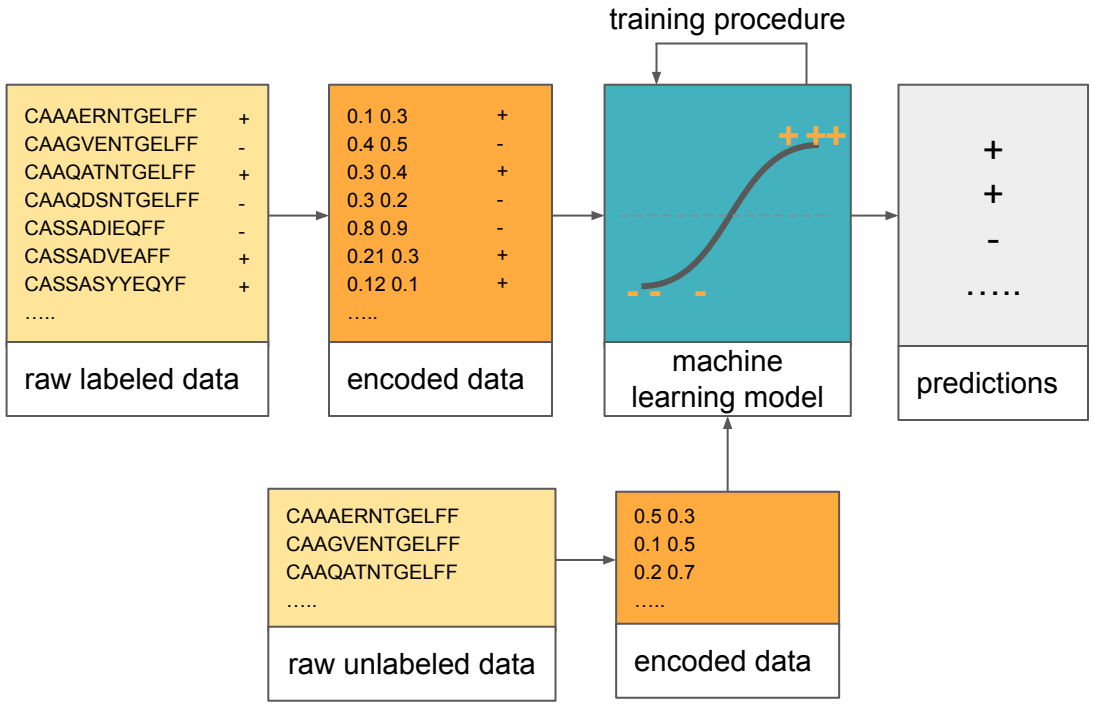
Machine learning is a powerful approach to discovering patterns in (biological) data

- A set of methods that allow for making inferences about the data
- Example: will the receptor bind to the virus or not? - we can fit a logistic regression model on receptor data and then predict binding for new receptors

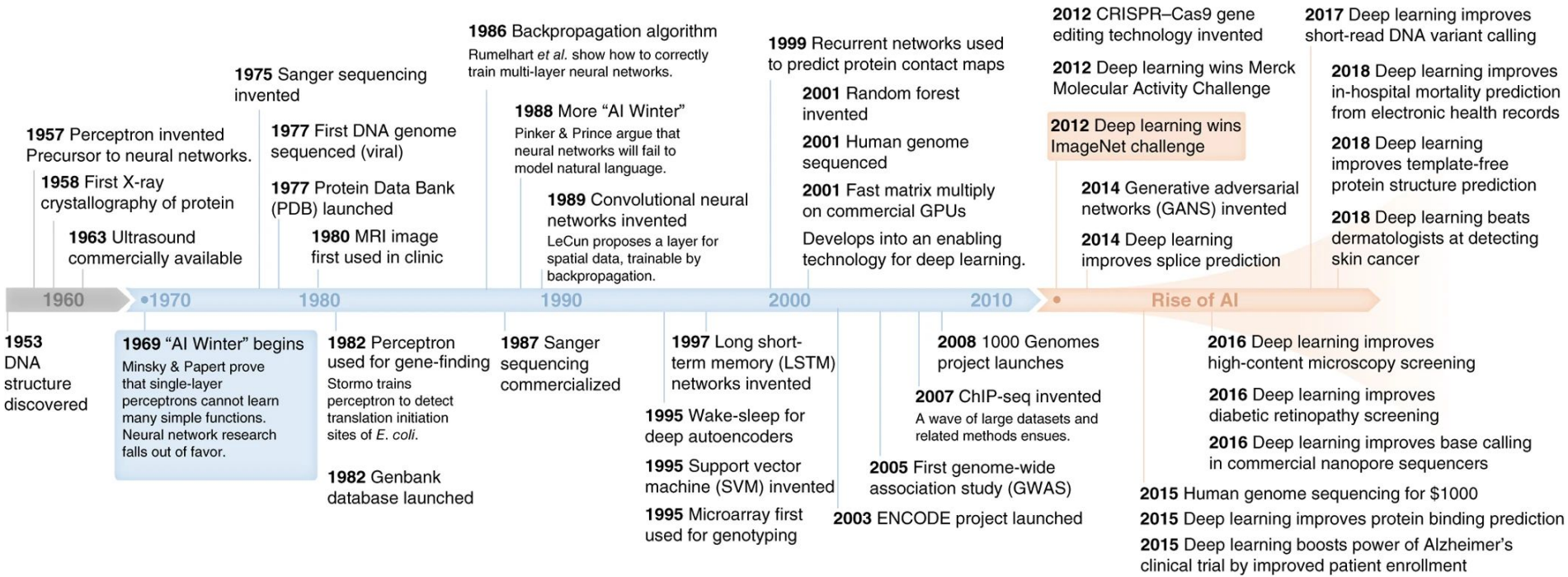


Machine learning is a powerful approach to discovering patterns in (biological) data

- A set of methods that allow for making inferences about the data
- Example: will the receptor bind to the virus or not? - we can fit a logistic regression model on receptor data and then predict binding for new receptors

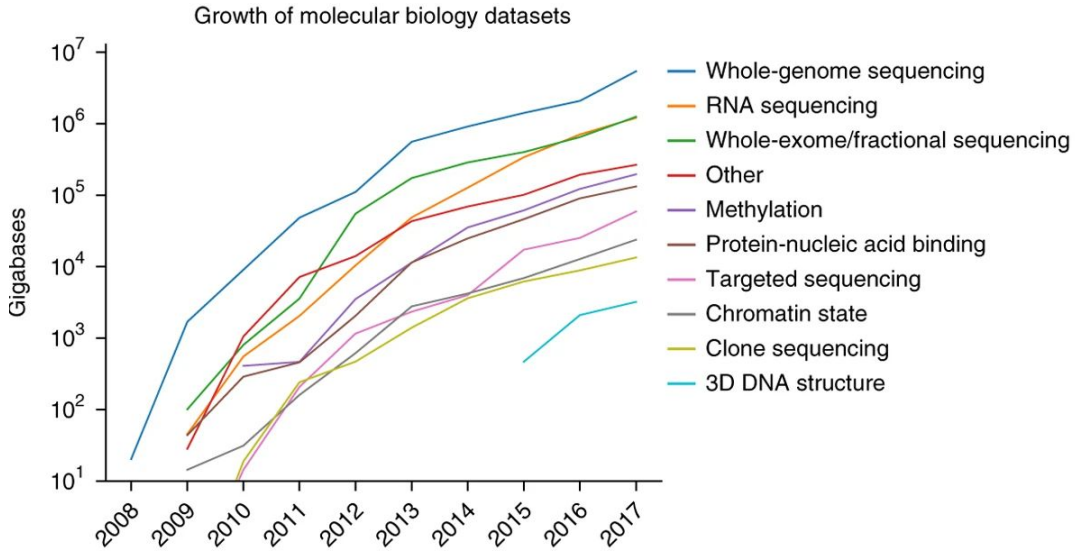


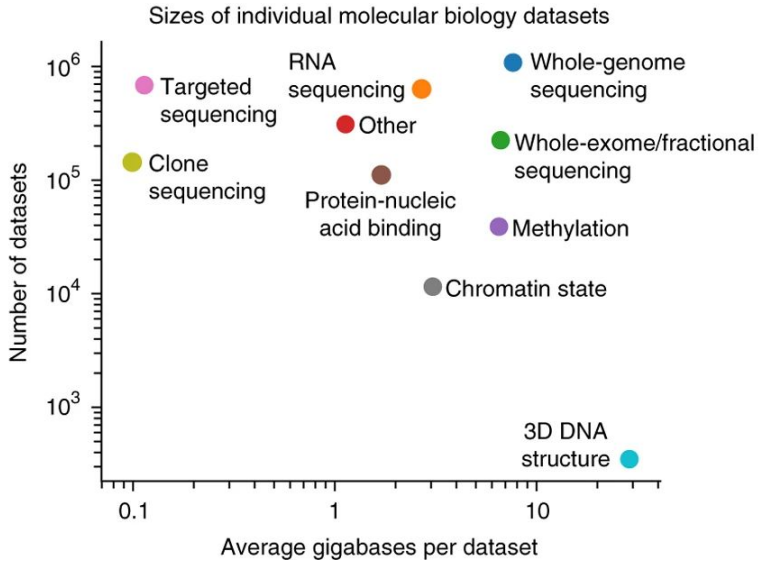
ML and computational biology development timeline



Wainberg et al. 2018

Data availability increased significantly in the recent years





Wainberg et al. 2018

A variety of research questions in computational biology can be tackled with machine learning

- Machine learning can be used for analysis of genomic data:
  - transcription factor binding
  - antigen binding
  - translation initiation site discovery
  - splicing prediction
  - single-cell RNA-seq clustering
- And also for obtaining genomic data:
  - variant calling

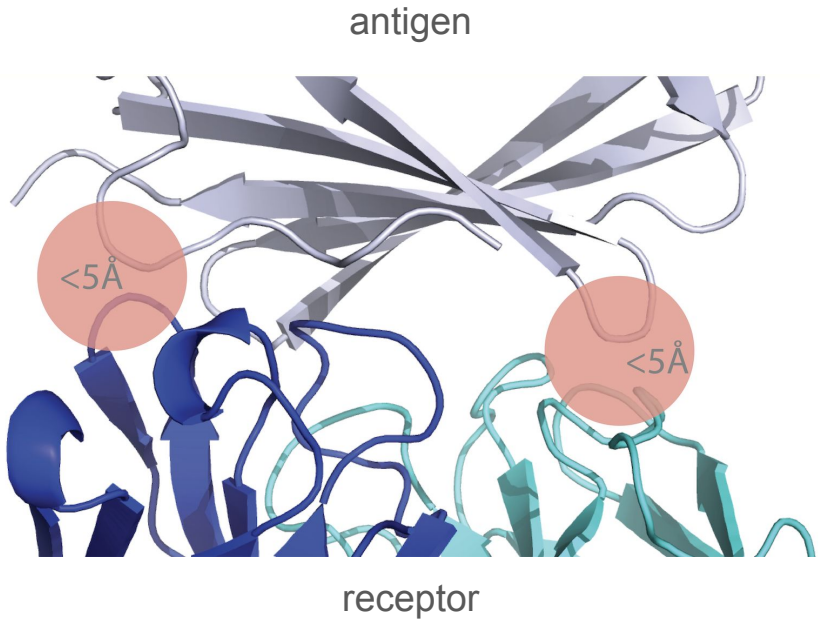
A variety of research questions in computational biology can be tackled with machine learning

Example: antigen binding prediction

Immune receptors (proteins) bind to antigens (e.g., parts of viruses) to help eliminate them

Given a set of receptors known to bind a given antigen, can we predict for the new receptor if it will bind to the antigen?

Classification problem!



Akbar et al. 2019

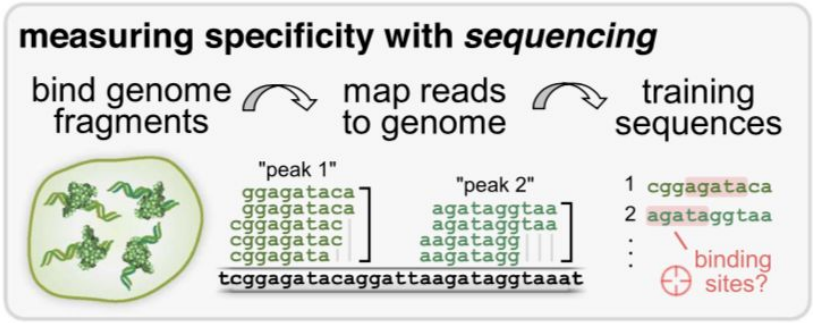
A variety of research questions in computational biology can be tackled with machine learning

Transcription factor binding prediction:

Transcription factors are proteins which bind to certain sites in DNA and regulate transcription of genes

Given a set of DNA sequences for which we know if they will bind or not, how can we predict if a transcription factor will bind to a new DNA sequence?

Classification problem!



Leung et al. 2016

Published: 27 July 2015

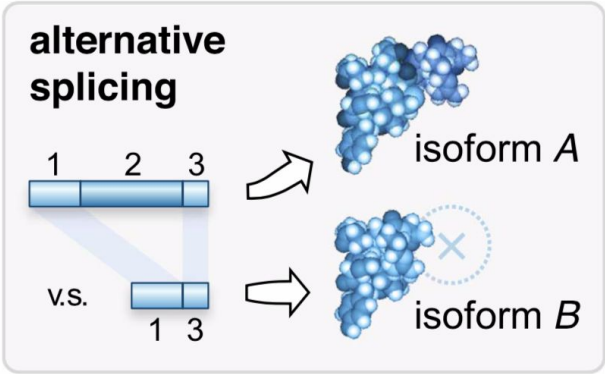
Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

Babak Alipanahi, Andrew Delong, Matthew T Weirauch & Brendan J Frey

Nature Biotechnology **33**, 831–838(2015) | Cite this article

A variety of research questions in computational biology can be tackled with machine learning

- Splicing: processing of precursor RNA that creates messenger RNA by removing non-coding regions (introns) and connects gene-coding regions (exons) together
- Predicting how often exons will be included in the transcripts is a regression problem



Leung et al. 2016

RESEARCH ARTICLE

The human splicing code reveals new insights into the genetic determinants of disease

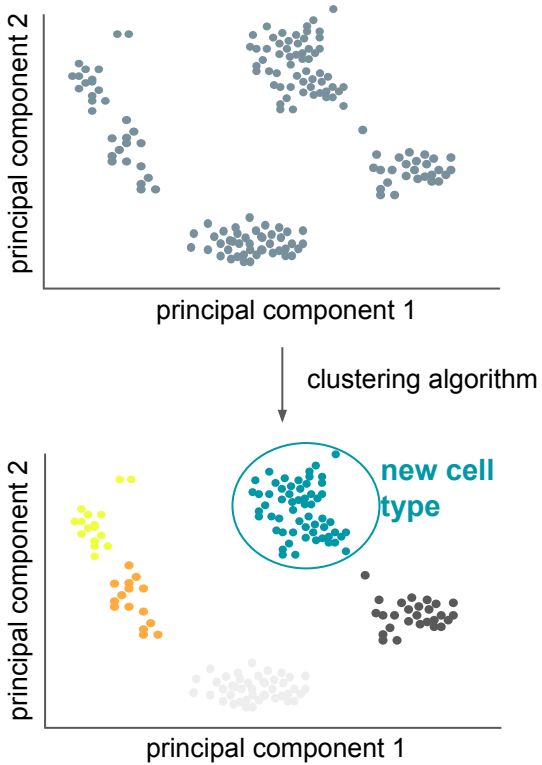
Hui Y. Xiong[1,2,3,*], Babak Alipanahi[1,2,3,*], Leo J. Lee[1,2,3,*], Hannes Bretschneider[1,3,4], Daniele Merico[5,6,7], Ryan K. C. Yuen[5,6,7], Yimin Hua[8], Serge Gueroussov[2,7], Hamed S. Najafabadi[1,2,3], Timothy R. Hughes[2,3,7], Quaid Morris[1,2,3,7], Yoseph Barash[1,2,9], Adrian R. Krainer[8], Nebojsa Jojic[10], Stephen W. Scherer[3,5,6,7], Benjamin J. Blencowe[2,5,7], Brendan J. Frey[1,2,3,4,5,7,10,†]

A variety of research questions in computational biology can be tackled with machine learning

Single-cell RNA-seq clustering for identification of cell types:

A dimensionality reduction technique is applied to normalized count data

Clustering the data (using e.g., k-means algorithm) can reveal new cell types



Review Article | Published: 07 January 2019

Challenges in unsupervised clustering of single-cell RNA-seq data

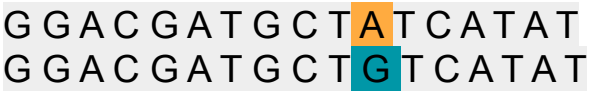
Vladimir Yu Kiselev, Tallulah S. Andrews & Martin Hemberg ✉

Nature Reviews Genetics **20**, 273–282(2019) | Cite this article

A variety of research questions in computational biology can be tackled with machine learning

Variant calling: finding genetic variants from sequence reads

From a pileup of the reference and read data around each candidate variant, ML models could determine the probabilities for each of the three diploid genotypes



Published: 24 September 2018

A universal SNP and small-indel variant caller using deep neural networks

Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, Sam S Gross, Lizzie Dorfman, Cory Y McLean & Mark A DePristo

Nature Biotechnology **36**, 983–987(2018) | Cite this article

Article | Open Access | Published: 01 March 2019

A multi-task convolutional deep neural network for variant calling in single molecule sequencing

Ruibang Luo, Fritz J. Sedlazeck, Tak-Wah Lam & Michael C. Schatz

Nature Communications **10**, Article number: 998 (2019) | Cite this article

A variety of research questions in computational biology can be tackled with machine learning

DeepGOPlus: improved protein function prediction from sequence
Maxat Kulmanov, Robert Hoehndorf ✉

Bioinformatics, Volume 36, Issue 2, 15 January 2020, Pages 422–429, https://doi.org/10.1093/bioinformatics/btz595

Sequence alignment using machine learning for accurate template-based protein structure prediction
Shuichiro Makigaki ✉, Takashi Ishida ✉

Bioinformatics, Volume 36, Issue 1, 1 January 2020, Pages 104–111, https://doi.org/10.1093/bioinformatics/btz483

Learned protein embeddings for machine learning FREE
Kevin K Yang, Zachary Wu, Claire N Bedbrook, Frances H Arnold ✉

Bioinformatics, Volume 34, Issue 15, 01 August 2018, Pages 2642–2648, https://doi.org/10.1093/bioinformatics/bty178

Graph neural representational learning of RNA secondary structures for predicting RNA-protein interactions
Zichao Yan, William L Hamilton ✉, Mathieu Blanchette ✉

Bioinformatics, Volume 36, Issue Supplement_1, July 2020, Pages i276–i284, https://doi.org/10.1093/bioinformatics/btaa456

Article | Published: 06 July 2020

Deep learning decodes the principles of differential gene expression

Shinya Tasaki ✉, Chris Gaiteri, Sara Mostafavi & Yanling Wang

Nature Machine Intelligence **2**, 376–386(2020) | Cite this article

Article | Open Access | Published: 11 May 2020

Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis

Xiangjie Li, Kui Wang, Yafei Lyu, Huize Pan, Jingxiao Zhang, Dwight Stambolian, Katalin Susztak, Muredach P. Reilly, Gang Hu ✉ & Mingyao Li ✉

Nature Communications **11**, Article number: 2338 (2020) | Cite this article

QDeep: distance-based protein model quality estimation by residue-level ensemble error classifications using stacked deep residual neural networks
Md Hossain Shuvo, Sutanu Bhattacharya ✉, Debswapna Bhattacharya ✉

Bioinformatics, Volume 36, Issue Supplement_1, July 2020, Pages i285–i291, https://doi.org/10.1093/bioinformatics/btaa455

Computational biology poses unique challenges for machine learning

- ❏ Dimensionality & dataset size
- ❏ Signal and noise in the data
- ❏ Unknown ground truth and weakly labeled datasets
- ❏ Selection bias

Keep in the data generation process in mind!

Machine learning in computational biology - outline

- Introduction to machine learning:
  - What is machine learning, types of problems, assumptions, workflow, generalization
- Machine learning models and algorithms:
  - Discriminative vs generative models, supervised models (logistic and linear regression, kNN, neural networks), unsupervised models (dimensionality reduction, clustering)
- Data representation:
  - Considerations and examples, one-hot encoding, feature engineering, representation learning
- Model comparison and uncertainty:
  - Model assessment, model selection, uncertainty, cross-validation
- Transparency and reproducibility

Machine learning in computational biology - outline

- **Introduction to machine learning:**
  - **What is machine learning, types of problems, assumptions, workflow, generalization**
- Machine learning models and algorithms:
  - Discriminative vs generative models, supervised models (logistic and linear regression, kNN, neural networks), unsupervised models (dimensionality reduction, clustering)
- Data representation:
  - Considerations and examples, one-hot encoding, feature engineering, representation learning
- Model comparison and uncertainty:
  - Model assessment, model selection, uncertainty, cross-validation
- Transparency and reproducibility

References

Wainberg M, Merico D, Delong A, Frey BJ. Deep learning in biomedicine. *Nature Biotechnology*. 2018;36(9):829-838. doi:10.1038/nbt.4233

Bagaev DV, Vroomans RMA, Samir J, et al. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res*. 2020;48(D1):D1057-D1062. doi:10.1093/nar/gkz874

Akbar R, Robert PA, Pavlović M, et al. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *bioRxiv*. Published online November 30, 2019:759498. doi:10.1101/759498

Leung MKK, Delong A, Alipanahi B, Frey BJ. Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. *Proceedings of the IEEE*. 2016;104(1):176-197. doi:10.1109/JPROC.2015.2494198