Seminar: Deep Learning for Molecular Biology

Alice McHardy, Giorgos Kallergis, Ehsannedin Asgari, Mohammad Hadi Foroughmand Araabi

Helmholtz Center for Infection Research & TU Braunschweig
Spring 2024

Seminar overview

Max. number of participants: 10

Language: English

Requirements:

- >5 page summary of the topic, with literature references (to be sent two weeks before seminar date)
- Student pairs with both practical (implementation in Python) and theoretical presentations
- 20 minutes of presentation, plus 10 discussion (Before July 13th, to be set via doodle)

Designated for Bachelor and Master students of Computer Science

Course Takeaways

Basic Knowledge of Machine Learning: Understanding of fundamental concepts in machine learning.

- Machine Learning:
  a. A hot topic in both scientific research and industry applications.
  b. Wide-ranging impact across various domains, from healthcare to finance.

Basic Knowledge of Deep Learning: Familiarity with the principles of deep learning.

- Deep Learning:
  - Specialized models built upon neural network architectures.
  - Remarkable success in tackling complex and challenging problems across domains.

Basic Knowledge of Bioinformatics:

- Understanding of bioinformatics principles and applications.
- Awareness of how machine learning and deep learning are applied in bioinformatics research.

Experience in Problem Solving:

- Practical experience in applying machine learning and deep learning techniques to bioinformatics problems.

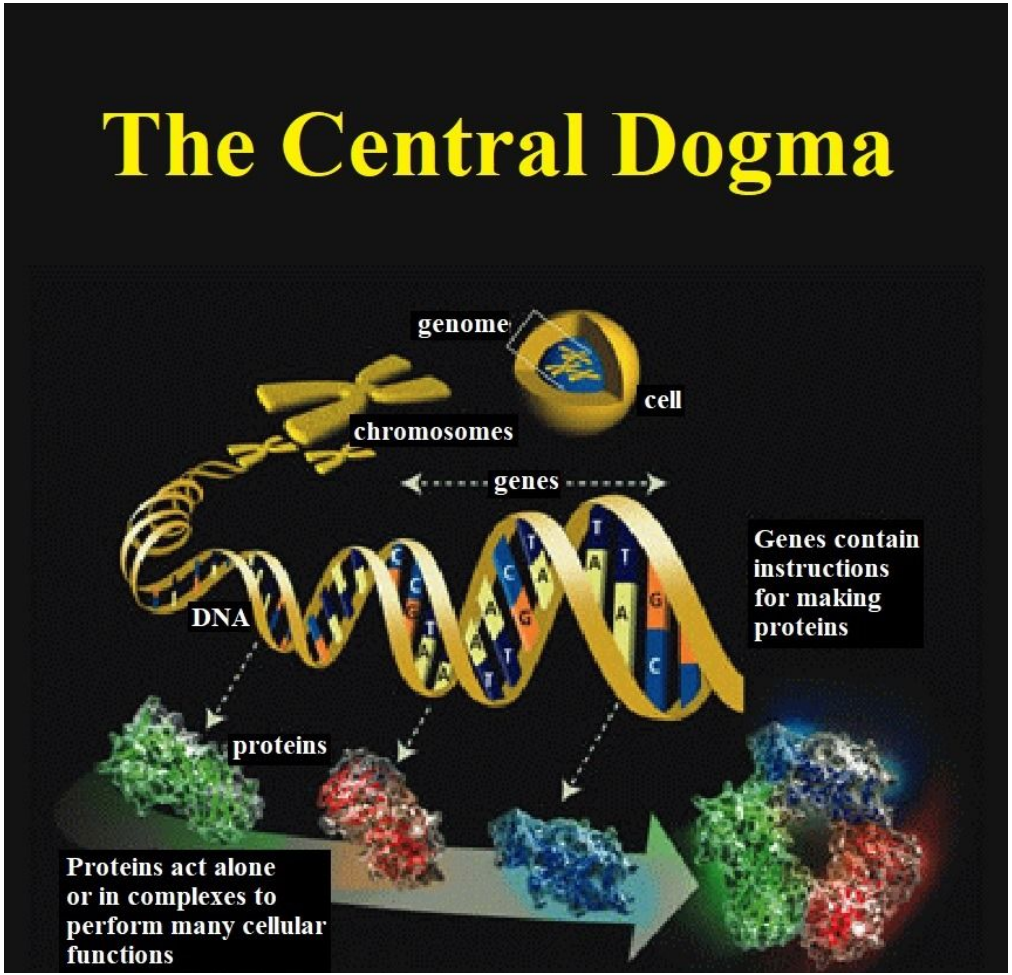
Introduction to Bioinformatics, Machine Learning, and Deep Learning

What is bioinformatics

(Current) Biology =

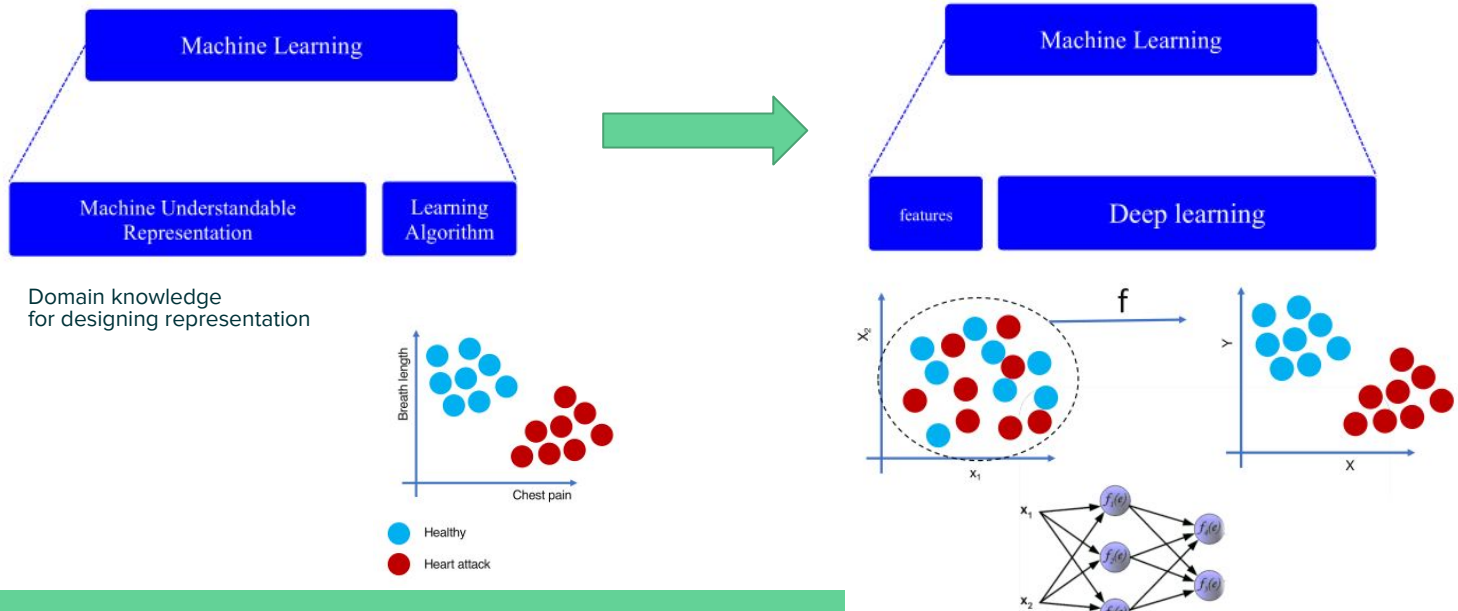
- DNA + RNA + Protein + Interactions

Bioinformatics: Computational analysis of the biological data



What is machine learning

- We alter data appearance to be interpretable by the audience.
  - Machine as the audience? Numerical values, vectors, matrices
- Finding a proper representation has been critical in machine learning



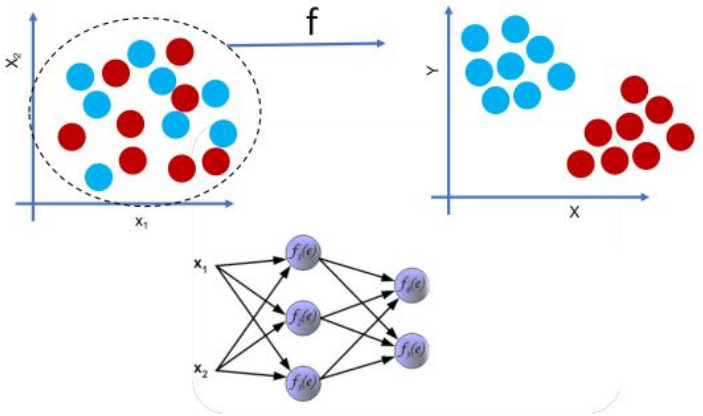
Machine Learning

Given an observed sample set, finding inherent structure of the data, in order to

- Understand what is happening there
- Predict some unknown features

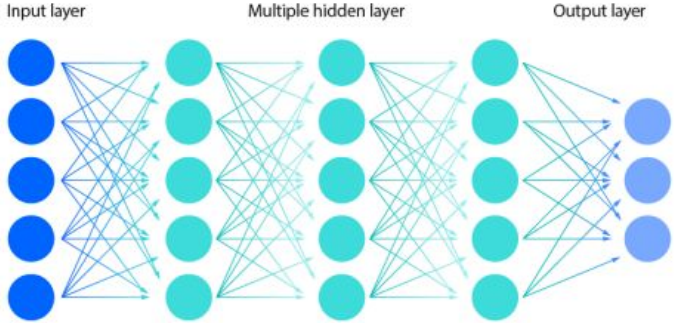
Representation learning:

- Data, could be represented as points in n-dimensional space (of features)
- Finding a transformation separating different data samples?
    - Then, we can solve several problems, e.g. classification.

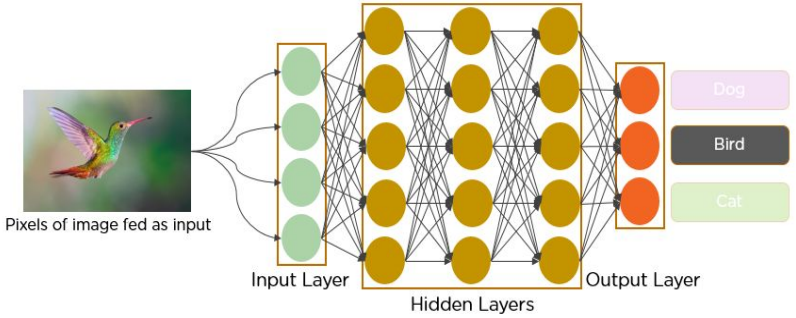


Specific ML models: NN, RNN, CNN

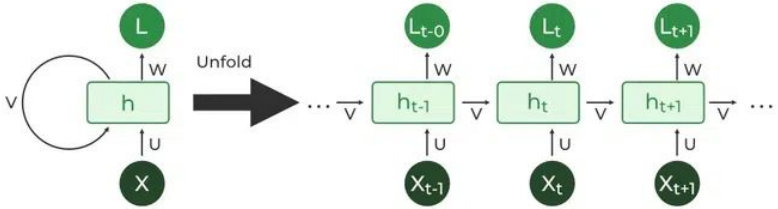
Neural network (NN): Inspired by brain



Convolutional Neural Network (CNN): NN with fewer (redundant) parameters



Recurrent Neural Network (RNN): Dealing with sequential information



What is a transformer?

- An architecture with breakthrough performances
- Encoder/Decoder architecture
- Used in a wide range of applications
- Effective
- Highly parallelizable
- Position encoding
- Ideal for transfer learning

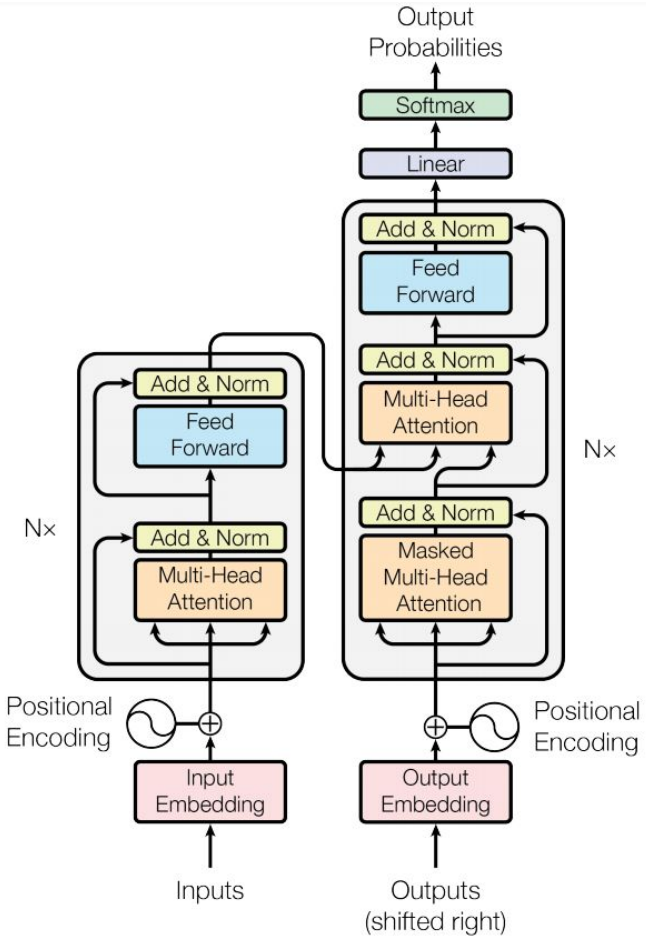
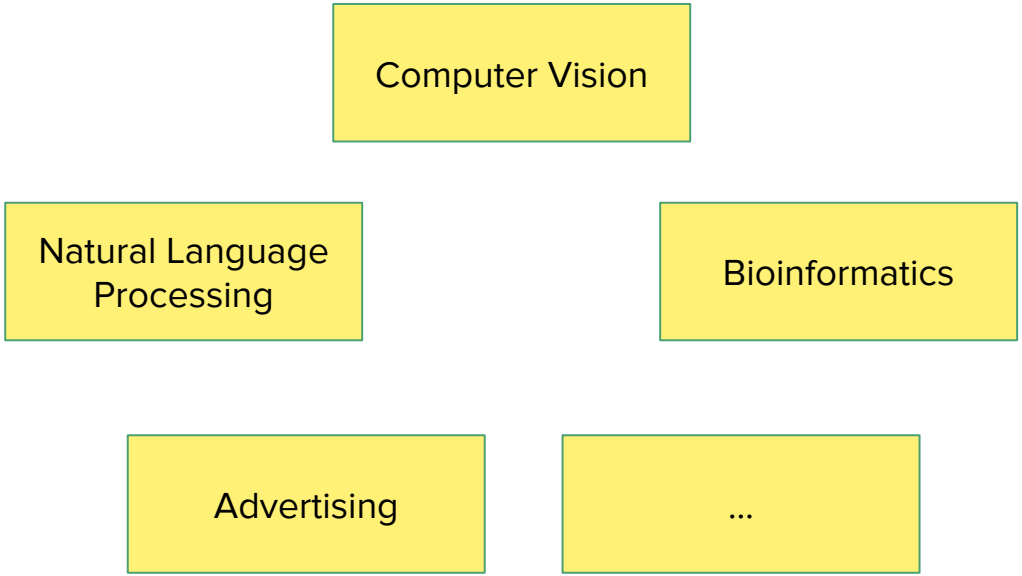


Figure 9. Transformers model architecture.

Vaswani et al, Attention is all you need. (NIPS 2017)

Applications of deep learning



Your task

- Make groups
- Chooses a topic (from the following list)
- Study the topic
  - Useful material is provided
  - Meeting and consultation with the lecturers (at least once)
- Implement and test your topic with the dataset (if applicable)
  - Evaluate (Metrics, precision/recall, TP, TN, ...)
- Create a written report
- Present your topic in the presentation day

Implementation Tasks

Frameworks: Pytorch, TensorFlow, Keras, etc.

Dataset: Toxic protein sequence prediction.

- Gacesa, Ranko, David J. Barlow, and Paul F. Long. "Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions." *PeerJ Computer Science* 2 (2016): e90.

Git repository: https://github.com/hzi-bifo/seminar-dlmb-2024

- Includes materials for the course

Dataset

- Proteins are biomolecules consist of sequences of amino acids e.g., “MESMFEDDISILTQEA…”
- Animal venoms comprise predominantly toxic peptides and proteins.
- Which ones are toxins?

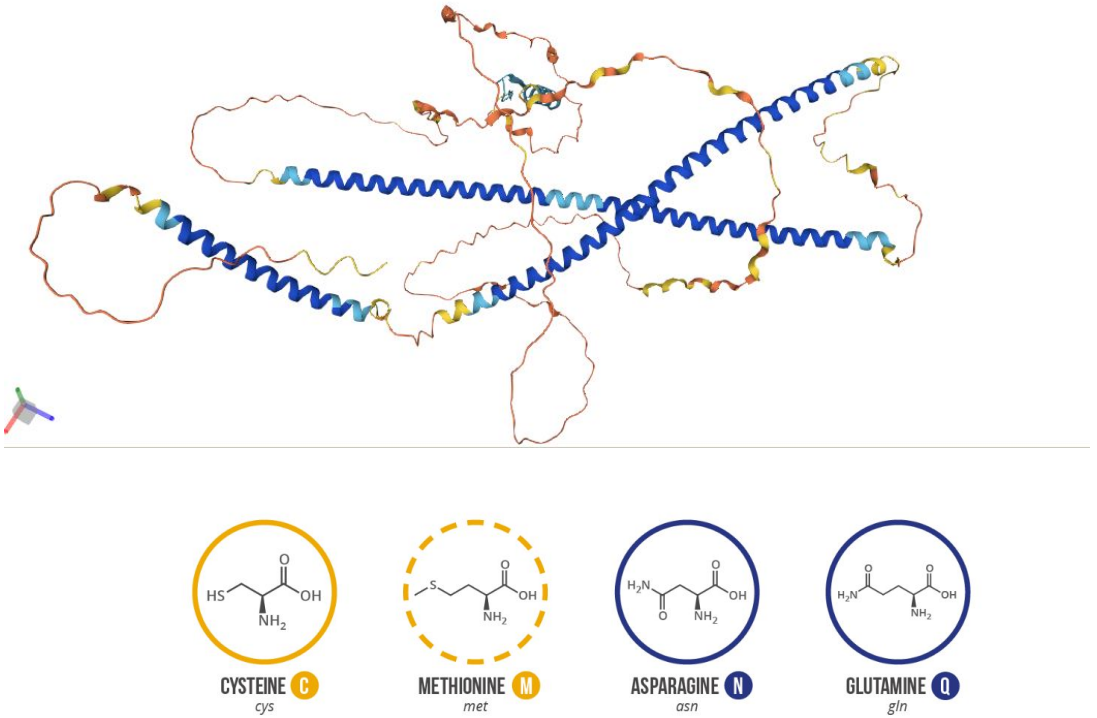


Figure 3. 3D structure of TANK-binding kinase 1-binding protein 1 and examples of amino acids

Topics

Topics

- Feed-forward Neural Networks and back propagation [2 students]
- Convolutional Neural Networks (CNN) [2 students]
- Recurrent Neural Networks and LSTMs [2 students]
- Representation learning - Encoders [2 students]
- Representation learning - Decoders, Encoder-Decoders [2 students]

Feed-forward Neural Networks and back propagation [2 students]

Goals: Getting familiar with the basics of neural networks

- Introduction to linear classification, multilayer perceptron (MLP), and back propagation algorithm
- Implementation in python: using MLP for toxin prediction

Suggested references

- Linear prediction: Lec. 1,2 at (https://bit.ly/1DIpc51) and (Chapter 4: https://stanford.io/2voWjra)
- G. Hinton's lecture 2, 3: https://bit.ly/3TNBPqw
- Lecture from U of Waterloo: https://bit.ly/2A2mzgN
- Stanford Tutorial: https://stanford.io/1FRrkZw
- Practicals: In Keras (keras.io) or Pytorch (pytorch.org)

Convolutional Neural Networks (CNN) [2 students]

Goals: Getting familiar with the convolutional neural network

- CNN
- Implementation in python: using CNN for toxin prediction

Suggested references

- MIT notes: https://tinyurl.com/3c8fk4mz
- G. Hinton's lecture: https://bit.ly/3TNBP qw
- Stanford Tutorial: https://stanford.io/1FRrkZw
- A more advanced reference: deeplearningbook.org
- Practicals: In Keras (keras.io) or Pytorch (pytorch.org), e.g. https://tinyurl.com/yfy56ay5

Recurrent Neural Networks [2 students]

Goals: Getting familiar with the RNNs and in particular LSTM

- Understanding “Vanilla” RNN
- Understanding the LSTM architecture (in particular read: https://bit.ly/1S6gmjZ)
- Implementation in python: using RNNs and LSTM for toxin prediction

Suggested references

- Lecture from U of Waterloo: https://bit.ly/2RCNEhn
- Understanding LSTMs: https://bit.ly/1S6gmjZ
- MIT notes: https://tinyurl.com/2x4z77fz
- G. Hinton’s lecture: https://bit.ly/3TNBPqw
- A more advanced reference: deeplearningbook.org
- Practicals: In Keras (keras.io) or Pytorch (pytorch.org)

Representation learning - Encoders [2 students]

Goals: Getting familiar with the concept of representation learning

- Encoder (e.g. Bert)
- Implementation in python: using Encoders for toxin prediction

Suggested references

- Representation learning: https://arxiv.org/pdf/1206.5538.pdf
- ProtVec paper: https://bit.ly/2REy8l9 and some slides (https://bit.ly/2ytOIeB)
- Transformer paper: https://arxiv.org/abs/1706.03762
- Simple explanation of transformers: https://jalammar.github.io/illustrated-transformer/
- Simple explanation of the details: https://serrano.academy/
- Practicals: In Keras (keras.io) or Pytorch (pytorch.org)
- Practicals: https://huggingface.co/docs/transformers/notebooks
- ProtBert https://huggingface.co/Rostlab/prot_bert, ESM (https://github.com/facebookresearch/esm)

Representation learning - Decoder-only and Encoder - Decoders [2 students]

Goals: Getting familiar with the concept of representation learning

- Decoder-only models (GPT)
- Encoder-decoder models
- Their applications on bioinformatics

Suggested references

- Representation learning: https://arxiv.org/pdf/1206.5538.pdf
- ProtVec paper: https://bit.ly/2REy8l9 and some slides (https://bit.ly/2ytOIeB)
- Transformer paper: https://arxiv.org/abs/1706.03762
- Simple explanation of transformers: https://jalammar.github.io/illustrated-transformer/
- Simple explanation of the details: https://serrano.academy/
- Practicals: In Keras (keras.io) or Pytorch (pytorch.org)
- Practicals: https://huggingface.co/docs/transformers/notebooks
- A survey of protein transformers: https://doi.org/10.1145/3388440.3412467, a survey on transformers in bioinformatics (https://doi.org/10.1093/bioadv/vbad001)

Further steps:

Send an email (to both of us):

- From one member
- CC all other members
- Send three preferred topics in order of preference
- Until 28th of April 2024

Final seminar date:

- Meeting and consultation with the lecturers (at least once)
- 11 July, 9:00 - 13:00 at BRICS.

Any question? Contact us:

- mohammad-hadi.foroughmand-araabi@helmholtz-hzi.de
- georgios.kallergis@helmholtz-hzi.de

The End

Protein models

ESM, MSA Transformer (tutorial for train, huggingface) (encoder?)
Unified rational protein engineering with sequence-only deep representation learning (LSTM), github
ProGen (decoder?) (github, huggingface?)
proteinBERT (like BERT, but different, somehow encoder)
ProtBert https://huggingface.co/Rostlab/prot_bert (Rostlab has also T5 models)
Yijia-Xiao/ProteinLM (huggingface, github)
TAPE-Transformer
ProtTrans
UDSMProt
UniRep
Sturmfels P. et al. (2020) Profile prediction: an alignment-based pre-training task for protein sequence models. arXiv Prepr arXiv201200195.
Yang J. et al. (2020) Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. USA, 117, 1496–1503.
Nambiar A. et al. (2020) Transforming the language of life: transformer neural networks for protein prediction tasks. In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. pp. 1–8.
DistilProtBert
ProLanGO2 (encoder-decoder)

References

- 3d structure of protein: https://www.uniprot.org/uniprotkb/A7MCY6/entry
- Amino acids: https://www.compoundchem.com/2014/09/16/aminoacids/
- Dataset comes from: https://peerj.com/articles/cs-90/