

# MACHINE LEARNING IN BIOINFORMATICS

## INTRODUCTION

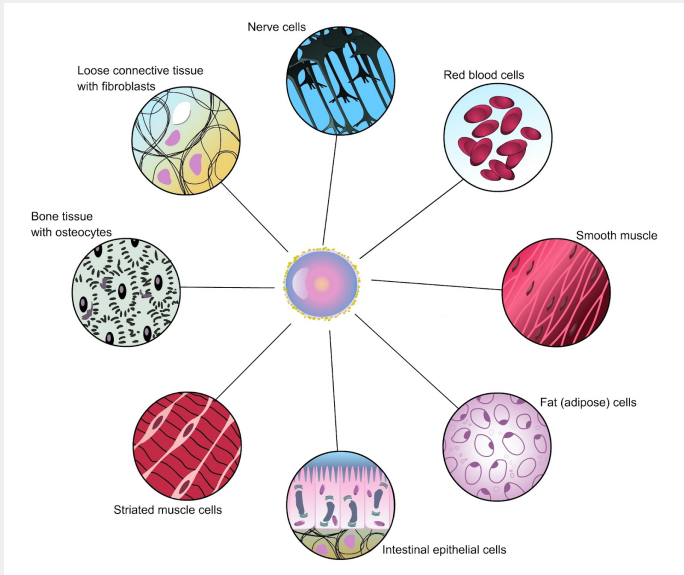
Philipp Benner  
*philipp.benner@bam.de*

VP.1 - eScience  
Federal Institute of Materials Research and Testing (BAM)

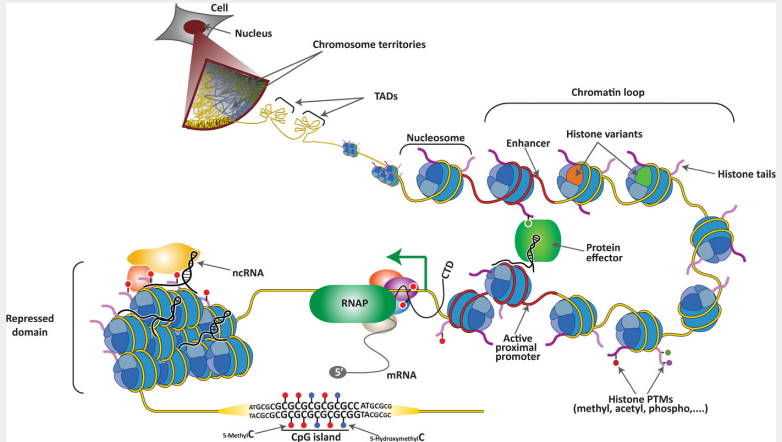
February 8, 2026

# **BIOLOGICAL BACKGROUND**

# BIOLOGICAL BACKGROUND

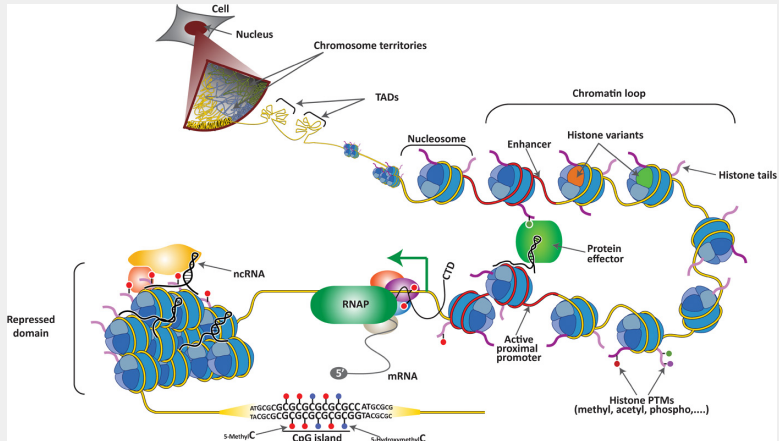


## BIOLOGICAL BACKGROUND



[Aranda et al., 2015]

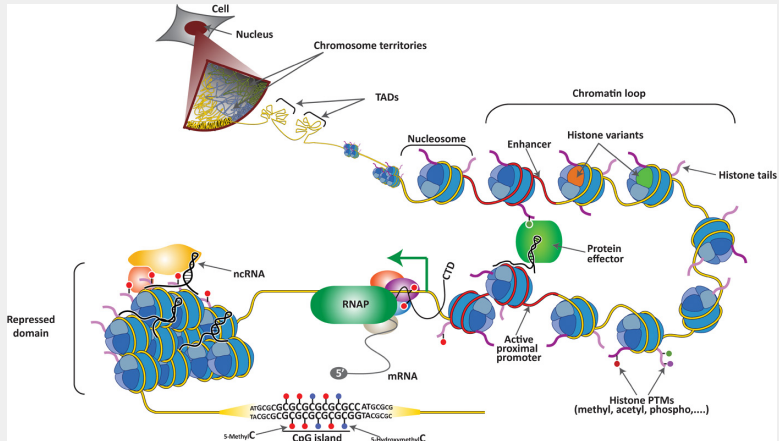
# BIOLOGICAL BACKGROUND



[Aranda et al., 2015]

Khorana, Holley and Nirenberg (1953-1965): Discovery of the Genetic Code

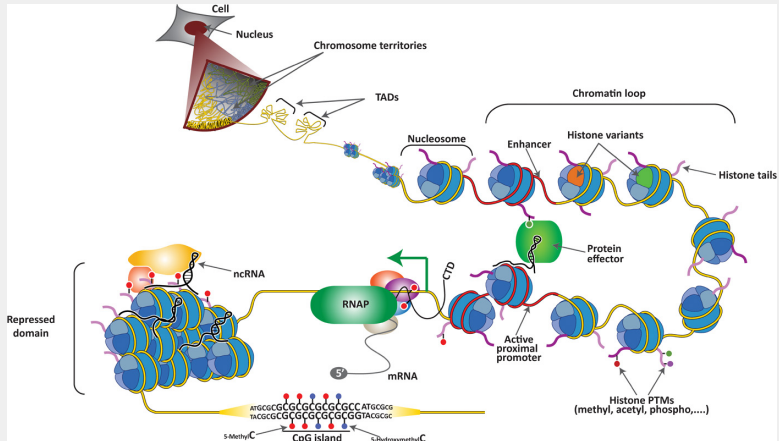
# BIOLOGICAL BACKGROUND



[Aranda et al., 2015]

Human Genome Project (1990-2003): Identify DNA sequence (3 billion basepairs)

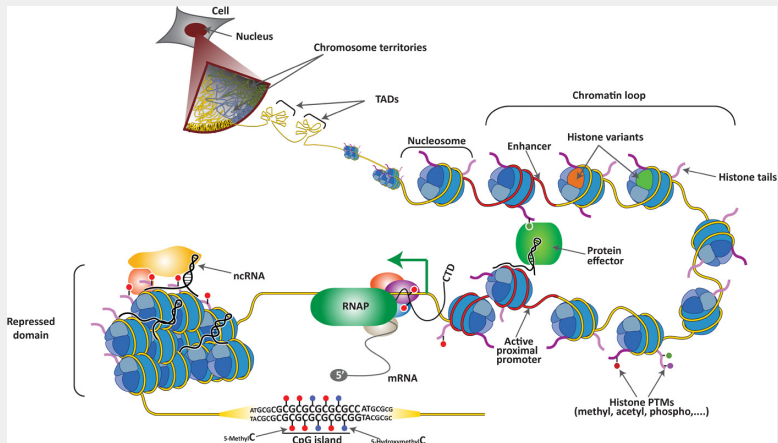
## BIOLOGICAL BACKGROUND



[Aranda et al., 2015]

GENCODE Project (since 2003): Identify location of genes (20,000 protein coding genes)

# BIOLOGICAL BACKGROUND

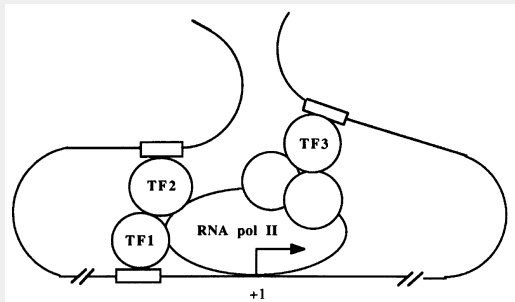


[Aranda et al., 2015]

ENCODE Project (since 2003): Identify cell type-specific epigenetic marks

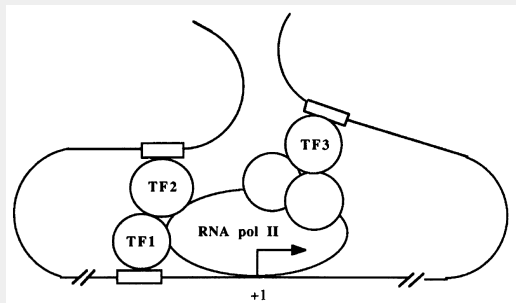


# GENE EXPRESSION REGULATION



[Mitchell and Tjian, 1989]

# GENE EXPRESSION REGULATION



[Mitchell and Tjian, 1989]

- Gene expression is regulated by promoters and enhancers
- Enhancer activity is highly cell type-specific
- Activation through transcription factors

# OBJECTIVES

If we knew...

- all transcription factors
- their binding preferences
- and interactions
- all promoters
- all enhancers and their targets

we should be able to predict cell type-specific gene expression

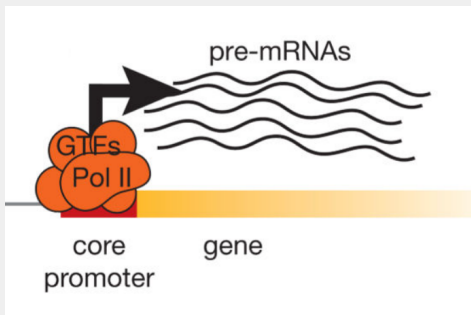
# APPLICATION OF MACHINE LEARNING IN UNDERSTANDING GENE REGULATION

- Study how the expression of genes is regulated
- How is information about gene expression encoded in the DNA?
- How do promoters control gene expression?
- What is the role of enhancers?
- When do enhancers get active?
- How do enhancers link to promoters?

# **APPLICATIONS OF ML**

# APPLICATION 1: PROMOTER ACTIVITY (REGRESSION)

How much control do promoters have over gene expression?



Approach: Develop machine learning method that predicts gene expression values from promoters

## APPLICATION 1: DATA

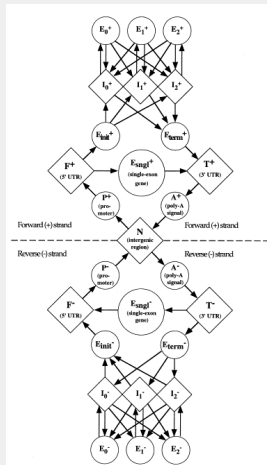
Expression values are cell type-specific. We only look at expression levels in liver:

Gene	Promoter sequence	Motif scores	Expression
1	CGTAGA...AGC	0.23,0.53,...,0.90	100
2	CTTGGA...CCC	0.03,0.87,...,0.93	328
...	...	...	...
N	GGACGA...AAT	0.69,0.21,...,0.43	0

Last column shows expression levels derived from total RNA-seq

# APPLICATION 1: HOW DO WE KNOW THE POSITION OF GENES?

- Option 1: RNA-seq experiments of all tissues
- Option 2: Predictions from DNA-sequence
  - Hidden Markov model for gene structure prediction

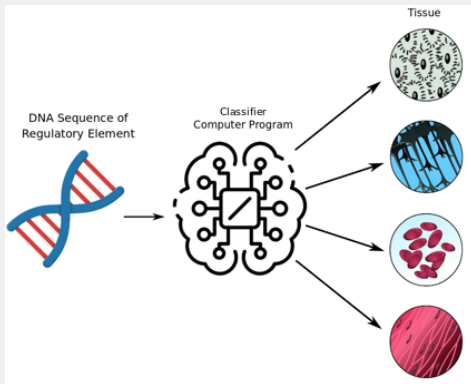


[Burge and Karlin, 1997]



## APPLICATION 2: ENHANCER ACTIVITY (CLASSIFICATION)

Understand to what extent and how cell type-specific enhancer activity is encoded in the DNA sequence



Approach: Develop machine learning method that identifies relevant patterns

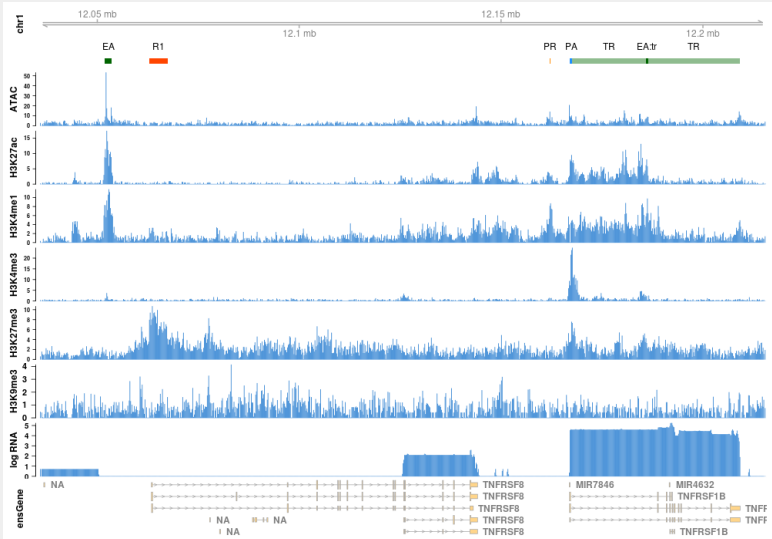
## APPLICATION 2: DATA

Enhancer activity is highly cell type-specific. We consider enhancers active in liver:

Enhancer	Sequence	Motif scores	Active
1	CGTAGA...AGC	0.23,0.53,...,0.90	1
2	CTTGGA...CCC	0.03,0.87,...,0.93	1
...	...	...	...
N	GGACGA...AAT	0.69,0.21,...,0.43	0

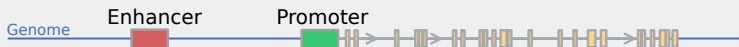
Last column encodes if an enhancer is active (1) or inactive (0) in liver

## IDENTIFICATION OF ENHANCERS



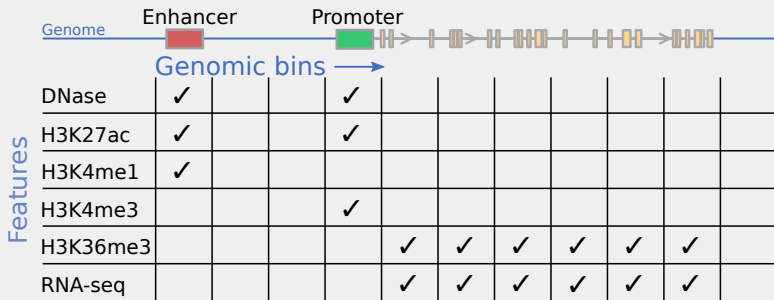
# IDENTIFICATION OF ENHANCERS

Enhancers are commonly identified from genome segmentations using hidden Markov models (HMMs):



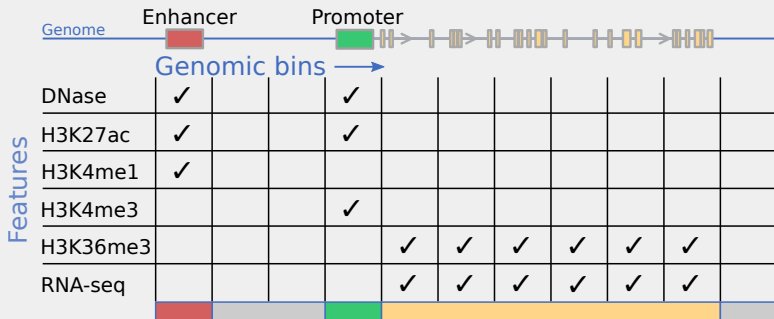
# IDENTIFICATION OF ENHANCERS

Enhancers are commonly identified from genome segmentations using hidden Markov models (HMMs):



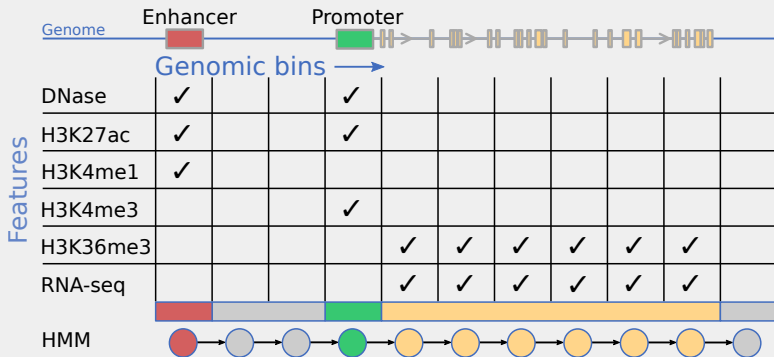
# IDENTIFICATION OF ENHANCERS

Enhancers are commonly identified from genome segmentations using hidden Markov models (HMMs):



# IDENTIFICATION OF ENHANCERS

Enhancers are commonly identified from genome segmentations using hidden Markov models (HMMs):



# ENCODE MOUSE EMBRYO DATA

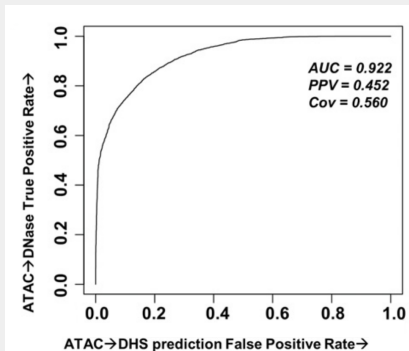
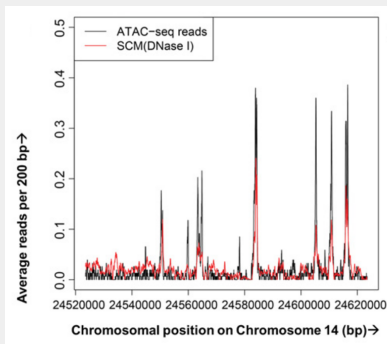
Enhancers are more difficult to identify. We need data from many different cell types:

	Forebrain	Midbrain	Hindbrain	Liver	Lung	Kidney	Heart	Limb
Day 11.5	✓	✓	✓	✓				
Day 12.5	✓	✓	✓	✓				
Day 13.5	✓	✓	✓	✓				
Day 14.5	✓	✓	✓	✓	✓	✓	✓	✓
Day 15.5	✓	✓	✓	✓	✓	✓	✓	✓
Day 16.5	✓	✓	✓	✓	✓	✓		

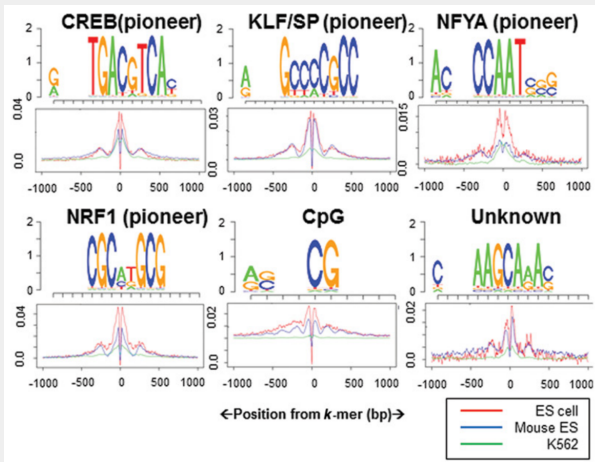


## **RELATED STUDIES**

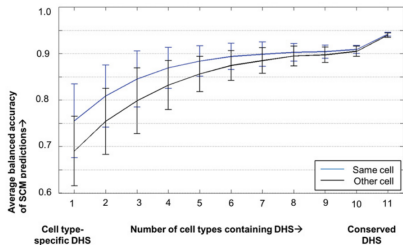
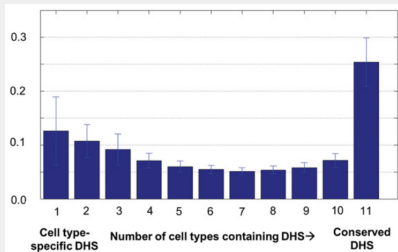
# PREDICTION OF ACCESSIBLE REGIONS FROM DNA SEQUENCE [HASHIMOTO ET AL., 2016]



# PREDICTION OF ACCESSIBLE REGIONS FROM DNA SEQUENCE [HASHIMOTO ET AL., 2016]



# PREDICTION OF ACCESSIBLE REGIONS FROM DNA SEQUENCE [HASHIMOTO ET AL., 2016]







# **SOFTWARE REQUIREMENTS**

# SOFTWARE REQUIREMENTS

- Python ( $\geq 3.7$ ) environment [Recommended: Anaconda]
- Python Packages:
  - ▶ Scikit-learn
  - ▶ Pandas
  - ▶ Numpy
  - ▶ PyTorch
- Editor for Jupyter Notebooks (e.g. VS Code)
- gkmSVM  
<https://www.beerlab.org/gkmsvm/>

# REFERENCES I

-  ARANDA, S., MAS, G., AND DI CROCE, L. (2015).  
**REGULATION OF GENE TRANSCRIPTION BY POLYCOMB PROTEINS.**  
*Science advances*, 1(11):e1500737.
-  BURGE, C. AND KARLIN, S. (1997).  
**PREDICTION OF COMPLETE GENE STRUCTURES IN HUMAN GENOMIC DNA.**  
*Journal of molecular biology*, 268(1):78–94.
-  HASHIMOTO, T., SHERWOOD, R. I., KANG, D. D., RAJAGOPAL, N., BARKAL, A. A., ZENG, H., EMONS, B. J., SRINIVASAN, S., JAAKKOLA, T., AND GIFFORD, D. K. (2016).  
**A SYNERGISTIC DNA LOGIC PREDICTS GENOME-WIDE CHROMATIN ACCESSIBILITY.**  
*Genome research*, 26(10):1430–1440.
-  MITCHELL, P. J. AND TJIAN, R. (1989).  
**TRANSCRIPTIONAL REGULATION IN MAMMALIAN CELLS BY SEQUENCE-SPECIFIC DNA BINDING PROTEINS.**  
*Science*, 245(4916):371–378.