Genome Language Modelling

Sonika Tyagi, PhD

Associate Professor (Digital Health and Bioinformatics)

School of Computational Technologies

What's next...

RMIT UNIVERSITY

Acknowledgements

2023 BRILLIANT WOMEN in Digital Health Awards
RECOGNISING WOMEN IN HEALTH & AGED CARE
WINNER

SUPERBUGAI

MONASH University

Australian Academy of Science

Australian Government
National Health and Medical Research Council

BIOPLATFORMS AUSTRALIA

The current and past members of the Tyagi Lab

Tyrone Chen
Yashpal Ramakrishnaiah
Navya Tyagi
Murali Aadithya MS
Imrad Nyeen
Eleanor Cummins
Naima Vahab
Lipika Singh
Melcy Phillip
Tarun Bonu
Alex Dubrovsky
Jasbir Dhaliwal
Esha Singh
Sarthak Chauhan

Outline

Big Genomic Data

- Computational Epigenomics
- Large scale data on DNA, RNA, and Protein

Healthcare data

- Electronic Medical Records

Data Integration

- Molecular
- Molecular + Healthcare

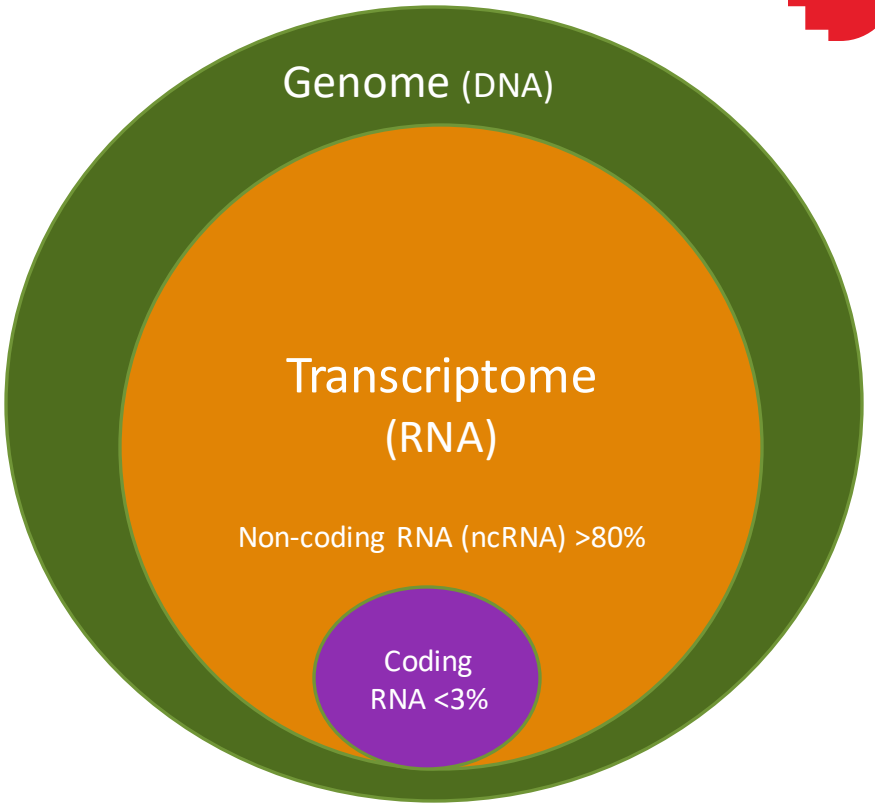
Big Genomic Data:

Biomolecules sequencing
Numerical measurements
Qualitative data

What's next...

DNA -> RNA -> Protein

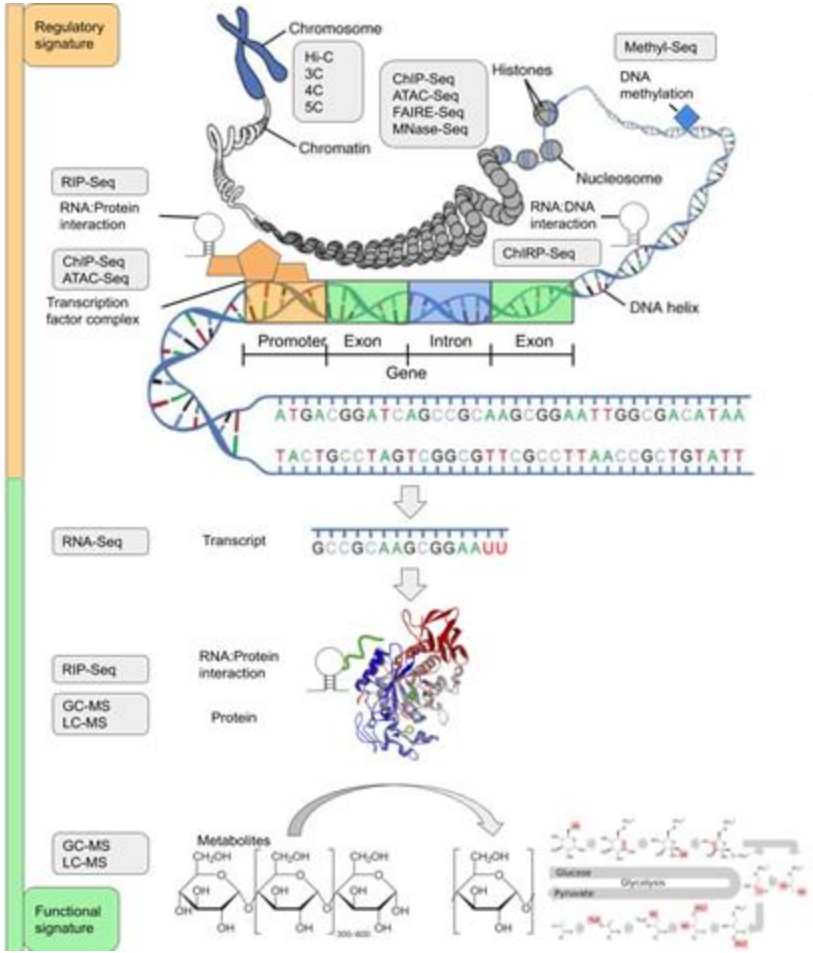
- Human DNA (3 Billion letters A, C, G, T)
- RNA (>200,000 transcripts)
- Protein (~20,000 proteins)



Big Genomic Data:

- From each high throughput assay thousands of data points are generated.
- High velocity, vol ume
- multidimensional and multimodal

Multi-omics



Genome structure-level

Open or Closed status

Genes data

RNA-level data

Protein-level data

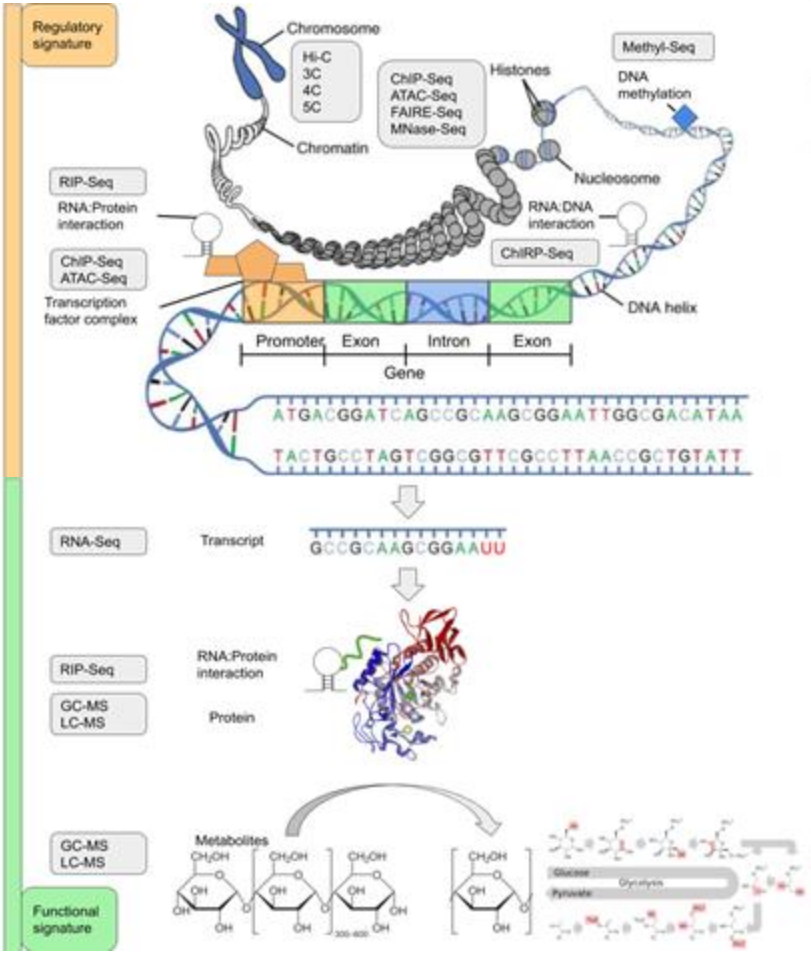
Metabolites

TyagiLab 2020

Big Genomic Data:

- From each high throughput assay thousands of data points are generated.
- High velocity, vol ume
- multidimensional and multimodal

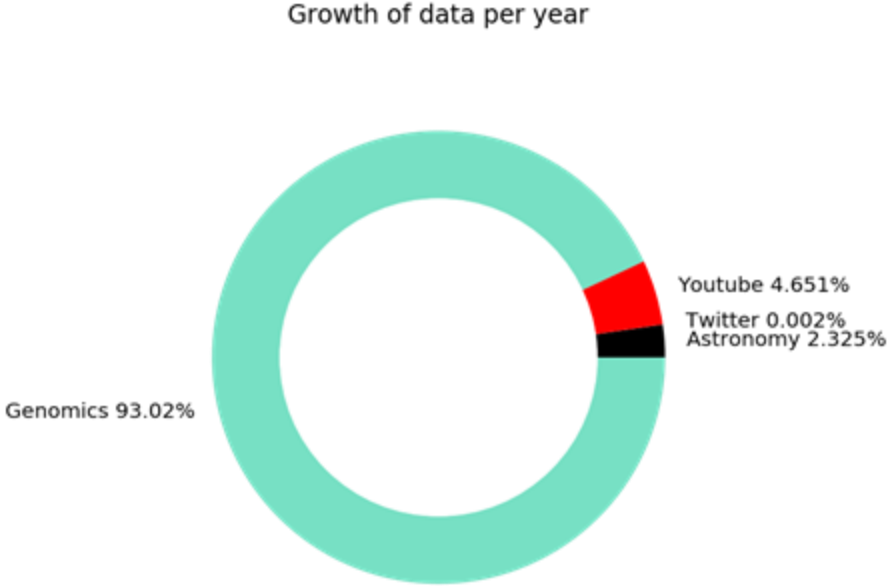
Multi-omics



DATA	Feature 1	Feature 2
Sample 1	3.142	2.7
Sample 2	10000	88.88

METADATA	Sample Type
Sample 1	Treatment A
Sample 2	Treatment B

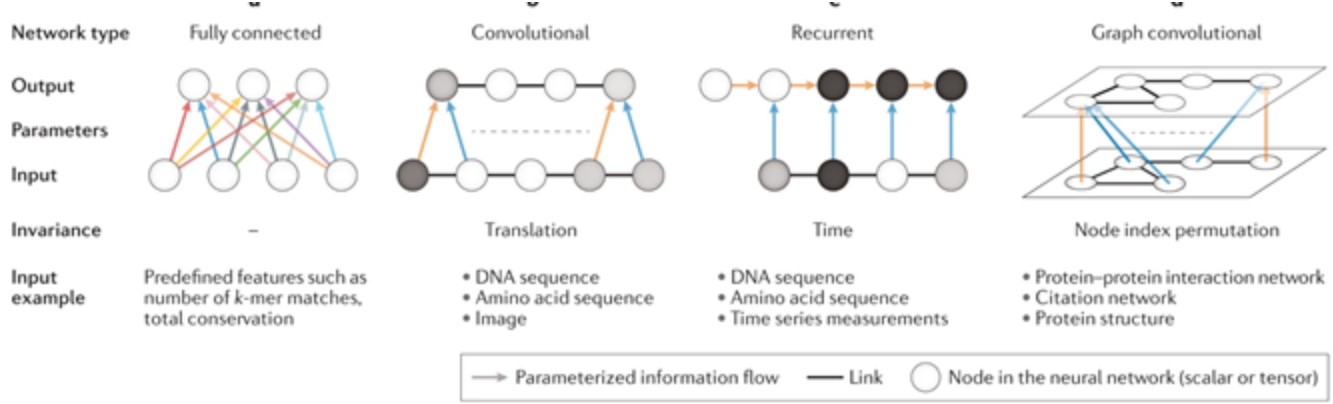
Main contributor to Big data growth is genomics



Tyrone Chen et al 2023, Tyagi Lab

Deep Learning Applications for Genomics

1. Pattern recognition
2. Predicting biomolecule structures
3. Classification or predictive modeling
4. Image analysis

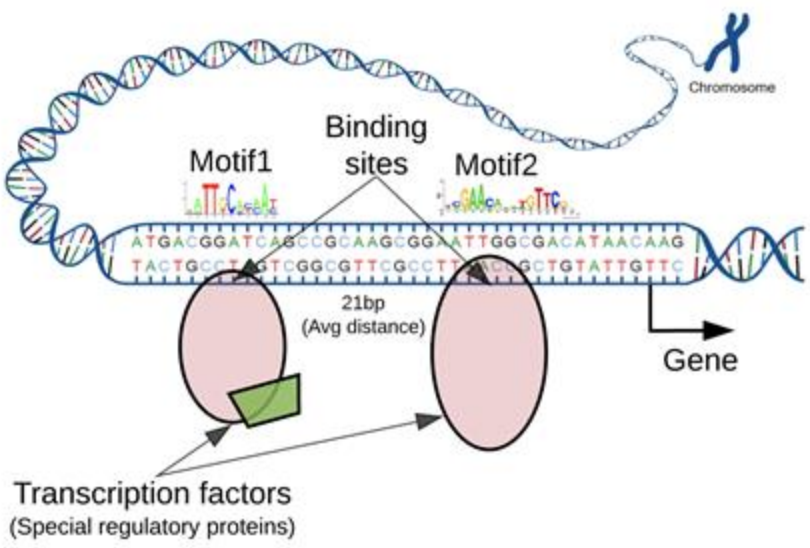


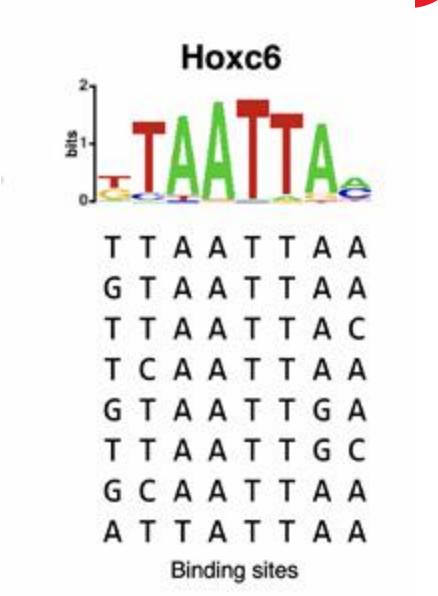
Eraslan *et al* 2019

Deep Learning Applications for Genomics

1. **Pattern recognition: e.g. Gene regulatory motifs**
2. Predicting Biomolecules:
    a. Modeling ncRNA structures,
    b. and functional motifs

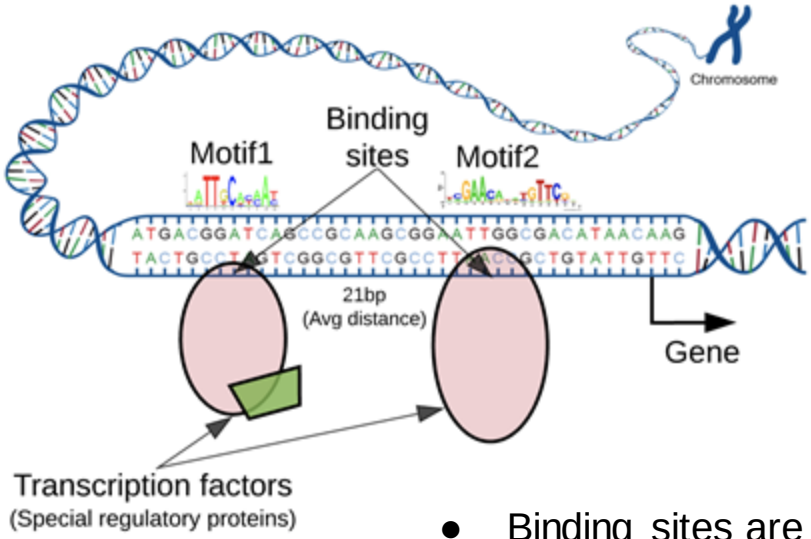
Gene regulatory code: Motifs

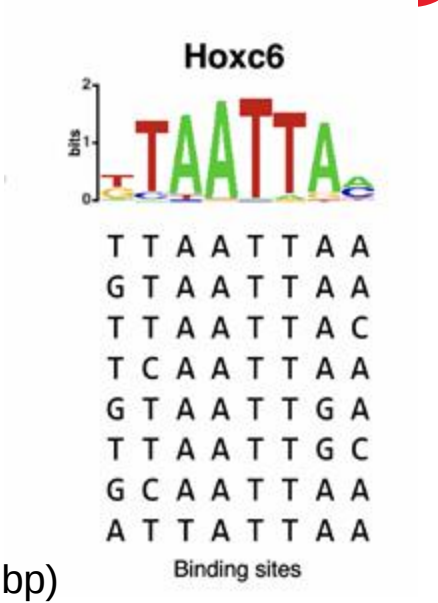




11

Gene regulatory code: Motifs





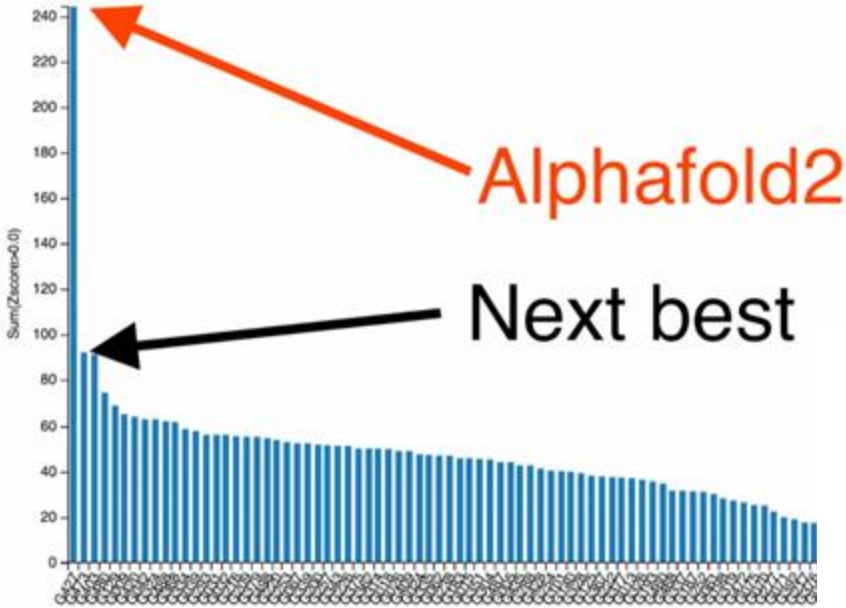
- Binding sites are very short patterns (5-12 bp)
- Genome is much longer (billion bp), results in very high false positive rate
- All combinations of motifs is exponential

Generally millions of binding sites for a TF are found but only a few thousands are bound

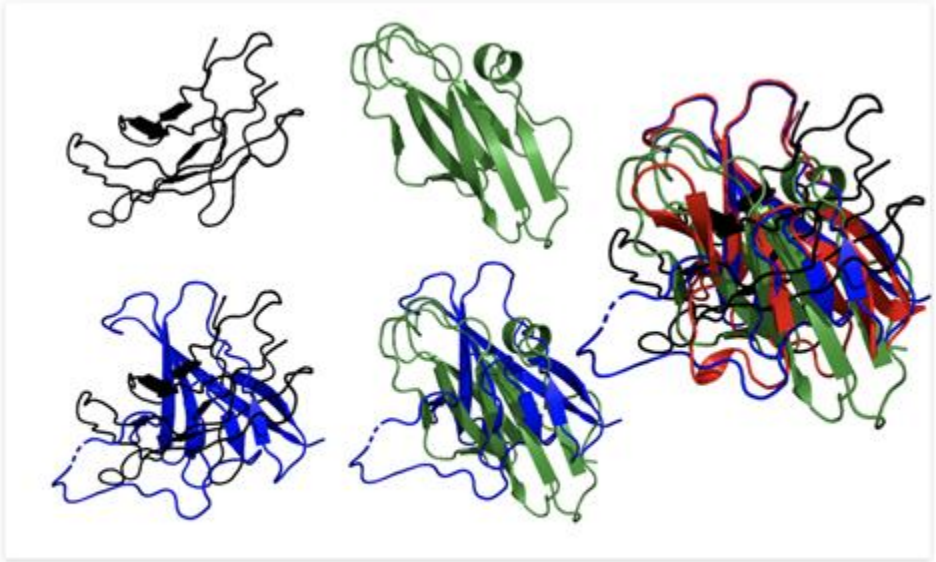
12

Deep Learning Applications for Genomics

1. Pattern recognition: e.g. Gene regulatory motifs
2. **Predicting Biomolecules:**
    a. **Modeling ncRNA structures,**
    b. **and functional motifs**

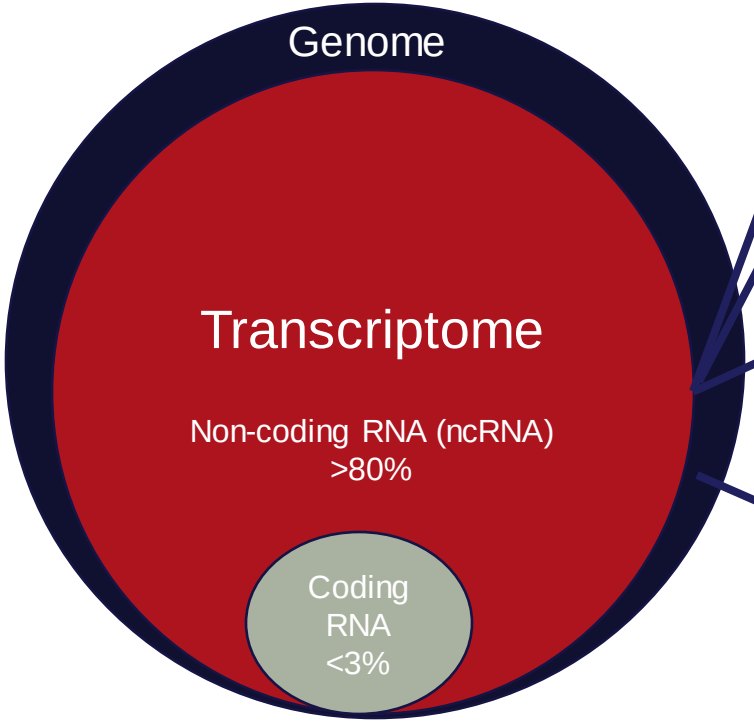


14th CASP Winner



Top: highest-ranked models for the target T1064 submitted by the Zhang (black) and Baker (green) human groups. Bottom: models aligned with the crystal structure. Right: all three models (Zhang, Baker and AlphaFold 2) aligned with the crystal structure. The submissions were obtained from the CASP14 webpage on Tuesday 1st December, 2020.

Non-coding Genome



tRNA, rRNA, other

Small ncRNA <200 nt

miRNA

siRNA

piRNA

snRNA

snoRNA

Regulatory Elements

Long ncRNA >200 nt

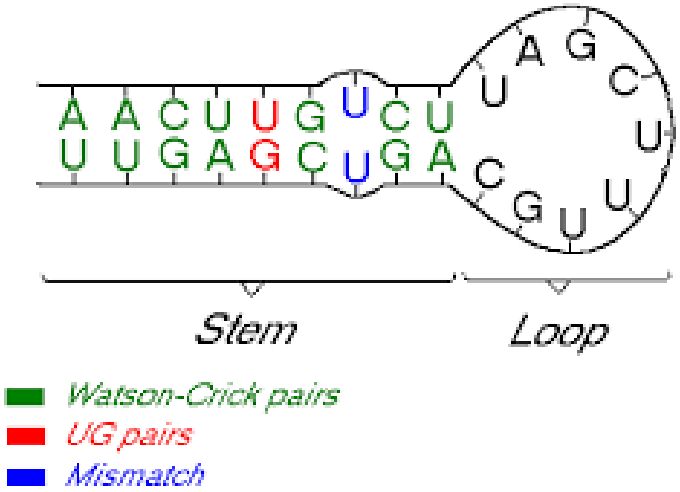
< GENCODE 31

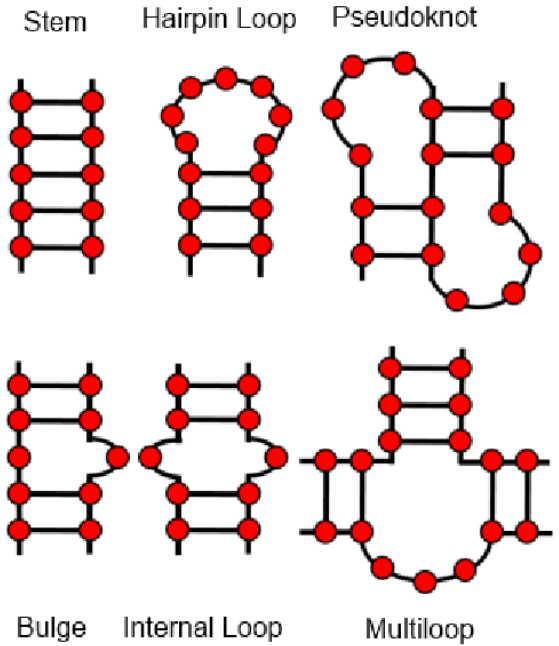
- lincRNA
- macroRNA
- Antisense
- Bidirectional
- 3' overlapping
- Promoter lincRNA
- Processed transcript
- Sense intronic
- Sense overlapping

lncRNA

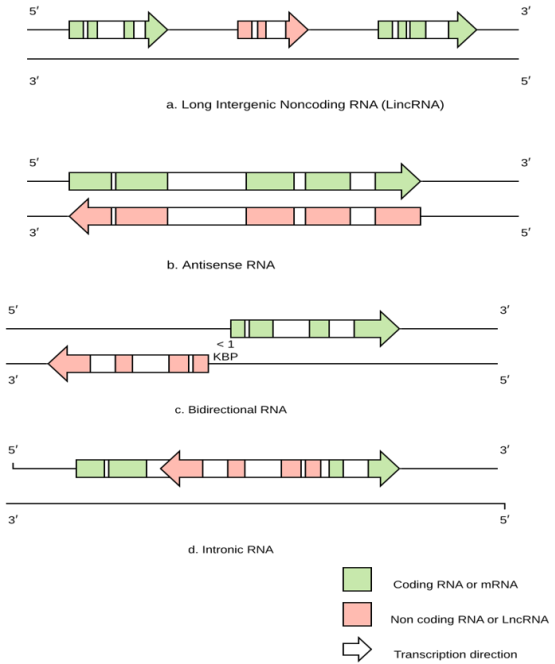
RNA SEQUENCE FOLDS ON TO ITSELF INTO SECONDARY CONFORMATIONS

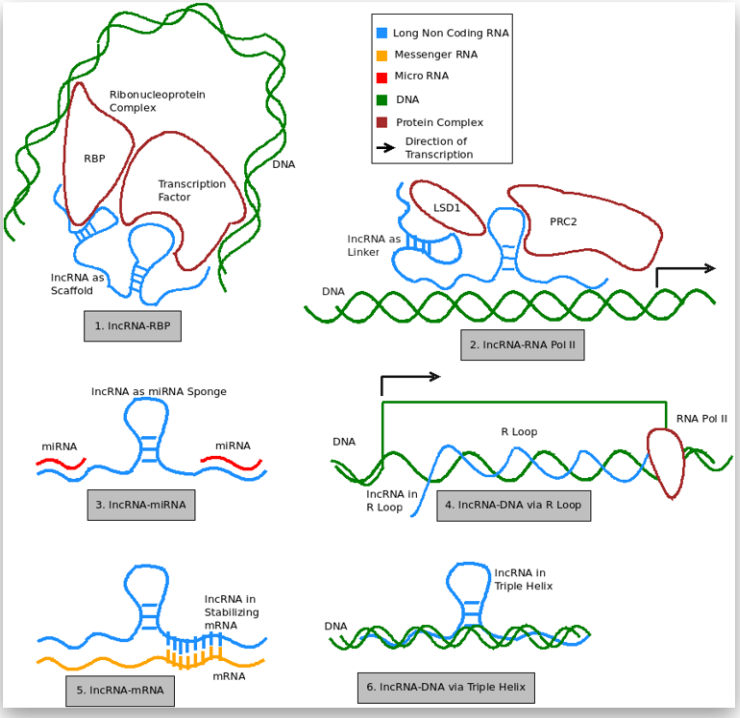
AACUGUCUUAGCUUUGCAGUCGAGUU





Long noncoding RNA (lncRNA)

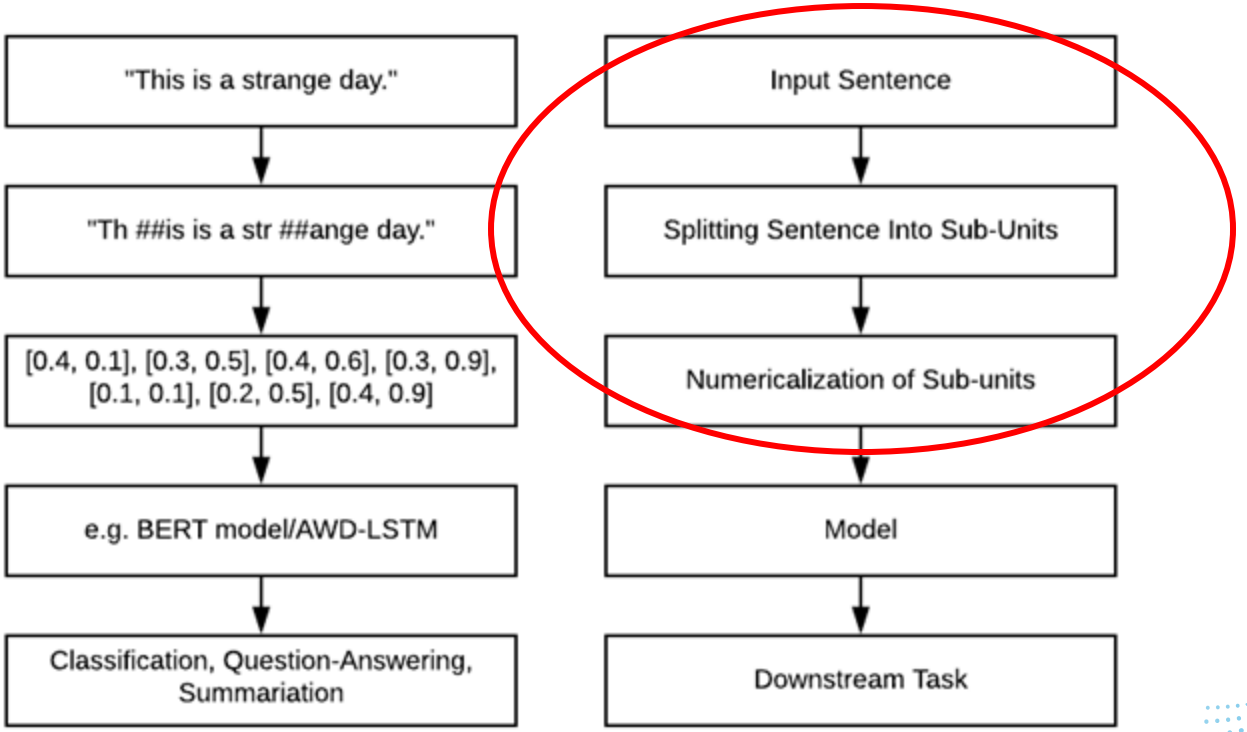




Ramakrishnaiah, Kuhlmann, and Tyagi CBM 2020

RMIT Classification: Trusted

Conventional NLP pipeline



Biological “words” can be learned from the data

[Rule-based] Predefined words

```
Tokens (EN): [Hello] [World]
Tokens (CN): [你好世界]
```

[Data-driven] Learned words

```
Tokens (EN): [Hello][_Wor][ld]
Tokens (CN): [你好][世界]
```

[Rule-based] Predefined k-mer/n-gram

```
Tokens (DNA): [ATCG][CGAT]
Tokens (RNA): [AUCG][CGAU]
```

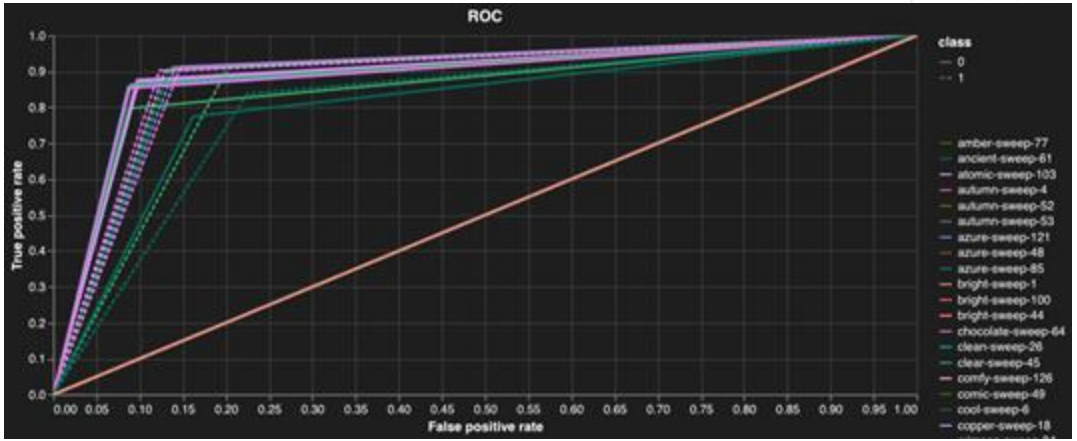
[Data-driven] Learned k-mer/n-gram

```
Tokens (DNA): [AT][CGCGAT]
Tokens (RNA): [AUC][GCGAU]
```

Kudo & Richardson, 2018; Chen et al, *unpublished*

genomicBERT model

Wandb sweeps



RMIT Classification: Trusted

No prior annotation is required

SgrT

Homology with part of
cis-regulatory element in Ecoli K12

Putative nucleoid protein
binding domain

Maps to part of cis-regulatory element in Ecoli K12

Short motifs

A composite signature of three short motifs corresponding to the **long motif**

Chen et al, unpublished

Precision Medicine

- Deep Learning has a huge potential for biomedicine.
- The biggest impact is in **Precision Medicine** that is a data driven approach:
  - Molecular data (e.g. omics)
  - Clinical data (e.g. Randomized trials)
  - Health data (e.g. Electronic medical records, wearables)

"To understand and treat disease by integrating multi-modal/multi-omics data from an individual to make patient-tailored decisions."

Genetic information

Health data

Other digital information

Diagnosis

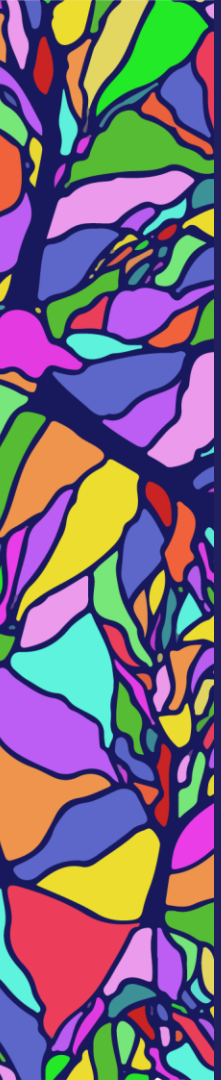
Treatment options

Prognosis

What's next...

Summary

- ★ New high throughput R&D activities generate measurements at scale.
- ★ New Big biomedical data present new challenges for ML
- ★ Genomics data often contain correlated data of common biological activities and integration of different data types provides a systems view.
- ★ Advances language models can be applied to genomic language directly to infer the grammar



Questions?