Lecture 15: Multi-Omics & Cancer Genomics Overview

Model overfitting

	Overfitting
Definition	- Statistically: The production of an analysis that corresponds too closely or exactly to a particular set of data and may therefore fail to fit additional data or predict future observation reliably. - Machine learning: the method is more complex than the problem, such that it can perform well on the training dataset but does not perform well on the testing dataset.
Cause	-Too long training time of the model makes the model learn irrelevant information. -The architectural complexity of model may cause noise to fit the data
Ways to detect	-Train-validation-test split • Train: 70% • Validation: 15% • Test: 15% *Reminder: Do not process the testing data during training or validation.  -Cross-validation • 5-fold validation • Leave-one-out • Reliable evaluation • Expensive and time-consuming
Indicator	-Loss function • The difference between training loss and validation loss • The performance may be OK even if overfitting

	-Performance • Precision, recall, F1 score • Make sure all of them have reasonable values (no bug) [*some overfitting issues may be caused by overfitting] • Performance on training dataset has the increasing trend • The difference between training dataset and validation dataset may increase over the time

Sources of over-complexity

1.Data: Too little, not reflect the true distribution

2.Model: Too large, too many useless parameters

3.Connectivity: Too strong, co-adaptation

4.Parameter value range: Too large, model too flexible

5.Training time: Too long, tend to overfitting

Techniques to deal with overfitting

1.Increase training dataset: training with more data.

2.Data augmentation: adding some noise to the input makes the data stable without affecting the data quality, while adding noise to the output makes the data more diverse.

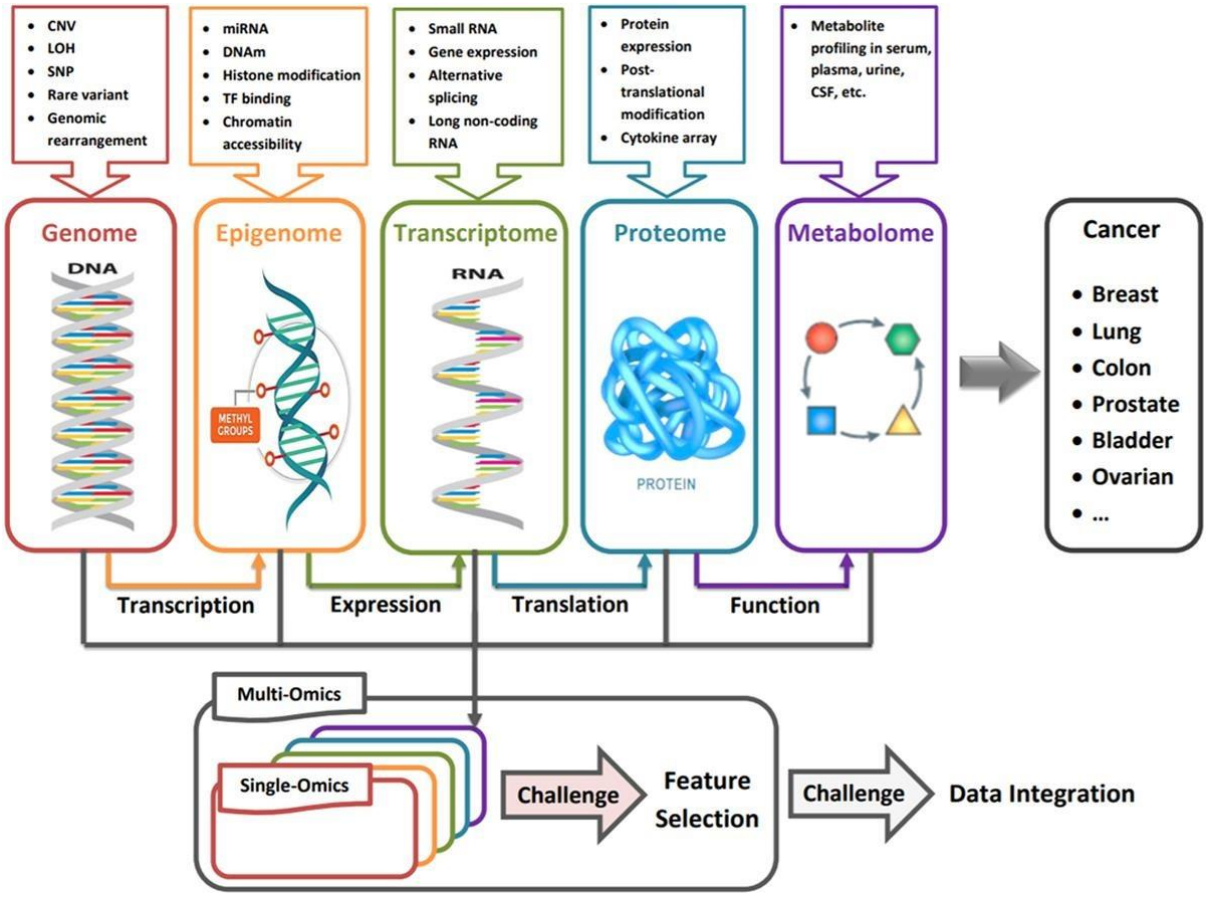
3.Regularization: penalizing over the large weight values to limit the model's variance.

4.Early stopping: pausing the model's training before memorizing noise and random fluctuations from the data

5. Simplify data: decreasing the complexity of the model, e.g. pruning a decision tree, reducing the number of parameters in a neural network and using dropout on a neural network.

Multi-omics

Multi-omics data broadly cover the data generated from genome, proteome, transcriptome, metabolome, and epigenome. The spectrum of omics can be further extended to other biological data such as lipidome, phosphoproteome, and glycol-proteome.



Statistical Testing

1. T-test

-Check there is a significant difference between two sets of data

-Calculate the test statistic based on the mean and variance of the data

Test statistic: $(\bar{x}_1 - \bar{x}_2) \ / \ s_p(\sqrt{1/n_1 + 1/n_2})$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, n_1 and n_2 are the sample sizes, and where s_p is calculated as:

$$\mathbf{s_p} = \sqrt{(n_1\text{-}1)s_1^2 + (n_2\text{-}1)s_2^2 \ / \ (n_1+n_2\text{-}2)}$$

-Calculate the p-value of the test statistic t

-Conclude that this gene expressed differently under two conditions if p-value smaller than 0.05

Different kinds of T-test

- Paired vs Unpaired
  For paired one, we cannot shuffle the values
- One-tailed vs Two-tailed test
  Two-tailed test: different or the same
  One-tailed test: greater, larger, smaller, at least

2. Fisher's exact test

-Calculate the p-value from the contingency table

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!(a+b+c+d)!}$$

	In gene set	Not in gene set	Total
In pathway	100 (a)	9000 (b)	9100
Not in pathway	113 (c)	11000 (d)	11113
Total	213	20000	20213