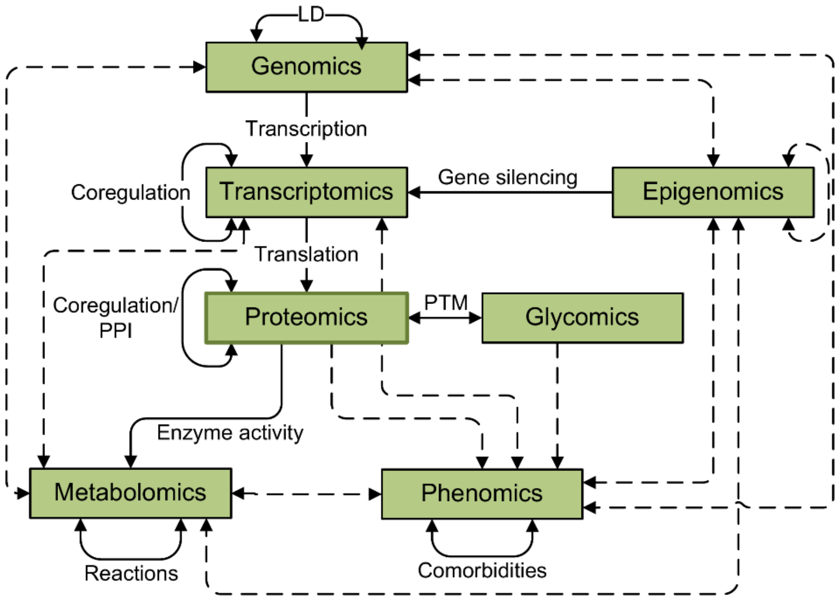
Morning session: overview

Unsupervised learning: dimension reduction and data integration

- Principal Component Analysis (PCA)
  - Maximal variance principle
- Partial Least Squares (PLS)
  - Maximal covariance principle
- Two-way Orthogonal PLS (O2PLS)
  - Multi-omics data integration

Background

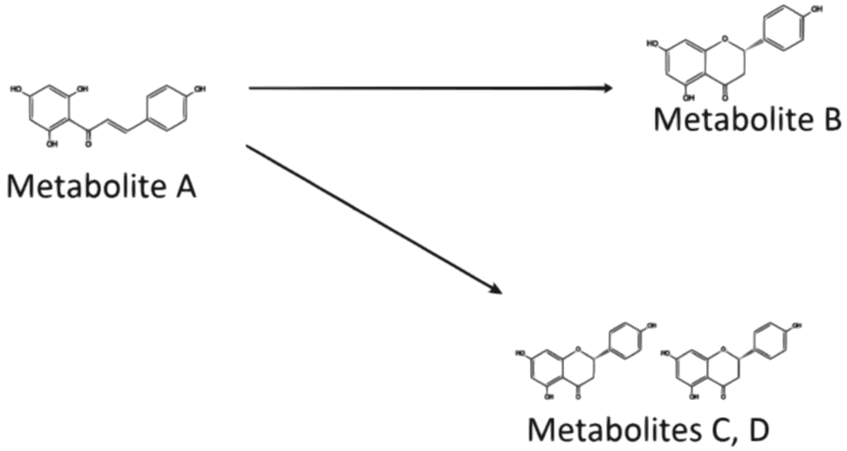


- Recent advances in technology provided many types of biological datasets (multi-omics data)
- Different levels of biological variation measured
- Need for integrative approaches: combine data and extract information

Integrative approach: aims

- How does variation between omics datasets relate?
- Which types of features induce this variation?
- Can we benefit from a joint/integrative analysis?

Metabolomics

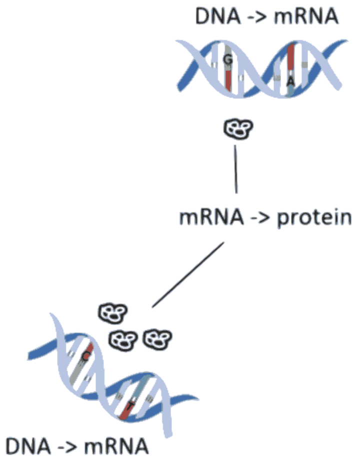


Metabolite A

Metabolite B

Metabolites C, D

Transcriptomics



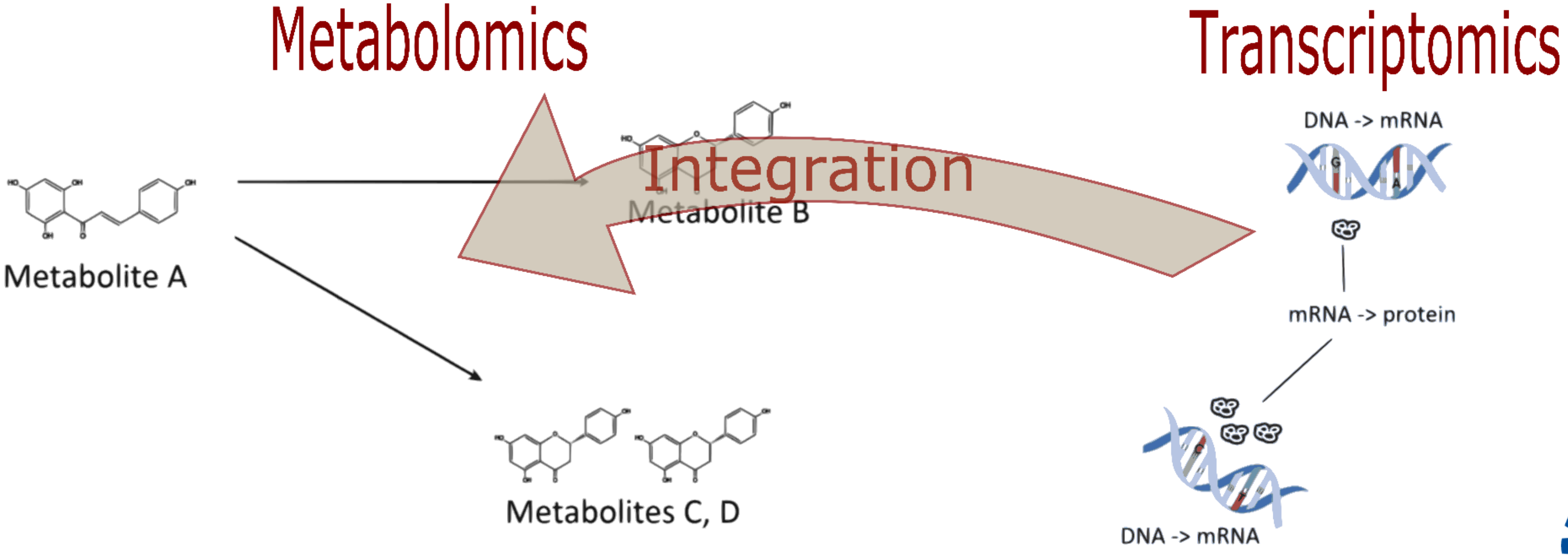
DNA -> mRNA

mRNA -> protein

DNA -> mRNA

Integrative approach: aims

- How does variation between omics datasets relate?
- Which types of features induce this variation?
- Can we benefit from a joint/integrative analysis?



Metabolomics

Transcriptomics

Integration

Metabolite A

Metabolite B

Metabolites C, D

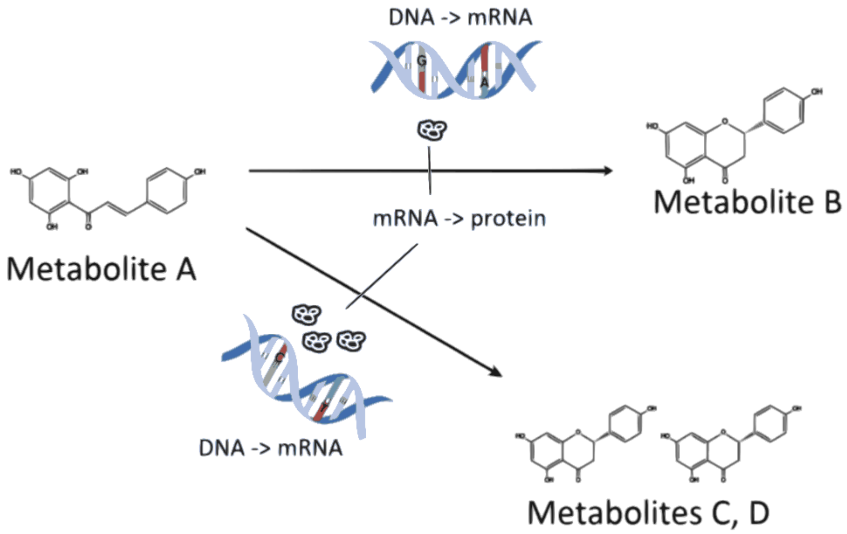
DNA -> mRNA

mRNA -> protein

DNA -> mRNA

Integrative approach: aims

- How does variation between omics datasets relate?
- Which types of features induce this variation?
- Can we benefit from a joint/integrative analysis?



Integrative approaches: means

There is typically

- High correlation *among* features
  - Genes are correlated
  - Metabolites are correlated
- Relation *between* features from two datasets

- Latent variable approach: few independent latent variables drive association

Example: bivariate data

- Suppose we have two genes, and two metabolites x_1, x_2, y_1, y_2
- The variance of x_1 is larger than of x_2
- The variance of y_2 is larger than of y_1
- Only x_2 and y_1 are correlated
- Which variables will get high weight with PCA? Why?
- Which variables should get high weight when you look at the relation between x and y?

Partial Least Squares (PLS)

- Let X and Y be two data matrices
  - Size: N times p and q, respectively
  - p and q can be very large
- Recall: in PCA, variance is maximized
- We are interested in the **co**variation between X and Y
- Consider covariance between X and Y: $Y^\top X$
  - Dimension: q times p

Partial Least Squares (PLS)

- Maximize covariance between projections of X and Y
  - Weights w for X and c for Y
  - Maximize $c^\top Y^\top X w$ such that $w^\top w = c^\top c = 1$
  - Lagrange: $c^\top Y^\top X w \ - \lambda_w w^\top w - \lambda_c c^\top c$
  - Take derivatives w.r.t. w and c separately, set to zero, and solve
- The solution is given by the singular value decomposition
  - Best w is the first right singular vector, best c is the first left singular vector of $Y^\top X$
- Similar interpretation as PCA, except that we focus on covariance
- The scores can again be calculated as: $t = Xw$ and $u = Yc$

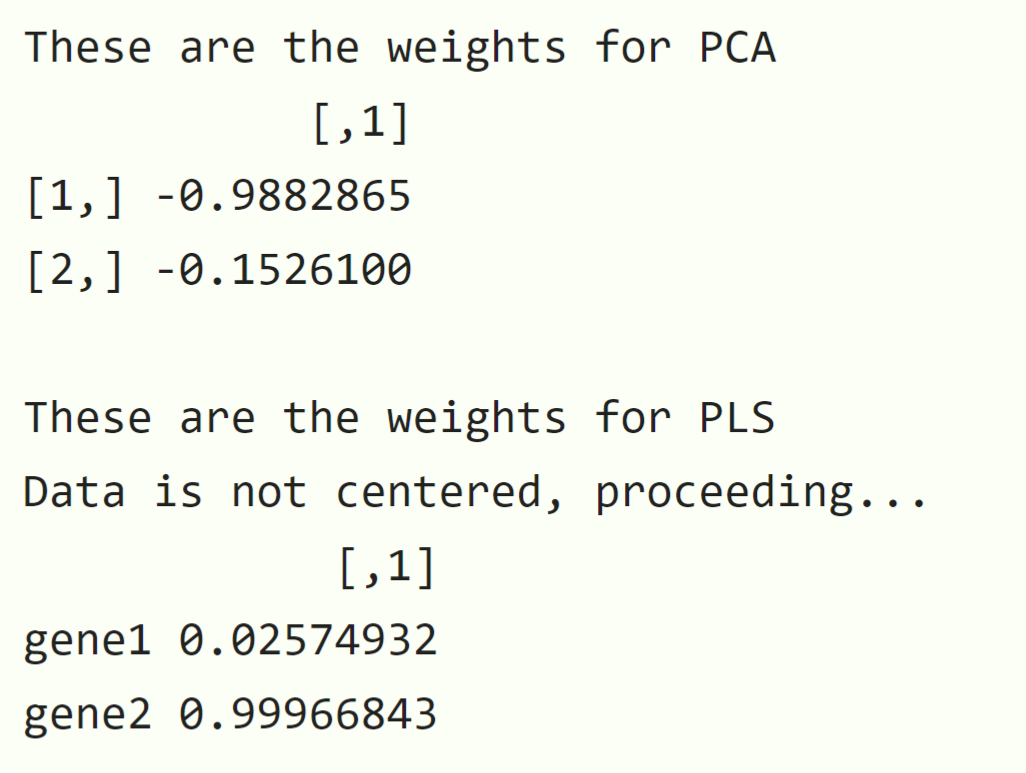
Example: PLS

```
library(OmicsPLS)

gene1 <- rnorm(100)
gene2 <- rnorm(100,sd=0.75)
metab1 <- rnorm(100)
metab2 <- gene2

X <- cbind(gene1, gene2)
Y <- cbind(metab1, metab2)

svd(X,0,1)$v
o2m(X, Y, 1, 0, 0)$W.
```



```
These are the weights for PCA
           [,1]
[1,] -0.9882865
[2,] -0.1526100

These are the weights for PLS
Data is not centered, proceeding...
            [,1]
gene1 0.02574932
gene2 0.99966843
```

Partial Least Squares: summary

- For given datasets X and Y, we want to inspect their relation
- We consider directions of maximal covariance
- This direction is represented by weights for each feature, calculated by a singular value decomposition of the covariance matrix $Y^\top X$
- The projections of the data onto these weights are called the scores.
- One can interpret or plot the weights and scores to understand which features/samples are most important

Morning session: overview

Unsupervised learning: dimension reduction and data integration

- Principal Component Analysis (PCA)
  - Maximal variance principle
- Partial Least Squares (PLS)
  - Maximal covariance principle
- Two-way Orthogonal PLS (O2PLS)
  - Multi-omics data integration