



Bioinformatika

Gene expression data analysis

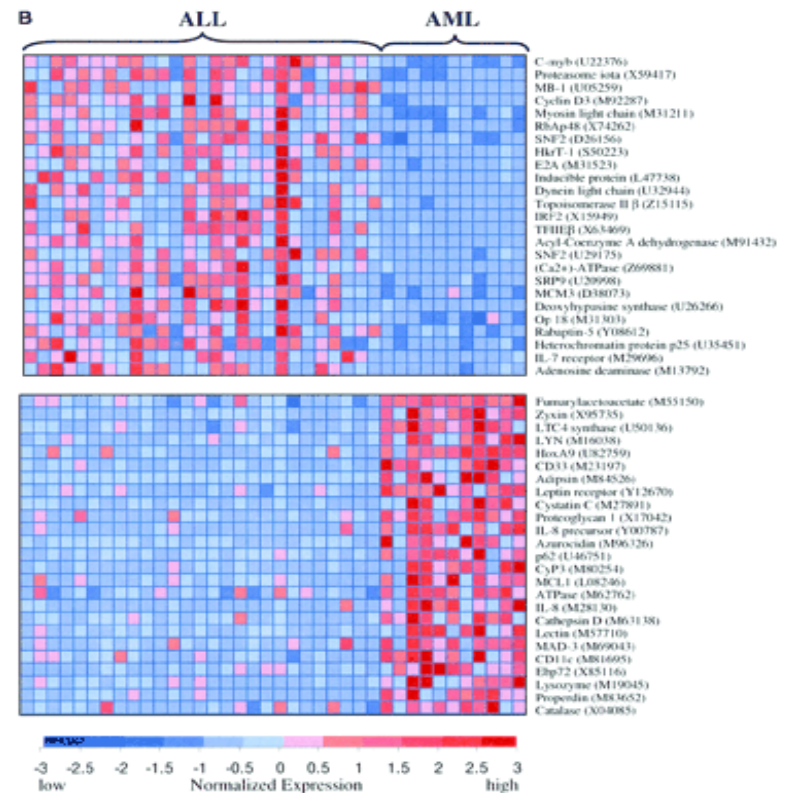


Michael Anděl

Classification task

Challenge:

↻ samples (10^1) x features (10^3)

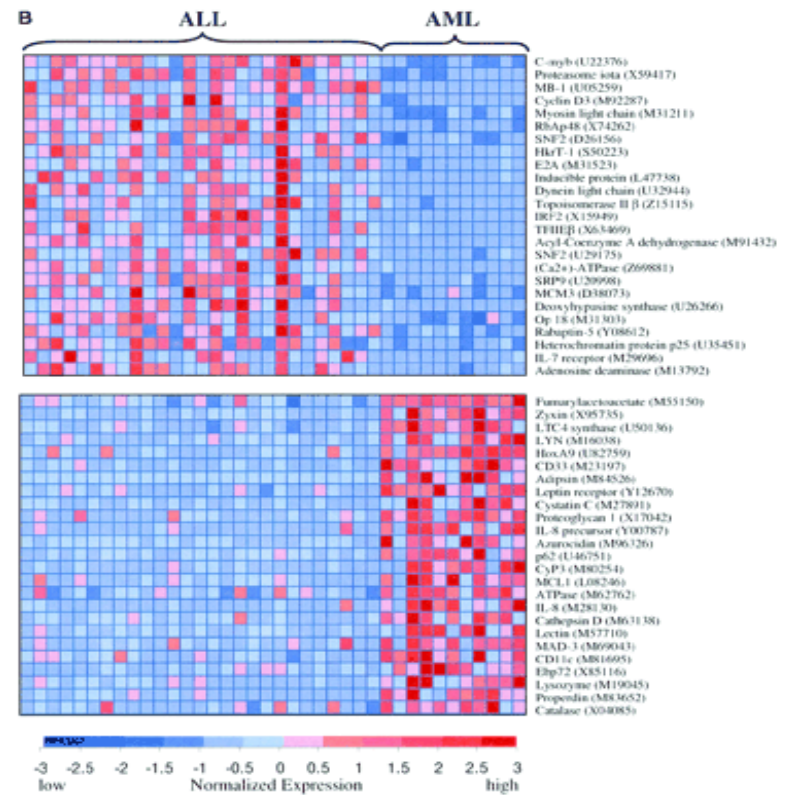


Golub et al.: *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.* Science, 1999

Classification task

Challenge:

- ↻ samples (10^1) x features (10^3)
- ↻ False hypotheses, overfitting



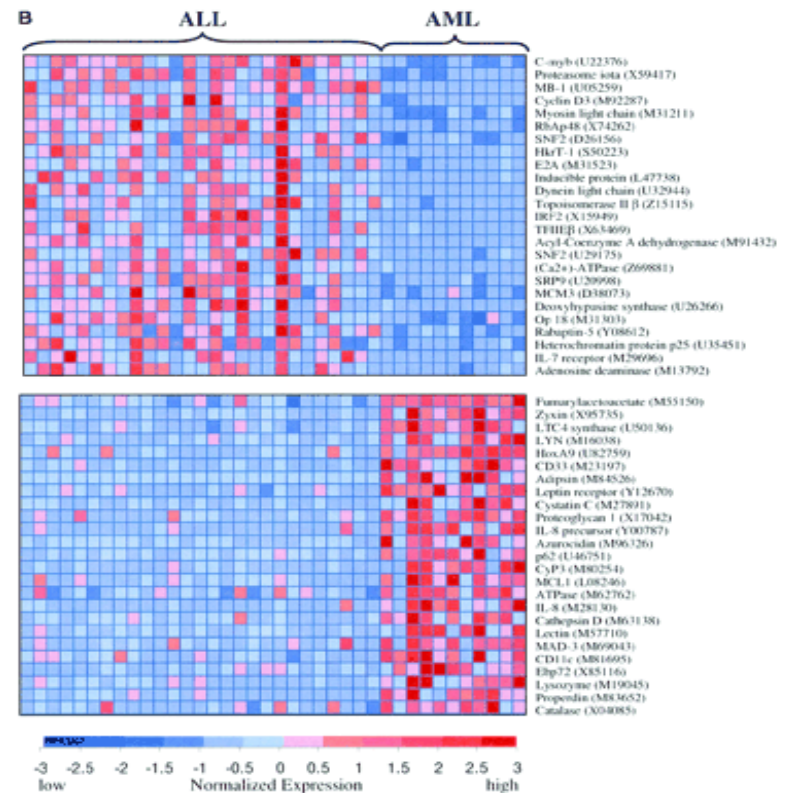
Golub et al.: *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.* Science, 1999

Classification task

Challenge:

- ↻ samples (10^1) x features (10^3)
- ↻ False hypotheses, overfitting
- ↻ Interpretability:
are the expressed genes the causal ones?

What to do?



Golub et al.: *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.* Science, 1999

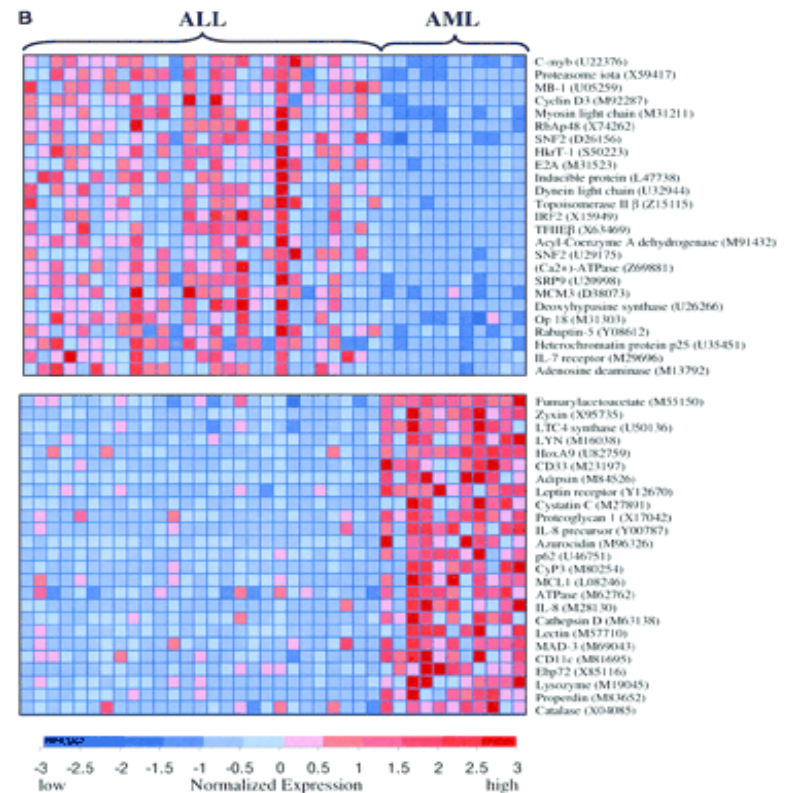
Classification task

Challenge:

- ↻ samples (10^1) x features (10^3)
- ↻ False hypotheses, overfitting
- ↻ Interpretability: are the expressed genes the causal ones?

What to do?

- ↻ Decrease number of hypotheses



Golub et al.: *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.* Science, 1999

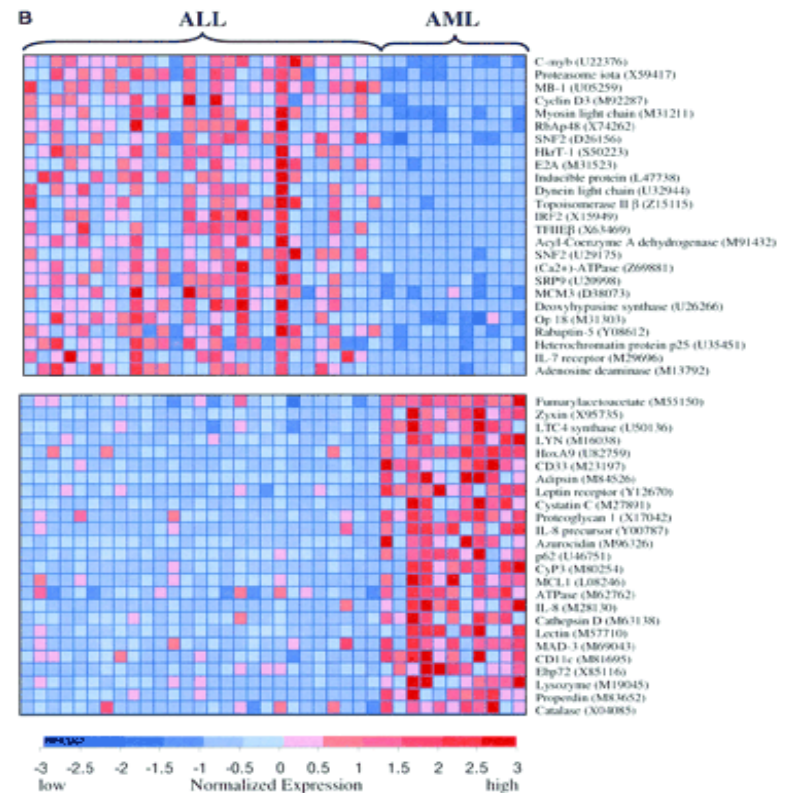
Classification task

Challenge:

- ↻ samples (10^1) x features (10^3)
- ↻ False hypotheses, overfitting
- ↻ Interpretability:
are the expressed genes the causal ones?

What to do?

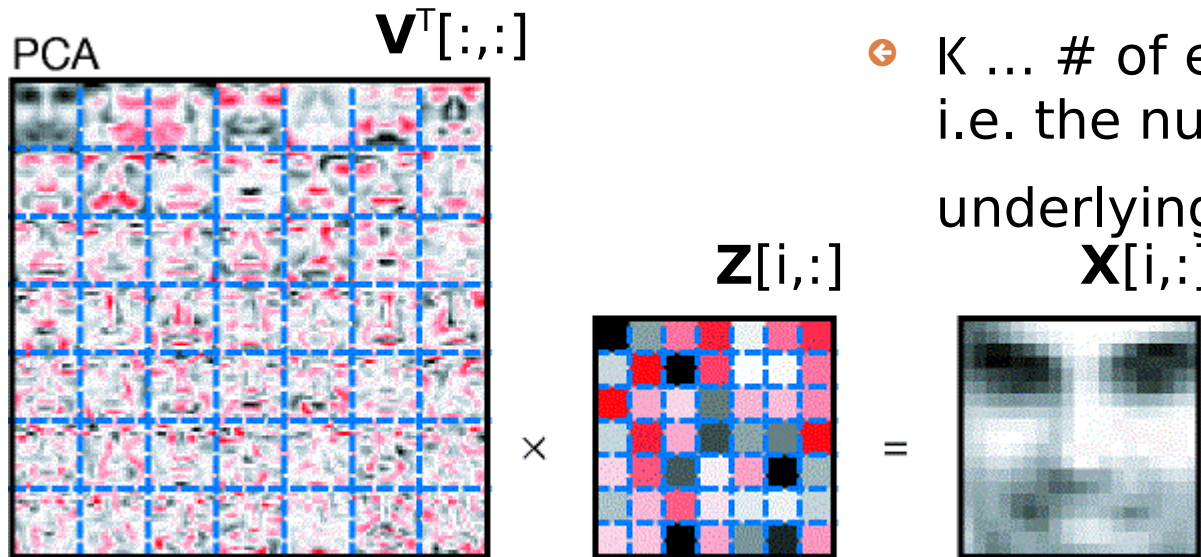
- ↻ Decrease number of hypotheses
- ↻ Analyze in terms of more *abstract entities* than genes, e.g. principal components



Golub et al.: *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.* Science, 1999

PCA - motivation

- ↻ M ... # genes
- ↻ N ... # samples
- ↻ $X (N \times M)$... GE data in the space of *genes*
- ↻ $V (M \times K)$... transformation basis, eigengenes
- ↻ $Z (N \times K)$... transformed GE data in the space of eigengenes
- ↻ K ... # of eigengenes, i.e. the number of underlying concepts



Lee et al.: *Learning the parts of objects by non-negative matrix factorization*. Science, 1999

Assignment

Data:

- 7,129 GE profiles of 72 patients
- 25 samples: acute myeloid leucaemia (AML)
- 47 samples: acute lymphoblastic leucaemia (ALL)

Golub, T., et al. *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*. Science 286.5439 (1999): 531-537.

Task:

- Construct decision model to differentiate these types of tumours. **Just complete the code in the script attached** → `ge_cv.m`
- **Deadline - 19.5.**

Assignment

Part I:

1. Learn a decision tree on subjected data. Use Matlab class `ClassificationTree` and its method `fit`.
2. Show the tree (method `view`) and enumerate its **training** accuracy.
3. How would you interpret this model? Which gene is crucial for the decision?
4. Is this gene really the one causing the cancer? Look up in the article *Golub et al., 1999*.
5. Estimate **real** accuracy of the tree. Use e.g., cross-validation (alternatively, you can split the data).
6. Compare it with the **training** accuracy.

Assignment

Part II:

1. Learn a basis-matrix V of the data. Use the **attached** function `pca.m`.
2. For for a **range** of component numbers K :
 - a) project the original data X to the top K components of V . The result are data Z with reduced dimensionality: $Z = XV_{1:K,1:}^T$
 - b) Create a tree out of these reduced data. Show it and enumerate its **training** accuracy.
3. Compare **all** the trees resulting from the reduced data and pick the “best” according to its accuracy and structure. Follow the Occam razor ;-)

Assignment

Part II:

4. Estimate the **real** accuracy of the “best” chosen tree. Again, by e.g. crossvalidation.
5. Extract the genes active in the discriminative components. The **discriminative components** are those vectors of basis-matrix V , which refer to the features **your tree consists of**. To extract the **active** genes from a component use the function `mineGenes`.
6. Resulting **gene-sets** related to each of the discriminative component shall hopefully refer to some abstract biological processes. Use [Gorilla](#) to enrich these gene sets in **Gene-ontology** terms.
7. Make a story ;-)