(1) Clustering genes

Problem

- Genes are responsible for protein synthesis in living cells.
- Genes interact with each other forming complex gene-networks.
- The function of most genes is still unknown.
- Grouping together genes with similar behaviour can help understanding their function.

(1) Clustering genes

Data

- The expression level (i.e. amount of proteins synthesized in a certain time) of genes can be measured by DNA microarrays.
- Expression level of genes from pathological cells are often measured relative to healthy cells (the control).
- The *leukemia* dataset consists of expression levels for 5147 genes in 72 patients: 47 affected by acute lymphoblastic leukemia (ALL), 25 by acute myeloid leukemia (AML).

(1) Clustering genes

Task

- Expression levels depend strongly on the gene considered $\Rightarrow$ need to be normalized
- Use information gain (see decision tree lesson) to choose the best threshold for binarizing each gene (classes are ALL vs AML).
- Select the 100 genes with highest information gain
- Cluster the 100 (binarized) genes by agglomerative hierarchical clustering.