Bioinformatics Project

Andrea Passerini
passerini@disi.unitn.it

Artificial Intelligence for Bioinformatics

Option 1: Building a Bayesian model of leukemia pathologies

Data collection

- 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).
- A small subset of 5 among the genes most correlated with one of the two pathologies have been selected.
- The datasets have been split into 58 patients for training and 14 patients for testing.

Option 1: Building a Bayesian model of leukemia pathologies

Building Bayesian Network

- Learn structure and parameters of the Bayesian network on the training set
- Evaluate performance of the learned network on the test set.
- Compare different networks:
  - hugin-lite structure learning (statistical-test based)
  - hugin-lite structure learning (score based)

Option 2: Modeling sequence families by profile HMM

Data collection

1. go to the database of protein families (PFAM) : `http://pfam.sanger.ac.uk/browse`
2. choose a protein family
3. go to its sequences (click on xxx sequences on top left menu, where xxx is number of sequences)
4. download sequences in Fasta format (check *You can also download a FASTA format file containing the full-length sequences for all sequences in the full alignment.*)

Option 2: Modeling sequence families by profile HMM

Building profile HMMs

1. Recover the SAM tool for profile HMM:
   - installed in the PC lab at `/usr/local/NEWSAM`
   - downloadable from `http://compbio.soe.ucsc.edu/sam.html`
2. Read the quick usage overview `https://compbio.soe.ucsc.edu/papers/sam_doc/node4.html`
3. Build a model for the protein family you choose
4. Align sequences on the model
5. Score sequences on the model
6. Write a small report with:
   1. Learned model
   2. Alignments of the sequences to the model
   3. Scores of the sequences on the model