TermBrain

the HPO term predictor

Jonathan Boidol, Rene Schoeffel, Yann Spöri

December 7, 2013

Homology based function prediction

General approach:

- Search for annotated similar sequences
  with blast and hhblits (hits)
- Build subgraph of HPO containing the found annotations
- Calculate confidence for every annotation from some distance measure to the hits

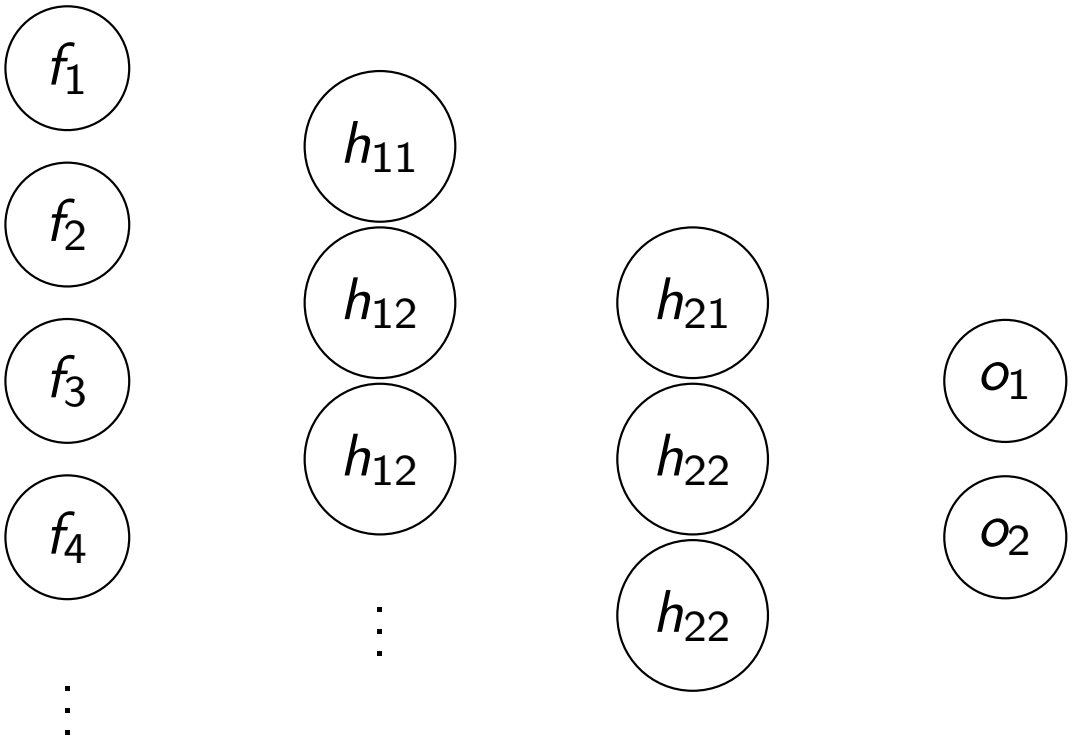
Preparations

- Prepare databases for annotated sequences
- Represent HPO Graph in predictor
- Merge trees corresponding to hits

Features

- use neural network to calculate confidence per node
- each node is assigned features derived from the merged tree
  - number of hits
  - min. E-value
  - avg. E-value [3, 0.0074, 0.45, 84, ...]
  - longest hit
  - ...

Network architecture



- two output nodes trained for the two possible predictions
- difference of predictions as confidence

Validation

- Inspection of the dataset shows: Most sequences have pairwise similarity $< 80\%$

  $\Rightarrow$ reduce set at 80%-level to remove highly similar clusters
- Crossvalidate over reduced set but allow non-reduced trainingset for similarity search during testing
- Calculate precision and recall per test sequence and average over all sequences

Results

Pre-Rec-Curve here

- F-measure *yy* (at confidence level x)
- Precision (at same confidence level x)
- Recall (at same confidence level x)