HyPnOBrain
your homology based HPO neural network predictor

Jonathan Boidol, Rene Schoeffel, Yann Spöri

December 10, 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
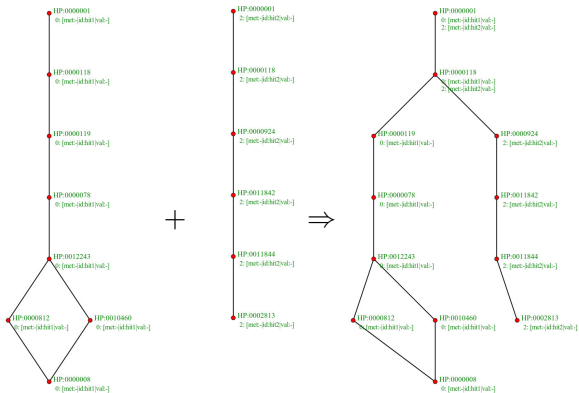
Features

- use neural network to calculate confidence per node
- each node is assigned 12 features derived from the merged tree
  - number of hits
  - min. E-value
  - avg. E-value
  - product of E-values
  - longest hit
  - . . .

  $[3, 0.0074, 0.45, 4.2e\text{-}7, 84, \dots]$

Network architecture
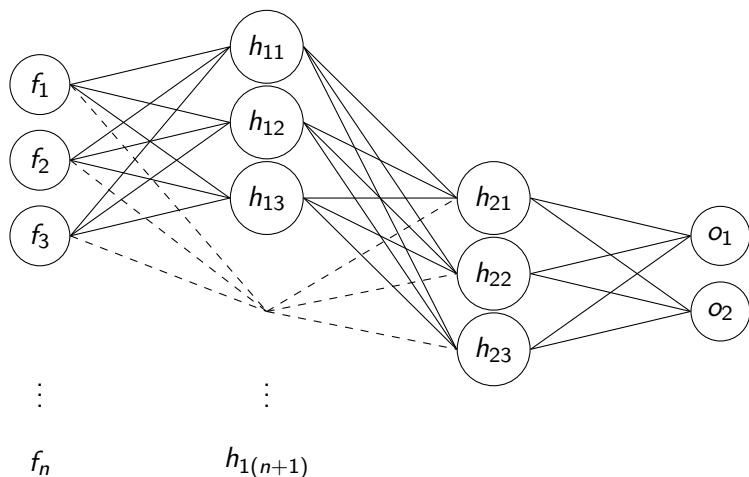
- ▶ fully connected net with two hidden layers
- ▶ two output nodes trained for the two possible predictions
- ▶ difference of predictions as confidence

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
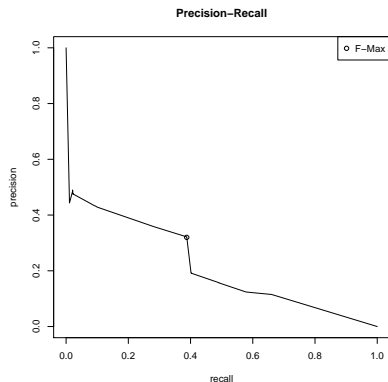
- ▶ F-measure 0.35 ± 0.03 (at confidence level 0.34)
- ▶ Precision 0.32 ± 0.04 (at same confidence level)
- ▶ Recall 0.39 ± 0.10 (at same confidence level)

Running predictions

```
e.g. ./hypnobrain -f sequence.fa -minConf 0.35

usage: hypnobrain [-h] (-s SEQUENCE | -f FASTAFILE)
                  [--svg] [-v VERBOSITY]
                  [-w {bash,plain,html}] [-p HPOFILE]
                  [-b BLASTDBFILE]
                  [-l HHBLITSDBFILE] [-c UNI2HPO]
                  [-e BLASTMINEVAL]
                  [-n NEURONALNET] [-k LOOKUPDB]
                  [--minConf MINIMALCONFIDENCE] [--fast]
```