

HyPnOBrain

your homology based HPO neural network predictor

Jonathan Boidol, Rene Schoeffel, Yann Spöri

January 23, 2014

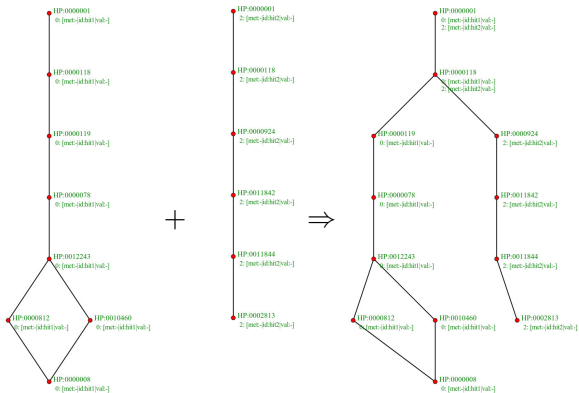
Homology based function prediction

General approach:

- ▶ Search for annotated similar sequences (hits) and transfer annotations
- ▶ HPO is hierarchical: Merge found annotations from different hits
- ▶ 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

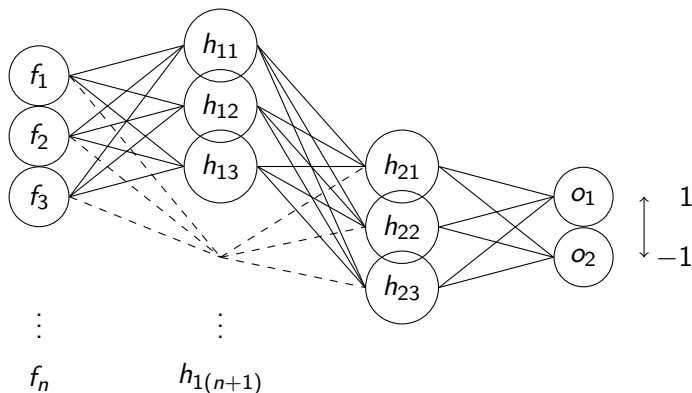


Features

- ▶ each node is assigned 12 features derived from the merged tree
 - ▶ length of query sequence
 - ▶ number of hits
 - ▶ longest hit
 - ▶ avg. hit
 - ▶ min. E-value
 - ▶ avg. E-value
 - ▶ product of E-values
 - ▶ best E-value from blast or hhblits
 - ▶ min. height in HPO-tree
 - ▶ max. height in HPO-tree
 - ▶ max overlap of query and all hits
 - ▶ length of best hit
- ▶ use neural network to calculate confidence per node

[3, 0.0074, 0.45, 4.2e-7, 84, ...]

Final network architecture

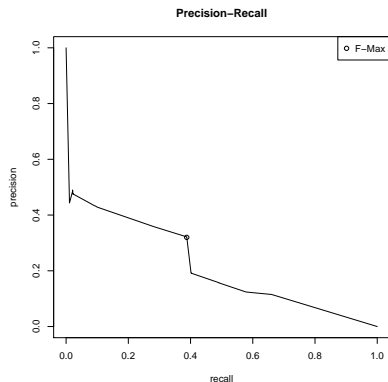


- ▶ different architectures evaluated
- ▶ fully connected net with two hidden layers
- ▶ two output nodes trained for the two possible predictions
- ▶ difference of predictions as confidence

Validation

- ▶ Inspection of the dataset shows: Most sequences have pairwise similarity $< 80\%$
 \Rightarrow no significant bias in the dataset caused by very close homologs
- ▶ 10-fold crossvalidation over 2815 sequences
- ▶ Calculate precision and recall per test sequence and average over all sequences
- ▶ Final model trained on 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)

HyPⁿOBrain

Your homology based HPO neuronal network predictor

Submit

Protein Sequence:

Either input your sequence to predict here:

```
>sp|Q03154|ACY1_HUMAN Aminoacylase-1 OS=Homo sapiens GN=ACY1 PE=1 SV=1
MTSKGPEEEHPSVTLFRQYLRI RTVQPKPDYGA AVAFEE TARQLGLGCQKVEVAPGYVV
TVLTWPGTNP TLLSILLNSHTDVVPVFKEHWSHDPFEAFKDSEGYIYARGAQDMKCVSIQ
YLEAVRRLKVEGHRFPRTIHM TFPD EEVGGHQGMELFVQRP EFHALRAGFALDEGIANP
TDAFTVFYSESPWWVRVTSTGRPGHASRFMEDTAAEKLHKV VNSILAFREKEWQRLQSN
PHLKEGSVTSVNLTKLEGGVAYNVIPATMSASFDFRVAPD VDFKAFEEQLQSWCQAAGEG
VTLEFAQKWMHPQVTP TDDSNPWWAAFSRVCKDMNLTLEPEIMPAATDNRYIRAVGVPAL
GFSPMNRTPVLLHDHDERLHEAVFLRGVDIYTRLLPALASVPALPSD
```

or upload a fasta file: Keine ausgewählt

Settings:

Performe fast prediction: ☒ yes ☐ no

- ▶ <https://dataminer.informatik.tu-muenchen.de/~spoeri/>
- ▶ file or text field input
- ▶ option to speed prediction up by restricting hits to 6 best

Productive waste of time aka: Future Improvements

- ▶ Incorporate SVG output of results in webinterface
- ▶ Feature evaluation and improvement
- ▶ Investigate effect of homologs in dataset
- ▶ Different network architectures
- ▶ Other machine learning devices