JPred4: JNet training (v.2.3.1) details

Authors: Alexey Drozdetskiy and Geoffrey J. Barton

CONTENTS

Selection of Training and Testing Data

As for JPred3 training datasets, we based our training and testing data by selecting representatives from SCOP superfamily level rather than using a simple sequence identity cutoff. Since SCOP is a structure-based classification, this reduces the likelihood of trivially detectable sequence similarities remaining in the training data or between training and blind test datasets.

Initial SCOPe/ASTRAL set

We started with 1987 single representative sequences from each superfamily in SCOPe v.2.04.

Sequence filtering

Removing low resolution structures

306 domain sequences from proteins with >2.5 Angstrom resolution structures were removed to leave 1681 sequences.

Sequence Length Filter

We filter sequences <30 residues long since they are unlikely to represent globular protein domains. We also remove sequences of > 800 residues in length, though this is largely for historical reasons due to the time required in building PSIBLAST profiles. This step removed 27 domain sequences to leave 1654 sequences.

Remove multi-chain or fragment domains

20 domains made up of multiple chains were removed to leave 1634 sequences.

Check for pairwise sequence redundancy

The 1634 sequences were compared pairwise by the AMPS algorithm with 100 randomisations and a Z-score threshold of 6.5. No sequences were filtered out at this stage.

Missing DSSP information filter

We checked for sequences, which would have missing DSSP information for >9 consecutive residues and removed 110 such sequences, left with 1524 sequences at this stage.

Further filters

A further 17 sequences were excluded by making conservative decisions about inconsistencies between PDB, DSSP and ASTRAL file definitions for some structures. This left a total of 1507 domain sequences.

Training and blind-test sets

At this stage (1507 sequences) we separated the set into two: a training set with 1357(In the 2015 NAR paper, this is incorrectly shown as 1358) sequences and blind-test set with 150 sequences which was not used at all in training the JNet algorithm. We randomly iterated the selection of 150 sequences until the percentage content of alpha helix, beta sheet, and coil was within 1% between the blind-test and training sets. This step is essential to allow the accuracy reported on the blind-test set to be a fair assessment of the method.

Training sub-sets

We used 7 fold cross-validation for initial training of the JNet Neural Network. The 1357 sequences set was divided into 7 approximately equal sub-sets. In each of 7 rounds 1 sub-set was used for a test and the 6 others for training.

Finally, the Neural Network was trained on the full training set of 1357 sequences. This is the final version, which became JNet v.2.3.1.

At every step we checked that reported accuracies are similar.

Final training set

Note that during training runs 9 sequences failed to produce PSI-BLAST hits and these 9 sequences were removed. So, **the final training set contains 1348 sequences**. It is for this final set we provide download links with details for your tests, or alternative training:

- sequences in FASTA format,
- DSSP info*,
- JNet full results,
- PSI-BLAST profiles,
- a summary table with more statistics and link to full JPred result directories with all visualizations, links to Jalview, etc.

Final blind-test set

During the blind-test run 1 sequence failed to produce PSI-BLAST hits and was removed. So, **the final blind-test set contains 149** sequences. It is for this set we provide download links and all the details as described for the final training set above.

* Note on DSSP information

In JPred/JNet we use 3-state notation, which is created using the following rules applied to a more complete 8-state original DSSP information:

- Alpha helix: H → H
- Beta Sheet: E,B → E
- Other: everything else → '-'

Further reading:

- Drozdetskiy A, Cole C, Procter J & Barton GJ. JPred4: a protein secondary structure prediction server, *Nucleic Acids Res.* Web Server issue [doi:10.1093/nar/gkv332]
- http://www.compbio.dundee.ac.uk/jpred4 web-site, in particular "About", "F.A.Q.", "Help & Tutorials" sections