Research stay workplan

Michał Burdukiewicz

October 10, 2016

1 Workplan

Aim: improve performance of signalP for atypical signal peptides using a reduced amino acid alphabet.

Since signalP 5.0 is not yet available, I want to instead use signalP 4.1.

1. Generate large number of reduced amino acid alphabets. I want to use different physicochemical features that are associated with the unique architecture of signal peptides, for example several measures of hydrophobicity. The final sets of alphabets will be based on around 20 properties and I will build alphabets based on all combination of all sizes of this properties receiving 1,048,575 alphabets. Most of them will be identical, so the real number of unique amino acid alphabets will be lower.

2. Adjust signalP 4.1 for reduced alphabets. I don't have an access to signalP 4.1 source code, but it is possible, that it has hardcoded alphabet of 20 amino acids, so a bit of programming might be necessary.

3. Train new iterations of signalP 4.1 on signalP 4.1 data set, but using sequences written in reduced amino acid alphabets.

4. Choose the best-performing iterations of signalP 4.1 considering different criteria (mostly AUC and specificity).

5. Benchmark best-performing iterations of signalP 4.1 on external data set(s) of atypical signal peptides and compare with normal signalP 4.1 and signalP 3.0.

The analysis presented above is independent of the predictor, so when signalP 5.0 is ready, it may easily replace signalP 4.1.

2 Concerns

It is possible that reduction of the alphabet may improve the general performance of signal peptide prediction, but lower accuracy of cleavage sites prediction. Cleavage sites seem to require more defined motifs than whole signal

peptide and may need larger alphabets, possibly even the full amino acid alphabet.