Projekt lab notes by Ezeddin Al Hakim

2016

Project

Classifier for identification of signal peptides.

---

19 December 2016

We have described the classification problem by a Hidden Markov Model. In the next days, we will focus to implement HMM, whether we have time we will implement RNN.

20 December 2016

We have installed *hmmlearn* module, it is a set of algorithms for inference of Hidden Markov Models for Python. Then we implemented the first version of the classifier using hmmlearn. The classifier works as follows:

- Training a model λ_p using only positive data
- Training a model λ_n using only negative data
- Given a sequence of amino acids $X_{1:n}$, we will calculate the probability for both models (i.e. $P(X_{1:n}|\lambda_p)$ and $P(X_{1:n}|\lambda_n)$) and choosing the model with highest probability.

But classifier was not so good, we got 61 % accuracy, 100 % recal and 58 % precission.

21 december 2016

Today we have tried another method. We have trained a model λ_a using all data, positive and negative data. To classify signal peptides, we calculate/find the most likely sequence of hidden states $Z_{1:n}$ (n-, h-, c-, C- and other-regions), i.e. maximizes the conditional distribution $P(Z_{1:n}|X_{1:n}, \lambda_a)$. To classify if the sequence is a signal peptide, we finding the n-, h-, c-, and C-regions in the hidden states.

The classifier was much better, we got 92 % accuracy, 92 % recal and 91 % precission.

22 December 2016

Today we tried to evaluate our classifier on two proteom, human and mouse. But we had a few bugs in our classifier, e.g. it was unknown characters in the sequences of the two proteom.

27 December 2016

Today we succeeded to evaluate our classifier on the two proteom, and we got the follows results:

———————-

HUMAN
All 215929
All true 12931
All positive 17181
True positive 10626
True negative 196443
Precission 0.618473895582
Recal 0.821746191323
Accuracy on test data: 95.9%

———————-

MOUSE
All 124168
All true 7756
All positive 9966
True positive 6246
True negative 112692
Precission 0.626730885009
Recal 0.805312016503
Accuracy on test data: 95.79%

———————-

28 December 2016

Today we tested our classifier with different length of random sequences and 92 % of them classified as non-signal peptides. Here are some outputs from Controls:

———————-

Evaluated on random data with length 2:
All true 0
All positive 0
True positive 0
True negative 1000
Accuracy on test data: 100.0%

———————-

Evaluated on random data with length 100:

All true 0
All positive 110
True positive 0
True negative 890
Accuracy on test data: 89.0%
———————-
Evaluated on random data with length 10000:
All true 0
All positive 118
True positive 0
True negative 882
Accuracy on test data: 88.2%
———————-
