1. What is the general procedure when classifying data with support vector machines?

SVM allows to categorize data in two (by default) distinct classes – either in 2-dimensional plane, or multidimensional plane. Classification usually refers to supervised learning approach while clustering refers to unsupervised.

1. A training dataset is obtained/constructed, equally representative of two classes.
2. Each training data point is a vector defined by it's features and class.
3. The algorithm attempts to fit the regression function that maximizes the distance between the two least distinct data points(the support vectors). If data cannot be split on a two-dimensional plane, kernel method can be used to represent the data on the multidimensional "fature" plane on which the data points can be separated.
4. New (now unlabeled, but known) test data can be used to validate accuracy of the trained SVM.
5. SVM is used to classify unlabeled data according based on known features.

2. Define with your own words supervised and unsupervised learning and point out the difference(s). Give 2 example methods for each.

Many general methods can be used labeled and unlabeled data, with some adjustment.

Supervised: Data is labeled, clustered or grouped. You tell the algorithm what is the answer, and the best approach to represent the question is calculated.

Examples: labeled SVM; feed-forward ANN adjusting weights based on positive/negative results.

Unsupervised: Data is unlabeled and unclassified. The algorithm will attempt to group/cluster the data based on similarity pattern according to statistical principles.

Examples: PCA; hierarchical clustering algorithms.

3. What is cross-validation?

Training data can be partitioned into several categories to evaluate fitness of the model and prevent over-training. The categories will alternate in being training/test data, and the general consensus from all arrangements will be evaluated. Generally more splits or categories leads to higher accuracy, but optimal ratio of accuracy and specificity needs to be adjusted.

4. What does a line in any of these files correspond to?

Information about classification (positive/negative example), values based on feature or amino acid, organized by feature number.

5. What is the meaning of a -1 in the first column in the file train25_mini?

-1 means the row represents a negative class

6. What is the meaning of the 3:1 in the first line of train25_mini?

It means the third feature (D or Aspartate?) has a value of 1 as a negative weight/contribution for training purposes in this sequence.

7. Train an SVM model on train25_mini. Then test the performance of this model on test25. What accuracy did you get?

87%

8. Use the svm_model from question 7 and test it on train25_mini. What is the accuracy? Is this a good way of testing an SVM model?

90%. Not a good way – because it's biased towards training sample. But it reveals how the model divides the training set. If accuracy is close to 100%, it could be a sign of overfitting, but the current value does still not exclude overfitting.

9. Train an SVM model on a larger training data set, train25, and then test this model on the set test25. What accuracy did you get?

94%

10. Train an SVM model on train100 and test it on test100. What is the accuracy?

75%

11. Do you get a better classification by training and testing with the first 25 residues or with the first 100 residues? How would you explain this result?

First 25.
The first 25 residues provide more relevant and distinctive information for classification purposes. The first 100 residues contain more information – some of which is not relevant and includes more than just the distinctive feature – adding noise/ diluting the significance of identifying feature.

12. There are different kernels that can be used when creating an SVM model using svm_learn (see different svm_learn options by running svm_learn --help). The svm_learn flag for selecting a kernel is called -t. Which kernel is used by default? Which kernel gives the highest accuracy when using train25 for building an SVM model and test25 for testing the SVM model?

Linear by default. Linear and polynomial equally good at 94%

13. A sequence LOGO is generally created to compare the different positions in a multiple sequence alignment in terms of information content. The higher the letters at a certain position in the LOGO, the more informative or conserved this position is in terms of sequence evolution. You can create sequence LOGOs using the online tool WebLogo. Create two LOGOs, one for the sequences in the attached file signal_sequences.txt and one for the sequences in nonsignal_seqs.txt. Submit these images together with the report. Compare these two LOGOs, can you observe any differences?

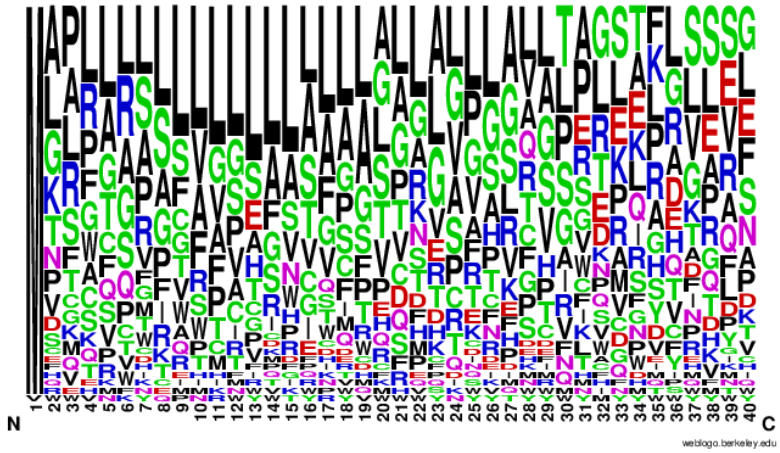


Illustration 1: Signal

Signal sequences appear to have much higher probability to begin with M, and to contain L or S residue/s close to the N terminus.

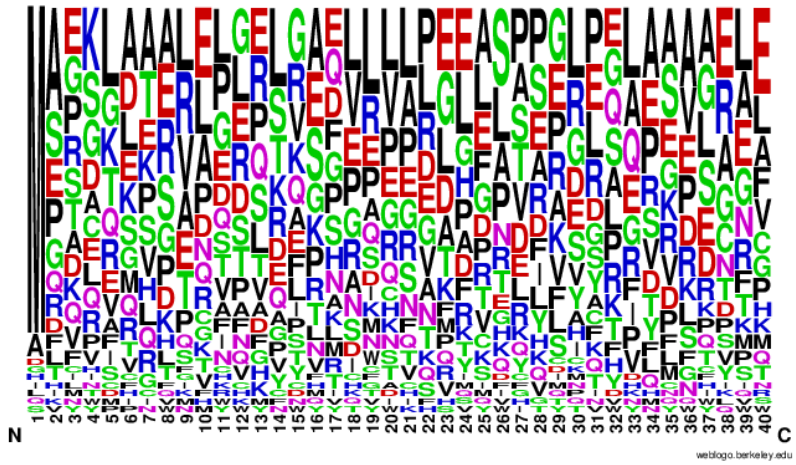
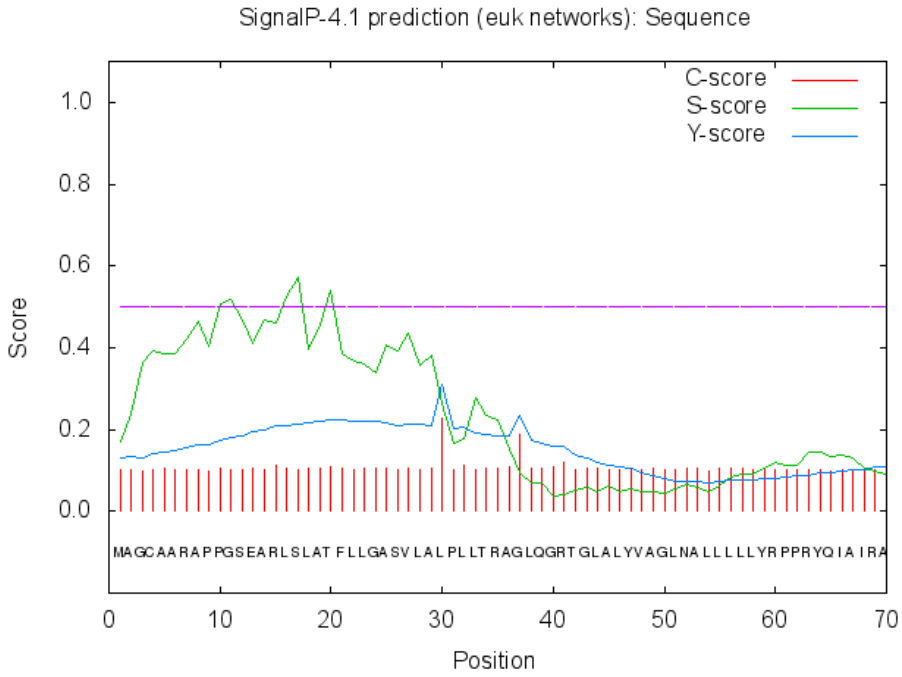


Illustration 2: Non-signal

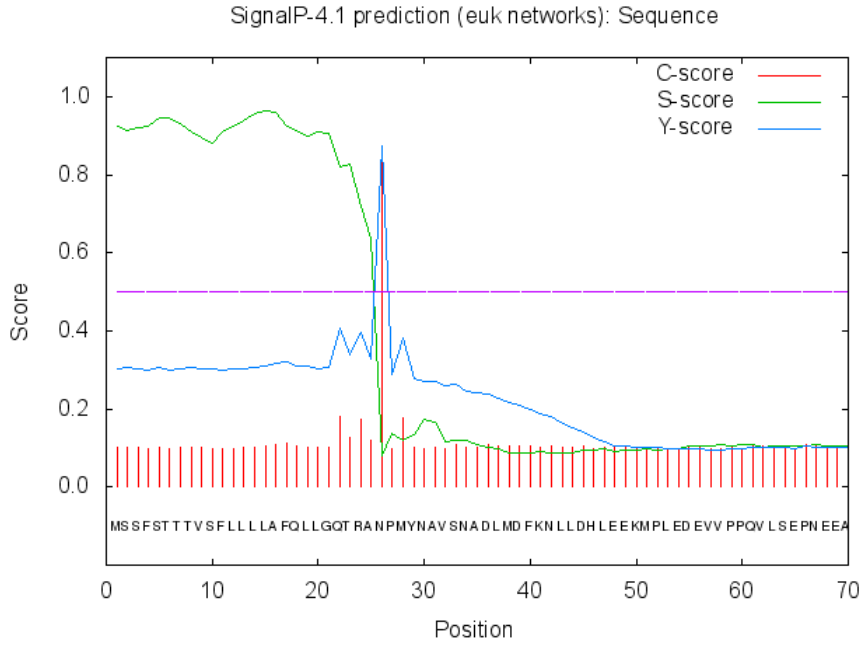
Non-signal peptides are much more likely to contain E, K near the N terminus, compared to signal.

14. There are various methods for predicting signal peptides in protein sequences. These tools have been trained on known signal peptide data and some of them perform really well. The tool SignalP is available online. Use it on the first two human proteins available in the file sequences.txt. Save the result plots of the SignalIP-NN and submit it together with the report.

All proteins in the file are likely of human origin according to Blast.



Sequence 1



Sequence 2

15. What do high and low S-scores indicate in the SignalP-NN result?

Signal peptide score – plotting signal peptide similarity over sequence length.
Overall low S score means less likely a signal peptide. But location also matters, for example sequence 2 appears to have high S score before the estimated cleavage site/C-score peak.

16. What are the D-scores for the two sequences?

1: 0.351
2: 0.885

17. The results from SignalP include something called cleavage sites. How do you interpret the term cleavage site in the context of signal peptides?

Signal peptides will be cleaved off from the main functional protein chain when the peptide is delivered to its destination. It is of interest for SignalP to help predict the signal peptide class and boundries