Signal Peptide Analysis Through Tensor Calculus

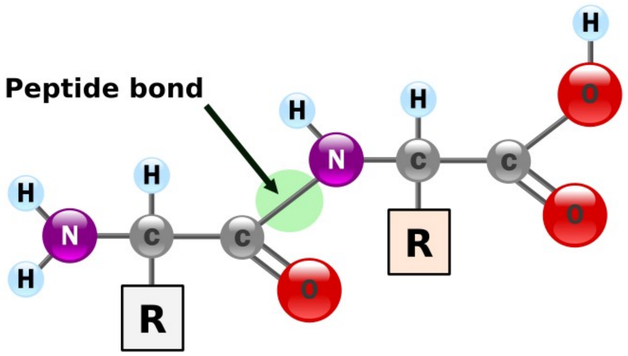
Alan Meeson, School of Computer Science

The goal of this project is to apply a Tensor LSA based classifier to the task of predicting signal peptide cleavage sites, and compare its performance to that of a classical LSA based classifier.

Background

Proteins

- A protein is made of one or more poly-peptide chains
- Amino acids are joined by covalent bonds to form poly-peptide chains.
- 20 different amino acids appear in the proteins considered in this project.
- The type and order of the amino acids in the poly-peptide chain contribute to determining the structure and function of the protein.
- All these amino acids are of similar structure, differentiated by their R groups.



Two amino acids joined by a covalent peptide bond.

Signal Peptides

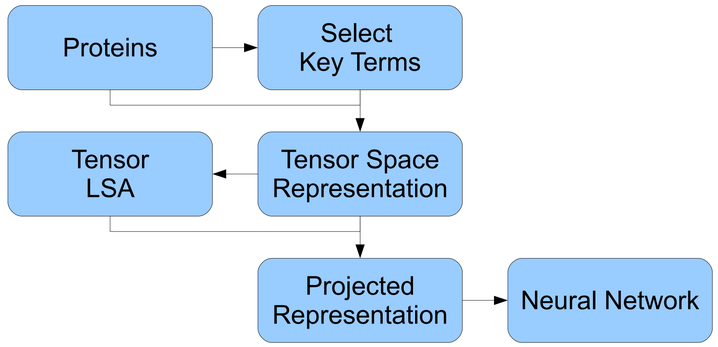
- They are a sequence of typically between 15 to 30 amino acids found at the start of certain proteins.
- They act as a 'postcode' directing the transport of the protein between and within cells.
- Once a protein has reached its destination the signal peptide is cleaved from the protein by the the enzyme signal peptidase.
- They can be useful in the creation of targeted protein based drugs.
- Due to its importance in the development of targeted protein based drugs, much research is being conducted into predictive analysis of proteins.[1] [2][3]

Method

Signal Peptide Cleavage sites are detected by classifying each point in the protein as either belonging to the signal peptide - thus being before the cleavage site - or belonging to the remainder of the protein, and thus occurring after the cleavage site.

Points are classified by considering a window of 20 amino acids on either side, and each amino acid in this neighbourhood is represented by its affinity to a set of key terms.

In order to allow the consideration this much information Tensor Latent Semantic Analysis is used to reduce the dimensionality and identify underlying structure. This allows the use of a standard neural network to perform classification.



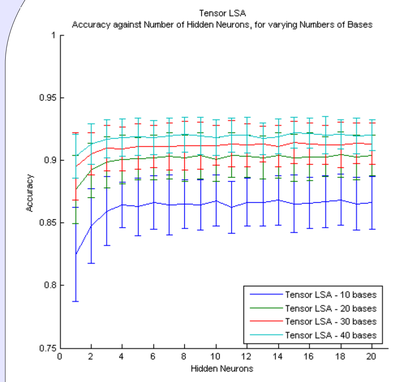
Tensor Latent Semantic Analysis

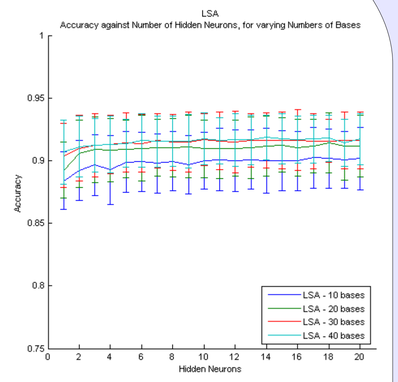
The Tensor Space Model & Tensor Latent Semantic Analysis were developed by Cai et al[4], based on the Vector Space Model and Latent Semantic Analysis, to combat the curse of dimensionality.

When Performing Latent Semantic Analysis on an N dimensional VSM representation the eigen-decomposition of an N by N matrix must be calculated, which can be very computationally intensive when the number of dimensions is high.

Tensor Latent Semantic Analysis handles this problem by using a Tensor Space Model, where each object is represented by an M by R matrix (where MR = N). These Matrices can be approximated by two vectors, and consequently bases can be found by performing the eigen-decomposition of two much smaller matrices (M by M, and R by R).

Results





- As seen in the graphs above the prediction accuracy grows with the number of hidden neurons until the 4th hidden neuron is added, after which it has little effect.
- Tensor LSA shows the best prediction accuracy when used with the 40 most important bases, with the accuracy being significantly better than the next most accurate.
- LSA shows the best prediction accuracy when used with the 40 most important bases, however the accuracy is not significantly better than the next most accurate.

	Mean Time	Standard Deviation
Tensor LSA	6064.9	610.4
LSA	16498.8	2678.5
P-Value	1.13E-09	

Times shown are measured in seconds and cover calculation of covariance matrices, eigen-decomposition, and calculation of basis importance.

Based on the validation results above the settings chosen for both methods were 4 hidden units and 40 bases. The tables below show the performance of the methods on a hold out test set using these settings.

	Mean Accuracy	Standard Deviation
Tensor LSA	0.9146	0.0027
LSA	0.9217	0.0011
P-Value	5.41E-07	

Conclusions

- In terms of prediction accuracy Tensor LSA is out performed slightly by LSA, and this difference is significant.
- In terms of computation time Tensor LSA significantly out performs LSA.

Acknowledgements

I would like to thank my supervisor Dr. Peter Tino for his support and guidance during this project, and the masters degree as a whole.
I would also like to thank Dr. Mikael Bodén for providing the dataset used in this project.

[1] Dyrløv Bendtsen, J.; Nielsen, H.; von Heijne, G. & Brunak, S. Improved Prediction of Signal Peptides: SignalP 3.0 Journal of Molecular Biology, 2004, 340, 783-795
[2] Nielsen, H. & Krogh, A. Prediction of signal peptides and signal anchors by a hidden Markov model In Proc. 6th Int. Conf. on Intelligent Systems for Molecular Biology, 1998, 6, 122-130
[3] Zhang, Z. & Henzel, W. J. Signal peptide prediction based on analysis of experimentally verified cleavage sites Protein Science, 2004, 13, 2819-2824
[4] Cai, D.; He, X. & Han, J. Tensor space model for document analysis SIGIR '06:
Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2006, 625-626