Computing the Language of Life: NLP Approaches to Feature Extraction for Protein Classification

Ananthan Nambiar[ab], Mark Hopkins[a] and Anna Ritz[c]

[a]Department of Mathematics, Reed College, Portland, OR 97202, USA

[b] Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

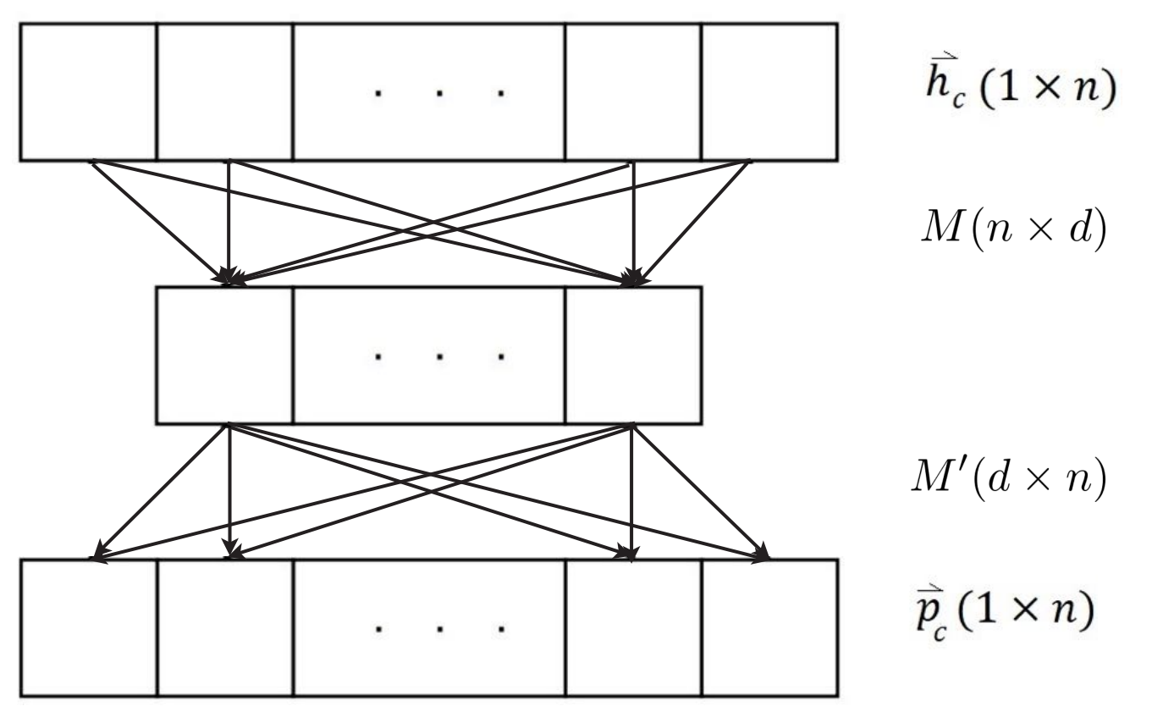
[c]Department of Biology, Reed College, Portland, OR 97202, USA

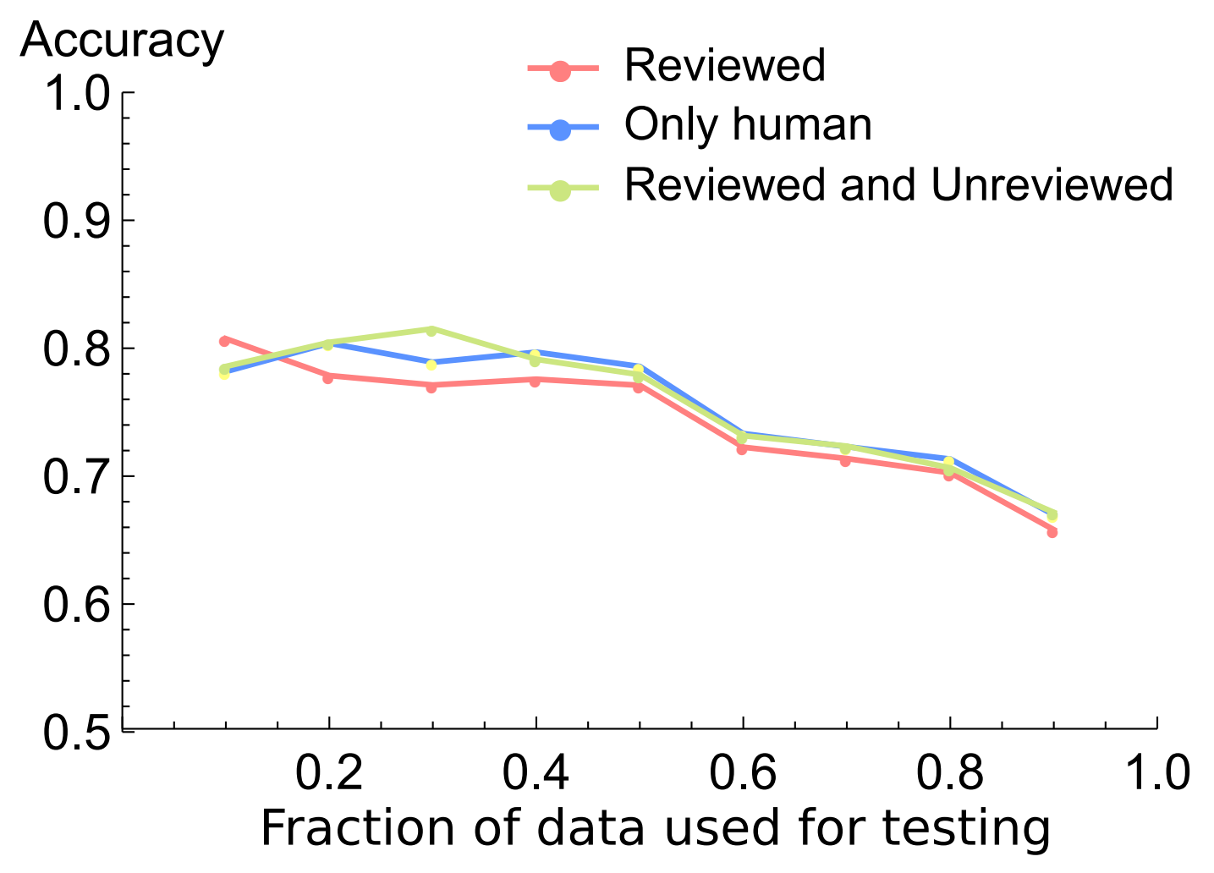
1. Introduction

We studied the effectiveness of several natural language processing methods (text embedding and convolutional neural networks), including existing methods such as ProtVec and DeepFam on protein family classification using two datasets from SwissProt and Clusters of Orthologous Groups (COGs).

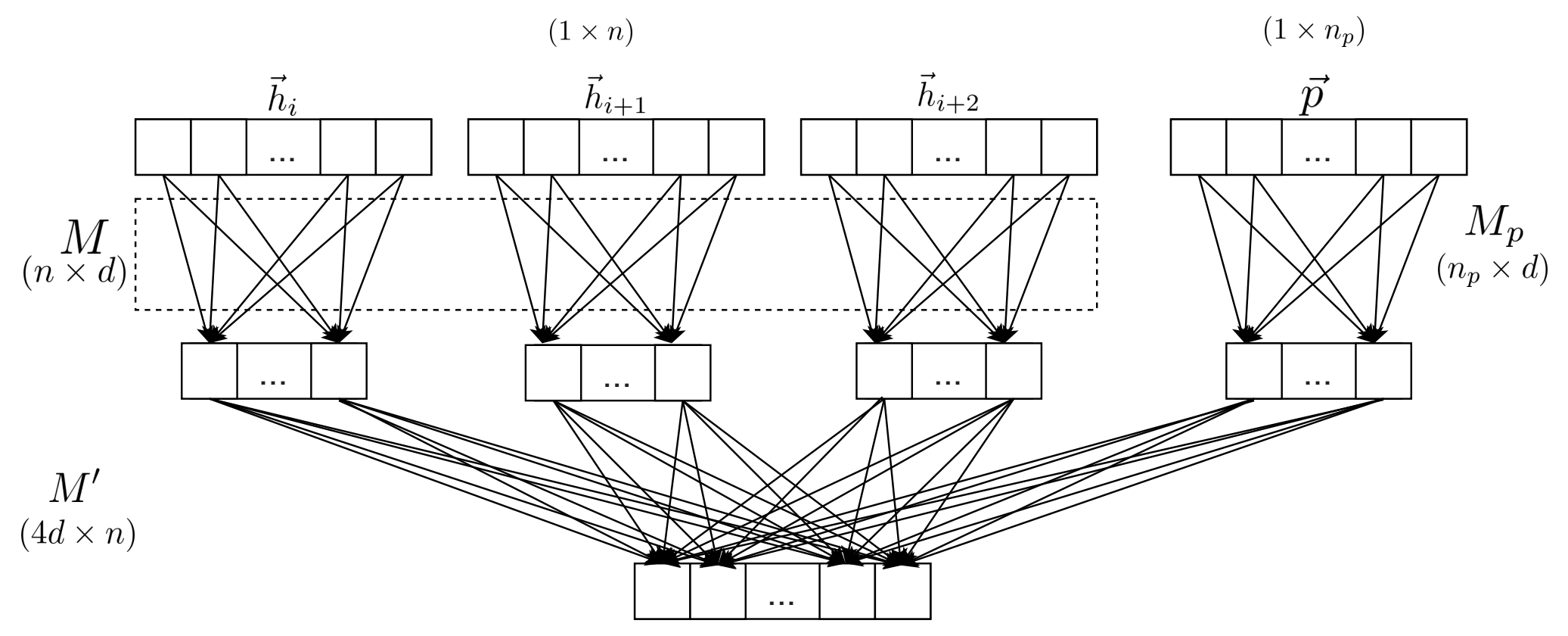
2. Text embeddings

Three methods based on text embedding were tested by converting proteins into sentences of trigrams. For example s_1 =AABRDAMEEAAM gives “AAB RDA MEE AAM", “ABR DAM EEA" and “BRD AME EAA". The vectors are then used by a logistic regression classifier to determine whether a protein belongs to a particular family. The first method, ProtVec [1] produces vector embeddings of triplets (or words) of amino acids and sums them up to produce a vector for a protein. The architecture this embedding is shown below with a graph that shows how the amount of training data has little impact on the model.



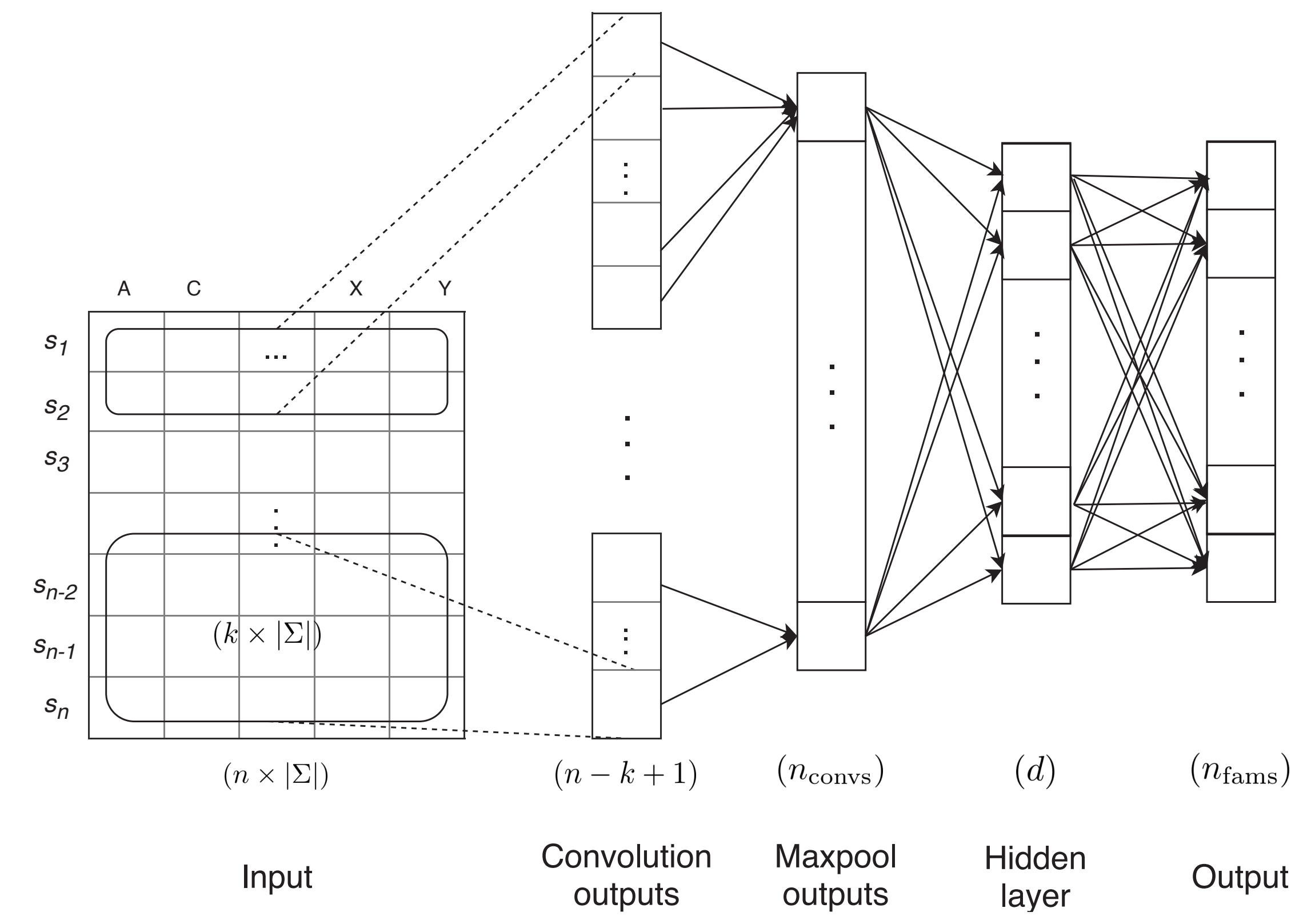


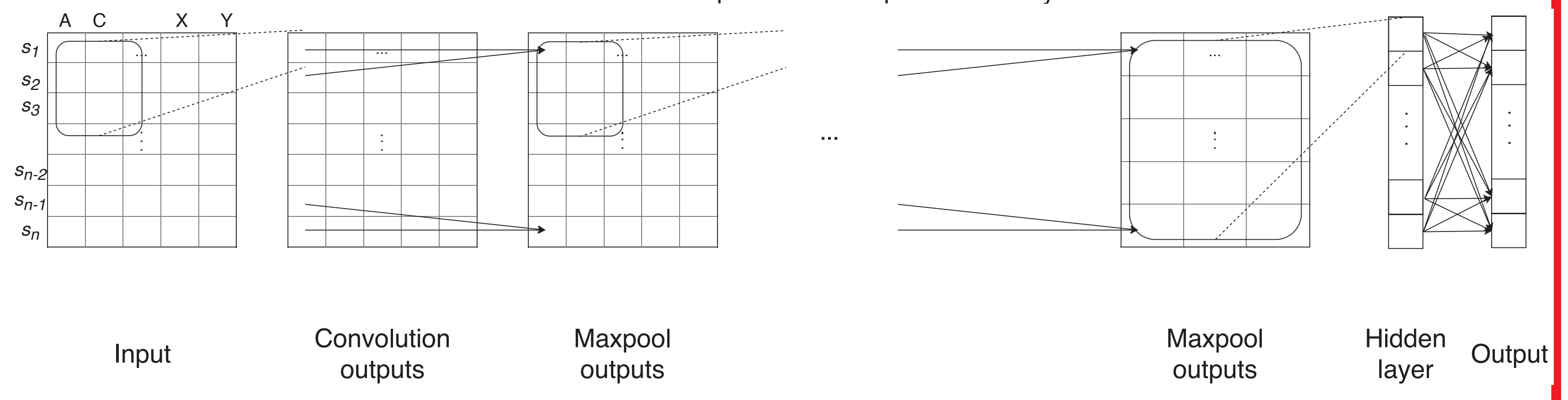
This lead to the idea that very little context information is extracted by ProtVec. We compared ProtVec to FreqVec, which produces a vector based on triplet frequency and ProtDocVec (shown below), which produces embeddings for the entire sequence. A method similar to ProtDocVec, known as seq2vec, has been independently developed [2].



3. Convolutional neural networks (CNNs)

There has been substantial work done in using CNNs for NLP [3] [4]. To study how CNNs extract features for protein classification, an established method known as DeepFam [5] (top) was compared to a more typical CNN (bottom).





DeepFam uses the information from the amino acid sequence at many different scales using convolution kernels of different sizes. However, the largest kernel only had a window width of 23 amino acids extracting only low level features. The features are then used together to predict the function of the protein. The second, typical network, attempts to extract higher level features with several layers of convolutions.

4. Results and Conclusion

The accuracy of protein family classification using each of the five methods is shown in the table below for both the SwissProt and COGs datasets.

Experiment data: UniProt proteins, COGs proteins
Classification: 1 vs all (logistic), multiclass (CNNs)
Cross validation: 70% train, 30% test

Method	Accuracy (UniProt/COGs)
ProtVec-logistic	0.89/0.81
ProtFreqVec-logistic	0.98/0.98
ProtDocVec-logistic	0.98/0.96
DeepFam	0.96/0.95
Simple CNN	0.72/0.65

An observation common to many of these methods is that low level features can be powerful enough to classify proteins with reasonable accuracy. This suggests that perhaps the protein family classification is not as difficult as expected. However, there is still work to be done before this can be said with certainty. For example, it is thought that while proteins that have a common near ancestor are easy to classify, those that do not share a common ancestor but have similar functions are more difficult to classify [6]. This hypothesis has yet to be tested using deep learning based methods.

5. Acknowledgements

Conference travel was supported by the Reed College Opportunity Grant and the National Science Foundation # 1750981. AN would like to thank Mark Bedau, Norman Packard and the Reed College Artifical Life Lab for insightful discussions and Desiree Odgers for inspiring the idea of taking a linguistic approach to a biological problem.

[1] E. Asgari, M. R. K. Mofrad, *PLOS ONE* **10**, 1 (2015).

[2] D. Kimothi, A. Soni, P. Biyani, J. M. Hogan (2016).

[3] R. Collobert, J. Weston, ICML ’08 (2008).

[4] N. Kalchbrenner, E. Grefenstette, P. Blunsom, *ACL* (2014).

[5] M. Oh, S. Seo, S. Kim, Y. Park, *Bioinformatics* **34**, i254 (2018).

[6] C. A. Orengo, J. M. Thornton, *Annual Review of Biochemistry* **74**, 867 (2005). PMID: 15954844.

Low level features can be powerful enough to classify proteins with reasonable accuracy