Protein Word Detection using Text Segmentation Techniques

G.Devi
Department of CSE
IIT Madras
Chennai-600036, India
gdevi@cse.iitm.ac.in

Ashish V. Tendulkar
Google Inc.,
Hyderabad-500084, India
ashishvt@google.com

Sutanu Chakraborti
Department of CSE
IIT Madras
Chennai-600036, India
sutanuc@cse.iitm.ac.in

Abstract

Literature in Molecular Biology is abundant with linguistic metaphors. There have been works in the past that attempt to draw parallels between linguistics and biology, driven by the fundamental premise that proteins have a language of their own. Since word detection is crucial to the decipherment of any unknown language, we attempt to establish a problem mapping from natural language text to protein sequences at the level of words. Towards this end, we explore the use of an unsupervised text segmentation algorithm to the task of extracting "biological words" from protein sequences. In particular, we demonstrate the effectiveness of using domain knowledge to complement data driven approaches in the text segmentation task, as well as in its biological counterpart. We also propose a novel extrinsic evaluation measure for protein words through protein family classification.

1 Introduction

Research works in the field of Protein Linguistics (Searls, 2002) are largely based on the underlying hypothesis that proteins have a language of their own. However, modeling of protein molecules using linguistic approaches is yet to be explored in depth. This might be due to the structural complexities inherent to protein molecules. Instead of resorting to purely wet lab experiments, we propose to make use of the abundant data available in the form of protein sequences together with knowledge from domain experts to model the protein language. From a linguistic point of view, the first step in deciphering an unknown language will be to identify the independent lexical units or words of the language. This motivates our current attempt to establish a problem mapping from natural language text to protein sequences at the level of words. Towards this end, we explore the use of an unsupervised word segmentation algorithm to the task of extracting "biological words" from protein sequences.

Many unsupervised word segmentation algorithms use compression based techniques ((Chen, 2013), (Hewlett and Cohen, 2011), (Zhikov et al., 2010), (Argamon et al., 2004), (Kityz and Wilksz, 1999)) and are largely centred around the principle of Minimum Description Length (MDL). We use the MDL based segmentation algorithm described in (Kityz and Wilksz, 1999) which makes use of the repeating subsequences present within text corpus to compress it. It is found that the segments generated by this algorithm exhibit close resemblances to words of English language. There are also other non-compression based unsupervised word segmentation and morphology induction algorithms in literature ((Mochihashi et al., 2009), (Hammarström and Borin, 2011), (Soricut and Och, 2015)). However, in this context of protein sequence analysis, we have chosen to use MDL based unsupervised segmentation because it resembles closely the first natural attempt of a linguist in identifying words of an unknown language i.e. looking for repeating subsequences as candidates for words.

As we do not have access to ground-truth knowledge about protein words, we propose to use a novel extrinsic evaluation measure based on protein family classification. SCOPe is an extended database of SCOP hierarchy (Murzin et al., 1995) which classifies protein domains based on the structural and sequence similarities. We have proposed a MDL based classifier for the task of

Proceedings of the BioNLP 2017 workshop, pages 238–246,
Vancouver, Canada, August 4, 2017.

automatic SCOPe prediction. The performance of this classifier is used as an extrinsic measure of the quality of protein segments.

Finally, the MDL based word segmentation used in (Kityz and Wilksz, 1999) is purely data driven and does not have access to any domain-specific knowledge source. We propose that constraints based on domain knowledge can be profitably used to improve the performance of segmentation algorithms. In English, we use constraints based on pronounceability rules to improve word segmentation. In protein segmentation, we use knowledge of SCOPe Class labels (Fox et al., 2014) to impose constraints. In both cases, constraints based on domain knowledge are seen to improve the segmentation quality.

To summarize, the main contributions of our work are the following :

1. We attempt to establish a mapping from protein sequences to language at the level of words which is a vital step in the linguistic approach to protein language decoding. Towards this end, we explore the use of an unsupervised text segmentation algorithm to the task of extracting "biological words" from protein sequences.

2. We propose a novel extrinsic evaluation measure for protein words via protein family classification.

3. We demonstrate the effectiveness of using domain knowledge to complement data driven approaches in the text segmentation task, as well as in its biological counterpart.

2 Related Work

Protein Linguistics (Searls, 2002) is the study of applying linguistic approaches to understand the structure and function of protein molecules. Research in the field of Protein Linguistics is largely based on the underlying assumption that proteins have a language of their own. David Searls draws many analogies between Linguistics and Molecular Biology to show how a linguistic metaphor can be seen interwoven into many problems of Molecular Biology. The fundamental analogy is that the 20 amino acids of proteins and 4 nucleotides of genes are analogous to the 26 letters in English alphabet.

Literature is abundant with parallels between language and biology (Bralley, 1996; Searls, 2002; Atkinson and Gray, 2005; Gimona, 2006; Tendulkar and Chakraborti, 2013). There are striking similarities between the structure of a protein molecule and a sentence in a Natural Language text some of which have been highlighted in Figure 1.

Gimona (2006) presents an excellent discussion on linguistics-based protein annotation and raises the interesting question of whether compositional semantics could improve our understanding of protein organization and functional plasticity. Tendulkar and Chakraborti (2013) also have drawn many parallels between biology and linguistics.

The wide gap between available primary sequences and their three dimensional structures leads to the thought that the current protein structure prediction methods might struggle due to lack of understanding of the folding code from protein sequence. If biological sequences are analogous to strings generated from a specific but unknown language, then it will be useful to find the rules of the unknown language. And, word identification is fundamental to the task of learning rules of an unknown language.

Motomura et. al ((2012),(2013)) use a frequency based linguistic approach to protein decoding and design. They call the short consequent sequences (SCS) present in protein sequences as words and use availability scores to assess the biological usage bias of SCS. Our approach of using MDL for segmentation is interesting in that it does not require prior fixing of word length as in (Motomura et al., 2012), (Motomura et al., 2013).

3 Word Segmentation

Word is defined as a single distinct conceptual unit of language, comprising inflected and variant forms[1]. In English, though space acts as a good approximation for word delimiter, proper nouns like *New York* or phrases like *once in a blue moon* make sense only when taken as a single unit. Therefore, space is not a good choice for delimiting atomic units of meaning.

Imagine a corpus of English text with spaces and other delimiters removed. Now, word segmentation is the problem of dividing a continuous piece of text into meaningful units. For example, imagine a piece of text in English with delimiters removed such as 'BIRDONTHETREE'. The contin-

[1]https://en.oxforddictionaries.com/definition/word

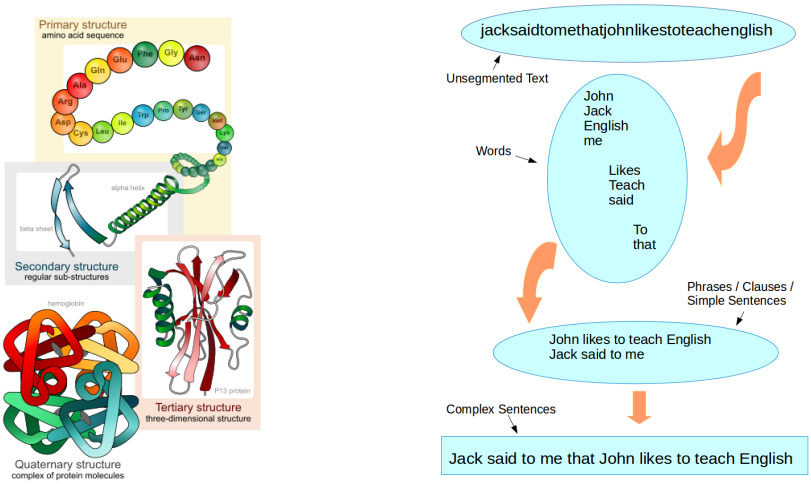

Figure 1: Structural levels in a Protein Molecule [Image source : (Wikipedia, 2017)] vs. Natural Language Sentence

uous text can be segmented into four meaningful units as 'BIRD','ON','THE','TREE'. Analogously, we define protein segmentation as the problem of dividing the amino acid sequence of a protein molecule into biologically meaningful segments. For example, the toy protein sequence 'MATGQKLMRAIRVFEFGGPEVLKLQSDVVVPVPQSHQ' can consist of three segments 'MATGQKLMRAIR', 'VFEFGGPEV', 'LKLQSDVVVPVPQSHQ'. For our work, we assume that the word segmentation algorithm does not have knowledge about English lexicon. The significance of this assumption can be understood in the context of protein segmentation. Since the ground truth about words in protein language is not known, we consider the problem of protein segmentation to be analogous to unsupervised word segmentation in English.

We begin this section by explaining why MDL can be a good model selection principle for learning words followed by description of the algorithm used and results obtained on Brown corpus.

3.1 MDL for Segmentation

According to the principle of Minimum Description Length (MDL),

Data compression → Learning

Any regularity present in data can be used to compress the data which can also be seen as learning of a model underlying the data (Grünwald, 2005). In an unsegmented text corpus, the repetition of words creates statistical regularities. Therefore, the key idea behind using MDL for word segmentation is that we can learn word-like segments by compressing the text corpus.

Description Length (DL) of a corpus X is defined as the number of bits needed to encode it using Shannon Fano coding [(Shannon, 2001),(Kityz and Wilksz, 1999)] and is expressed as given below.

$$DL(X) = -\sum_{x \epsilon V} c(x) \log \frac{c(x)}{|X|} \quad (1)$$

where, V is the language vocabulary, $c(x)$ is the frequency of word x in the given corpus and $|X|$ is total number of words in X.

As an unsupervised learning algorithm does not have access to language lexicon, the initial DL of the corpus is calculated by using the language alphabet as its vocabulary. When the algorithm learns word-like segments, we can expect the DL of corpus to get reduced. According to MDL, the segmentation model that best minimizes the combined description length of *data* + *model* (i.e. *corpus*+ *vocabulary*) is the best approximation of the underlying word segmentation.

An exponential number of candidate segmentations is possible for a piece of unsegmented text. For example, some candidate segmentations for the text 'BIRDONTHETREE' are given below.

'B','IRDONTHETREE'
'BI','RD','ONTHET','R','E','E'
'B','I','R','D','ONTHET','REE'
'BIR','D','ONT','HE','TREE'
'BIRDON','THE','TREE'
'BIRD','ON','THETREE'

(Kityz and Wilksz, 1999) define a goodness measure called *Description Length Gain* (DLG) to quantify the compression effect produced by a candidate segmentation. DLG of a candidate segmentation is equal to the sum of DLGs of individual segments within it. DLG of a segment s is defined as the reduction in description length achieved by retaining this segment as a single lexical unit while aDLG stands for the average description length gain as given below.

$$DLG(s) = DL(X) - DL(X[r \rightarrow s] \oplus s)$$
$$aDLG(s) = \frac{DLG(s)}{c(s)}$$

where, $X[r \rightarrow s]$ represents the new corpus obtained by replacing all occurrences of the segment s by a single token r, $c(s)$ is the frequency of the segment s in corpus and $\oplus$ represents the concatenation of two strings with a delimiter in between. This is necessary because MDL minimizes the combined DL of corpus and vocabulary. (Kityz and Wilksz, 1999) uses Viterbi algorithm to find the optimal segmentation of a corpus. Time complexity of the algorithm is $O(mn)$ where n is the length of the corpus and m is the maximal word length.

3.2 Imposing Language Constraints

MDL based algorithm as described in (Kityz and Wilksz, 1999) performs uninformed search through the space of word segmentations. We propose to improve the performance of unsupervised algorithm by introducing constraints based on domain knowledge. These constraints help to improve the word-like quality of the MDL segments. For example, in English domain, we have used the following language constraints, mainly inspired by the fact that legal English words are pronounceable.

1. Every legal English word has at least one vowel in it
2. There cannot be three consecutive consonants in the word beginning except when the first consonant is 's'
3. Some word beginnings are impossible. For example, 'db', 'km', 'lp', 'mp', 'ns', 'ms', 'td', 'kd', 'md', 'ld', 'bd', 'cd', 'fd', 'gd', 'hd', 'jd', 'nd', 'pd', 'qd', 'rd', 'sd', 'vd', 'wd', 'xd', 'yd', 'zd'
4. Bigrams having high probability of occurrence at word boundaries are obtained apriori from a knowledge base to facilitate splitting of long segments

3.3 MDL Segmentation of Brown Corpus

The goal of our experiments is twofold. First, we apply an MDL based algorithm to identify word boundaries. Second, we use constraints based on domain knowledge to further constrain the search space and thereby improving the quality of segments.

The following is a sample input text from Brown corpus (Francis and Kucera, 1979) used in our experiment.

> *implementationofgeorgiasautomobiletitlelaw*
> *wasalsorecommendedbytheoutgoingjury*
> *iturgedthatthenextlegislatureprovideenab*
> *lingfundsandresettheeffectivedatesothata*
> *norderlyimplementationofthelawmaybeeffect*

The output segmentation obtained after applying MDL algorithm is given below. It can be seen that the segments identified by the MDL algorithm are close to the actual words of English language.

> *implementationof georgias automobile title l a w wasalso recommend edbythe outgoing jury i tur g edthat thenext legislature provide enabling funds andre s et theeffective d ate sothat anorderly implementationof thelaw maybe effect ed*

The segments generated by MDL are improved by applying the language constraints listed in previous section. Sample output is shown below. We can observe the effect of constraints on segments, for example, [l][a][w] is merged into [law] ; [d][ate] is merged into [date].

> *implementationof georgias automobile title law wasalso recommend edbythe outgoing jury i tur ged that thenext legislature provide enabling funds andre set theeffective date sothat anorderly implementationof thelaw maybe effect ed*

Segmentation results are evaluated by averaging the precision and recall over multiple random samples of Brown Corpus. A segment is declared as

Algorithm	Precision	Recall
MDL (Kityz and Wilksz, 1999)	79.24	34.36
MDL + Constraints	82.57	41.06

Table 1: Boundary Detection by MDL Segmentation

Algorithm	Precision	Recall
MDL(Kityz and Wilksz, 1999)	39.81	17.26
MDL + Constraints	52.94	26.36

Table 2: Word Detection by MDL Segmentation

a correct word only if both the starting and ending boundaries are identified correctly by the segmentation algorithm. Word precision and word recall are defined as follows.

$$\text{Word Precision} = \frac{\text{No. of correct segments}}{\text{Total no. of segments}}$$

$$\text{Word Recall} = \frac{\text{No. of correct segments}}{\text{Total no. of words in corpus}}$$

Boundary precision and boundary recall are defined as follows.

$$\text{Boundary Precision} = \frac{\text{\# correct segment boundaries}}{\text{\# segment boundaries}}$$

$$\text{Boundary Recall} = \frac{\text{\# correct segment boundaries}}{\text{\# word boundaries}}$$

The performance of our learning algorithm averaged over 10 samples of size 10,000 characters (from random indices in Brown corpus) is shown in Tables 1 and 2. The reported results are in line with our proposed hypothesis that domain constraints help in improving the performance of unsupervised MDL segmentation.

4 Protein Segmentation

In this section, we discuss our experiments in protein domain. Choice of protein corpus is very critical to the success of MDL based segmentation. If we look at the problem of corpus selection from a language perspective, we know that similar documents will share more words in common than dissimilar documents. Hence, we have chosen our corpus from databases of protein families like SCOPe and PROSITE. We believe that protein sequences performing similar functions will have similar words.

4.1 Qualitative Analysis

The objective of our experiments on PROSITE database (Sigrist et al., 2012) is to qualitatively analyse the protein segments. It can be observed that within a protein family, some regions of the protein sequences have been better conserved than others during evolution. These conserved regions are found to be important for realizing the protein function and/or for the maintenance of its three dimensional structure. As part of our study, we examined if the MDL segments are able to capture the conserved residues represented by PROSITE patterns.

MDL segmentation algorithm was applied to 15 randomly chosen PROSITE families containing varying number of protein sequences. [2] Within a PROSITE family, some sequences get compressed more than others. An interesting observation is that the less compressed sequences are those that have evolved over time and hence have low sequence similarity with other members of the protein family. But, they have the conserved residues intact and MDL segmentation algorithm is able to capture those conserved residues.

For example, consider the PROSITE pattern [3] for Amidase enzyme (PS00571) G-[GAV]-S-[GS](2)-G-x-[GSAE]-[GSAVYCT]-x-[LIVMT]- [GSA]-x(6)-[GSAT]-x- [GA]-x-[DE]-x-[GA]-x-S- [LIVM]-R-x-P-[GSACTL] . The symbol 'x' in a PROSITE pattern is used for a position where any amino acid is accepted. 'x(6)' stands for a chain of five amino acids of any type. For patterns with long chains of x, MDL algorithm captures the conserved regions as a series of adjacent segments. For example, in the protein sequence with UniProtKB id O00519, the conserved residues and MDL segments are shown in Figure 2.

As another example, consider the family PS00319 with pattern G-[VT]-[EK]-[FY]-V-C-C-P . This PROSITE pattern is short and does not contain any 'x'. In such cases, the conserved residues can get captured accurately by MDL segments. The protein sequence with UniProtKB id P14599 has less sequence similarity but its conserved residues GVEFVCCP are captured exactly in a single MDL segment. We also studied the distribution of segment lengths among the PROSITE families. A single corpus was created combining the sequences from

[2] The output segments are available at https://1drv.ms/f/s!AnQHeUjduCq0ae9rWhuoybZoA-U

[3] A PROSITE pattern like [AC]-x-V-x(4)-AV is to be translated as: [Ala or Cys]-any-Val-any-any-any-any-Ala-Val

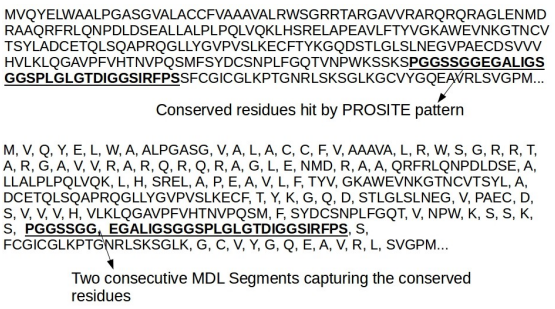

Figure 2: Conserved residues and MDL segments of a protein sequence (UniProtKB id O00519) in PROSITE family PS00571

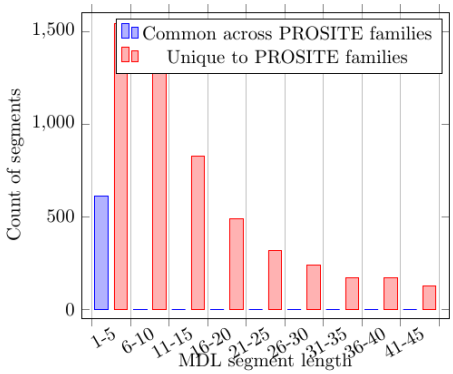

Figure 3: Distribution of MDL segment lengths among PROSITE families PS00319, PS00460, PS00488, PS00806 and PS00818

5 randomly chosen PROSITE families and the distribution of segment lengths is shown in Figure 3. Protein segments that were common among the families were typically four or five amino acids in length. However, within each individual family there were longer segments unique to that family. Very long segments (length >15) are formed when the corpus contains many sequences with high sequence similarities.

4.2 Quantitative Analysis

Unlike in English language, we do not have access to ground truth about words in proteins. Hence, we propose to use a novel extrinsic evaluation measure based on protein family classification. We describe a compression based classifier that uses the MDL segments (envisaged as words in proteins) for SCOPe predictions.The performance of the MDL based classifier on SCOPe predictions is used as an extrinsic evaluation measure of protein segments.

4.2.1 MDL based Classifier

Suppose we want to classify a protein sequence p into one of k protein families, the MDL based classifier is given by,

$$\text{family (p)} = \underset{\text{family}}{\operatorname{argmax}}\, DLG(p, \text{family}_{1\ldots k}) \quad (2)$$

where DLG(p,family$_i$) is the measure of the compression effect produced by protein sequence p in the protein corpus of family$_i$. We hypothesize that a protein sequence will be compressed more by the protein family it belongs to, because of the presence of similar words among the same family members.

Experimental Setup The dataset used for protein classification is ASTRAL Compendium (Chandonia et al., 2004). It contains protein domain sequences for domains classified by the SCOPe hierarchy. ASTRAL 95 subset based on SCOPe v2.05 is used as training corpus and the test set is created by accumulating the protein domain sequences that were newly added in SCOPe v2.06. Performance of the MDL classifier is discussed in four SCOPe levels - *Class, Fold, Superfamily and Family*. At all levels, we consider only the protein domains belonging to four SCOPe *classes* A,B,C and D representing *All Alpha, All Beta, Alpha+Beta, Alpha/Beta* respectively. The blind test set contains a total of 4821 protein domain sequences.

SCOPe classification poses the problem of *class imbalance* due to the non-uniform distribution of domains among different classes at all SCOPe levels. Due to this problem, we use macro precision and macro recall (Yang, 1999) as performance measures and are given by the below equations.

$$Precision_{macro} = \frac{1}{q}\sum_{i=1}^{q}\frac{TP_i}{TP_i + FP_i} \quad (3)$$

$$Recall_{macro} = \frac{1}{q}\sum_{i=1}^{q}\frac{TP_i}{TP_i + FN_i} \quad (4)$$

4.2.2 Performance of MDL Classifier

Class Prediction Out of 4821 domain sequences in the test data, the MDL classifier abstains from prediction for 71 sequences due to multiple classes giving the same measure of compression. The MDL Classifier achieves a macro precision of 75.64% and macro recall of 69.63% in *Class* prediction.

SCOPe level	Macro Average Precision	Macro Average Recall
Class	75.64	69.63
Fold	60.59	45.08
Super family	56.65	43.73
Family	43.25	37.7

Table 3: Performance of MDL Classifier in SCOPe Prediction

SCOPe level	Weighted Average Precision	Weighted Average Recall
Class	76.38	69.77
Fold	**81.49**	49.25
Super family	72.80	48.23
Family	45.02	35.85

Table 4: Performance of MDL Classifier in SCOPe Prediction - Weighted Measures

Fold Prediction SCOPe v2.05 contains a total of 1208 *folds* out of which 991 folds belong to *classes* A,B,C and D. The distribution of protein sequences among the *folds* is non-uniform ranging from 1 to 2254 sequences with 250 *folds* containing only one sequence. MDL Classifier achieves a macro precision of 60.59% and macro recall of 45.08% in *fold* classification.

Impact of Corpus Size The number of protein domains per class decreases greatly down the SCOPe hierarchy. The *folds* (or *families, superfamilies*) that have very few sequences should have less contribution in the overall prediction accuracy. We weighted the macro measures based on the number of instances which resulted in the weighted averages reported in Table 4. The MDL classifier achieves a weighted macro precision of 81.49% in SCOPe *fold* prediction which is higher than the precision at any other level. This observation highlights the quality of protein segments generated by MDL algorithm. It is also important to note that *fold* prediction is an important sub task of protein structure prediction just as how word detection is crucial to understanding the meaning of a sentence.

4.3 MDL Classifier as a Filter

The *folds* which are closer to each other in the SCOPe hierarchy tend to compress protein sequences almost equally. Instead of returning a single *fold* giving maximum compression, if the MDL classifier returns the top-k candidates, then we can reduce the search space for manual or high cost inspections. We define utility of the MDL classifier when used as a filter as given below.

$$\text{Utility} = \frac{\text{No. of predictions where correct fold is in top-k list}}{\text{Total no. of predictions}}$$

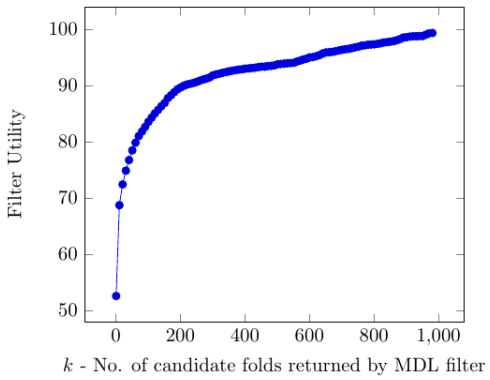

Figure 4: Variation of Filter Utility with Filter Size k

Figure 4 shows the k versus utility on test data. It can be seen from the graph that at k=400 (which is approximately 33% of the total number of folds), top-k predictions are able to give 93% utility. In other words, in 93% of the test sequences, MDL filter can be used to achieve nearly 67% reduction in the search space of 1208 folds.

4.4 Impact of Constraints based on Domain Knowledge

Similar to experiments in English domain, the MDL algorithm on protein dataset can also be enhanced by including constraints from protein domain knowledge. For example, in a protein molecule, hydrophobic amino acids are likely to be found in the interior, whereas hydrophilic amino acids are more likely to be in contact with the aqueous environment. This information can be used to introduce checks on allowable amino acids at the beginning and end of protein segments. Unlike in English, identifying constraints based on protein domain knowledge is difficult because there are no lexicon or protein language rules readily available. Domain expertise is needed for getting explicit constraints.

As proof of concept, we use the SCOPe *class* labels of protein sequences as domain knowledge and study its impact on the utility of the MDL filter. After introducing *class* knowledge, MDL filter achieves an utility of 93% at k=100, i.e., in 93%

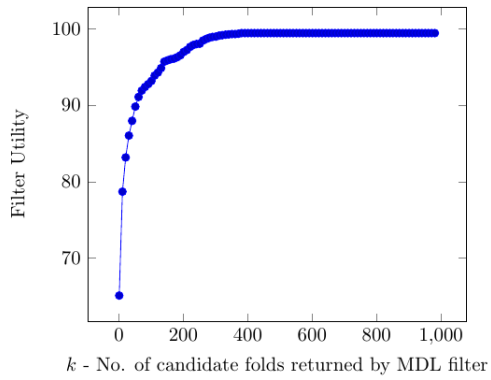

Figure 5: Variation of Filter Utility with Filter Size k after adding constraints based on SCOPe Class labels

of the test sequences, MDL filter can be used to achieve nearly 90% reduction in the search space of 1208 folds. In the absence of *class* knowledge, the same filter utility was obtained at k=400 which is only 67% reduction of search space (Figure 5). Through this experiment, we emphasize that appropriate domain knowledge can help in improving the quality of word segmentation in protein sequences. Such domain knowledge could be imposed in the form of constraints during unsupervised learning of protein words. We would like to emphasize the fact that introducing domain knowledge in the form of class labels as in supervised or semi-supervised learning frameworks may not be appropriate in protein sequences due to our current ignorance of the true protein words.

5 Discussion

In the words of Jones and Pevzner (Jones and Pevzner, 2004), "It stands to reason that if a word occurs considerably more frequently than expected, then it is more likely to be some sort of 'signal' and it is crucially important to figure out the biological meaning of the signal". In this paper, we have proposed protein segments obtained from MDL segmentation as the signals to be decoded.

As part of our future work, we would like to study the performance of SCS words (Motomura et al., 2012), (Motomura et al., 2013) in protein family classification and compare it against MDL words; We would also like to measure the availability scores of MDL segments. It may also be insightful to study the co-occurrence matrix of MDL segments.

6 Conclusion

Given the abundance of unlabelled data, data driven approaches have witnessed significant success over the last decade in several tasks in vision, language and speech. Inspired by the correspondence between biological and linguistic tasks at various levels of abstraction as revealed by the study of Protein Linguistics, it is only natural that there would be a propensity to extend such approaches to several tasks in Computational Biology. A linguist already knows a lot about language however, and a biologist knows lot about biology; so, it does make sense to incorporate what they already know to constrain the hypothesis space of a machine learner, rather than make the learner rediscover what the experts already know. The latter option is not only demanding in terms of data and computational resources, it may need us to solve riddles we just do not have answers to. Classifying a piece of text as humorous or otherwise is hard at the state of the art; there are far too many interactions between variables than we can model, not only do the words interact between them, they also interact with the mental model of the person reading the joke. It stretches our wildest imaginations to think of a purely bottom up Deep Learner that is deprived of common-sense and world knowledge to learn such end-to-end mappings reliably by looking at data alone. The same is true in biological domains where non-linear interactions between a large number of functional units make macro-properties "emerge" out of interactions between individual functional units. We feel that a realistic route is one where top down (knowledge driven) approaches complement bottom up (data driven) approaches effectively. This paper would have served a modest goal if it has aligned itself towards demonstrating such a possibility within the scope of discovering biological words, which is just one small step in the fascinating quest towards deciphering the language in which biological sequences express themselves.

References

Shlomo Argamon, Navot Akiva, Amihood Amir, and Oren Kapah. 2004. Efficient unsupervised recursive word segmentation using minimum description length. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, page 1058.

Quentin D Atkinson and Russell D Gray. 2005. Cu-

rious parallels and curious connectionsphylogenetic thinking in biology and historical linguistics. *Systematic biology* 54(4):513–526.

Patricia Bralley. 1996. An introduction to molecular linguistics. *BioScience* 46(2):146–153.

John-Marc Chandonia, Gary Hon, Nigel S Walker, Loredana Lo Conte, Patrice Koehl, Michael Levitt, and Steven E Brenner. 2004. The ASTRAL compendium in 2004. *Nucleic acids research* 32(suppl 1):D189—-D192.

Ruey-Cheng Chen. 2013. An improved mdl-based compression algorithm for unsupervised word segmentation. In *ACL (2)*. pages 166–170.

Naomi K Fox, Steven E Brenner, and John-Marc Chandonia. 2014. Scope: Structural classification of proteinsextended, integrating scop and astral data and classification of new structures. *Nucleic acids research* 42(D1):D304–D309.

W Nelson Francis and Henry Kucera. 1979. The brown corpus: A standard corpus of present-day edited american english. *Providence, RI: Department of Linguistics, Brown University [producer and distributor]* .

Mario Gimona. 2006. Protein linguisticsa grammar for modular protein assembly? *Nature Reviews Molecular Cell Biology* 7(1):68–73.

Peter Grünwald. 2005. A tutorial introduction to the minimum description length principle. *Advances in minimum description length: Theory and applications* pages 23–81.

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics* 37(2):309–350.

Daniel Hewlett and Paul Cohen. 2011. Fully unsupervised word segmentation with bve and mdl. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pages 540–545.

Neil C Jones and Pavel Pevzner. 2004. *An introduction to bioinformatics algorithms*. MIT press.

Chunyu Kityz and Yorick Wilksz. 1999. Unsupervised learning of word boundary with description length gain. In *Proceedings of the CoNLL99 ACL Workshop. Bergen, Norway: Association f or Computational Linguistics*. Citeseer, pages 1–6.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, pages 100–108.

Kenta Motomura, Tomohiro Fujita, Motosuke Tsutsumi, Satsuki Kikuzato, Morikazu Nakamura, and Joji M Otaki. 2012. Word decoding of protein amino acid sequences with availability analysis: a linguistic approach. *PloS one* 7(11):e50039.

Kenta Motomura, Morikazu Nakamura, and Joji M Otaki. 2013. A frequency-based linguistic approach to protein decoding and design: Simple concepts, diverse applications, and the scs package. *Computational and structural biotechnology journal* 5(6):1–9.

Alexey G Murzin, Steven E Brenner, Tim Hubbard, and Cyrus Chothia. 1995. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology* 247(4):536–540.

David B Searls. 2002. The language of genes. *Nature* 420(6912):211–217.

Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5(1):3–55.

Christian JA Sigrist, Edouard De Castro, Lorenzo Cerutti, Béatrice A Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenarios. 2012. New and continuing developments at prosite. *Nucleic acids research* page gks1067.

Radu Soricut and Franz Josef Och. 2015. Unsupervised morphology induction using word embeddings. In *HLT-NAACL*. pages 1627–1637.

Ashish Vijay Tendulkar and Sutanu Chakraborti. 2013. Parallels between linguistics and biology. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing, Sofia, Bulgaria: Association for Computational Linguistics*. Citeseer, pages 120–123.

Wikipedia. 2017. Protein structure — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Protein%20structure&oldid=774730776`. [Online; accessed 22-April-2017].

Yiming Yang. 1999. An evaluation of statistical approaches to text categorization. *Information retrieval* 1(1-2):69–90.

Valentin Zhikov, Hiroya Takamura, and Manabu Okumura. 2010. An efficient algorithm for unsupervised word segmentation with branching entropy and mdl. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 832–842.