Words in Biology

Rahul Kejriwal, CS14B023 Srinidhi Prabhu, CS14B028

Indian Institute of Technology, Madras

Outline

Motivation

- Proteins are strings composed of amino acids.
- Subsequences of amino acids can be thought of as functional units in proteins.
- Functional units in proteins can be seen as words of an unknown language.[1]

[1] [Tendulkar and Chakraborti, 2013]

Motivation

- Proteins are strings composed of amino acids.
- Subsequences of amino acids can be thought of as functional units in proteins.
- Functional units in proteins can be seen as words of an unknown language.[1]
- Can we use existing methods to identify the words in the language of proteins?

[1] [Tendulkar and Chakraborti, 2013]

Approach

- To identify words from natural language texts, we use word segmentation algorithms.
- Word segmentation algorithms can be of two types:
  1. **Supervised:** The possible set of words is known beforehand and stored as a dictionary.
  2. **Unsupervised:** The possible set of words are not known beforehand, and the segments must be identified by looking at repeating patterns in the corpus.
- We primarily focus on unsupervised word segmentation algorithms, because the words in the protein domain are not known to us.
- We then use the extracted segments to perform classification of proteins. The classification measures like precision and recall are used to measure the goodness of segmentation.

MDL-based approaches

- Works on the intuition that the segmentation that best compresses the string is actually the true segmentation.
- Works because language generally has repeating patterns and not all character n-grams are equiprobable.
- Formally,

$$segmentation = argmin_{s \in S} \ \Sigma_i \ wc_i log(\frac{wc_i}{N}) + \Sigma_j \ c_j log(\frac{c_j}{M})$$

where wc_i is count of word w_i in the corpus, $N = \Sigma_i wc_i$, c_j is the count of the jth character in the codebook and $M = \Sigma_j c_j$

- Can be thought of as a noisy channel process

$$segmentation = argmax_{s \in S} \ P(OS|s)P(s)$$

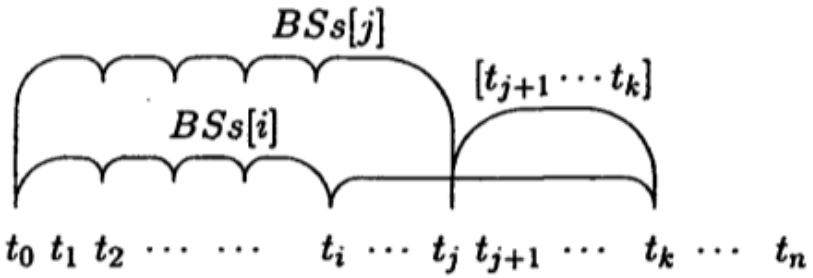
where S is set of segmentations that generate the given string OS.

Approach 1: Segmentation using Description Length gain

- $O(2^n)$ segmentations for a string of length n exist.
- Idea is to effectively traverse the search space using a heuristic.
- Use dynamic programming to find the optimal segmentation till the kth index by optimizing on a metric called Description Length Gain (DLG)

$$DLG(OS[j]) = \Sigma_{s \in OS[j]} \left(DL(X) - DL(X[r \rightarrow s] \oplus s)\right)$$

- Can be efficiently done in O(nlogn)



Approach 2: Segmentation using Branching Entropy and MDL

- Uses the intuition that uncertainty of next character in a stream is higher at word boundaries than within words
- Formally, branching entropy is

$$H(X_k|x_{k-1},..,x_{k-n}) = -\Sigma_{x \in X} P(x|x_{k-1},..,x_{k-n}) log_2 P(X_k|x_{k-1},..,x_{k-n})$$

- Generally, we use bidirectional variant of branching entropy
- Uses a 3 step approach:
  - Find initial segmentation by finding a good threshold for branching entropy
  - Try possible splitting/merging of segments (local changes) in order of their costs and accept if DL decreases
  - Try possible splitting/merging of segment types (global changes) in order of their costs and accept if DL decreases

Evaluation

- We compare the two approaches by measuring precision and recall of word boundary detection on the first half of the text *"Alice in Wonderland"*

Approach	**Precision**	**Recall**	**F1-score**
Approach 1	0.445	0.884	0.592
Approach 2	0.564	0.832	0.673

Table: Performance metrics of word segmentation algorithms

Approach 3: Segmentation as Search

- Segmentation can be modeled as finding the best possible segmentation in a space of possible segmentations.
- AI search techniques can be used to search for these segmentations as the search space is exponential.
- We attempt to search for good segmentations by using Genetic Algorithms (GA).
- The components of the algorithm are:
  - The fitness function is the negative of Description Length.
  - An individual is represented as the indices at which the string has to be segmented.
  - Crossover and mutation operations are defined appropriately on these individuals.

Protein Classification

- Classification using the segments as features is used as an extrinsic measure to evaluate the segmentations[2].
- We use two approaches to classification:
  1. Deep Learning techniques
  2. Dictionary-based segmentation followed by classification[3]

[2] [Devi et al., 2017]
[3] [Yang et al., 2008]

Approach 1: Classification using Deep Learning techniques

- Proteins are sequential in nature, motivating the use of recurrent neural networks to build classification systems.
- The following architecture was used:
  - Embed amino acids to a vector space
  - Use an LSTM network to compute representation of protein
  - Use a feedforward network to classify protein from the protein representation
- We obtain an average precision and average recall of 0.89 when trained over 25,000 proteins and tested over 5,000 proteins.

Approach 2: Dictionary-based Segmentation

The steps in dictionary-based segmentation[4] followed by classification are as follows:

- Build a dictionary with maximum word length of 4 amino acids.
- Use a Dynamic Programming based algorithm to perform segmentation.
- The words of the segments of a given protein are used as features. Each protein is represented as a vector of features.
- An SVM with RBF kernel is trained over a train set and then tested.
- We obtain an average precision and average recall of 0.75 when trained over 25,000 proteins and tested over 5,000 proteins.

[4] [Yang et al., 2008]

References

[Devi et al., 2017] Devi, G., Tendulkar, A. V., and Chakraborti, S. (2017).
Protein word detection using text segmentation techniques.
BioNLP 2017, page 238.

[Tendulkar and Chakraborti, 2013] Tendulkar, A. V. and Chakraborti, S. (2013).
Parallels between linguistics and biology.
ACL 2013, page 120.

[Yang et al., 2008] Yang, Y., Lu, B.-l., and Yang, W.-Y. (2008).
Classification of protein sequences based on word segmentation methods.
In *Proceedings of the 6th Asia-Pacific Bioinformatics Conference*, pages 177–186. World Scientific.