NLP7-8780 Project: **Latent Semantic Analysis for Classifying Genomic Sequences**

Instructor: Dr. Vasile Rus
Student: Quang Tran

Spring 2017

1 Introduction

1.1 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a technique in natural language processing, in particular distributional semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. In the field of information retrieval, LSA techniques are used to identify text documents with related underlying concepts. In this project, we extend these approaches to infer the biological structure through which a collection of organisms are related; we would like to identify and partition evolutionarily similar genomic sequences.

1.2 Genomic K-mer Sequences

The term k-mer typically refers to all possible subsequences of length k that are contained in a larger sequence. For instance, we can say ATGCCCATTA is a k-mer of length 10 or a "10-mer". And this "10-mer" contains the 4-mers: ATGC, TGCC, GCCC, CCCA, CCAT, CATT, and ATTA. Since the nucletotides in DNA contain four different bases: A, T, G, and C, these 4-mers above are just a small subset of the $4^k = 4^4 = 256$ possible k-mers.

1.3 Bag-of-k-mers Model

A DNA sequences can be hypothesized by some statistical presentation of the k-mers from which it was comprised. LSA techniques are based on the bag-of-words model for prepresenting text documents By treating k-mers as the words, we modify the original model to bag-of-k-mers model in genetic language.

2 Project Plan

- Week 1: Study LSA, implement LSA in python
- Week 2: Collect genomic data, compute k-mers on data
- Week 3: Integrate LSA methods to biological data
- Week 4,5,6: Implement and improve the integration
- Week 7: Finalize the proposed method, revise project report
- Week 8: Review overall

References

[1] V. Rus, "Lecture slides for natural language processing course nlp7-8780," 2017.

[2] J. H. Martin and D. Jurafsky, "Speech and language processing," *International Edition*, vol. 710, 2000.

[3] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.

[4] S. T. Dumais, "Latent semantic analysis," *Annual review of information science and technology*, vol. 38, no. 1, pp. 188–230, 2004.