Rare and frequent n-grams in whole-genome protein sequences

**Madhavi Ganapathiraju, Judith Klein-Seetharaman,
Roni Rosenfeld, Jaime Carbonell and Raj Reddy** [1]

Keywords: probabilitistic modeling, biology-language analogy, protein folding

1 Introduction.

The precise relationship between a primary protein sequence, its three-dimensional structure and its function in a complex cellular environment is one of the most fundamental unanswered questions in biology. Unprecedented amounts of genomic and proteomic data create an opportunity for attacking the sequence-structure-function mapping problem with data-driven methods. The mapping of biological sequences to form and function of proteins is conceptually similar to the mapping of words to meaning. This analogy is being studied by a growing body of research ([1] and pointers thereof). Thus, n-gram analysis (statistical analysis of co-occurrence of words in a text) has found applications to biological sequences, using various types of "vocabulary", for example nucleotides and amino acids. Here, we investigate n-gram statistics in whole-genome sequences to address the following questions: How characteristic is the amino acid n-gram distribution for specific organisms? Do different organisms tend to use different "phrases"? What is the "meaning" of a rare sequence in a protein? The long-term goal is to provide a useful starting point to derive language models with defined vocabulary and phrase preferences and grammatical rules for protein sequences of different organisms.

2 Development of a tool-kit for biological language modeling.

Statistical analysis of biological sequence data requires n-gram string matching and string searches, a computationally challenging problem for large-sized genomic data. Therefore, data structures like suffix trees [2] and suffix arrays [3], have been used, also for biological data [4], alone or complemented with other data arrays, e.g. the Least Common Prefix (LCP) array [3] and/or Rank arrays [5] for added functionality. A sub-string of length P in a string of length N can be searched in O(P+log N) time, and requires O(N) space for construction, which is competitive with those of suffix trees [3]. Preprocessed suffix arrays can be used to efficiently extract global n-gram statistics and compare it amongst various genomes. We have assembled a tool-kit that combines the following functions: *(1) Protein number and length; (2) n-gram statistics; (3) n-grams of specified length; (4) Relative frequencies of specific n-grams across organisms; (5) Longest repeating sequences; (6) (co-)Localization of n-grams for grouping proteins; (7) N-gram neighbors; (8) n-gram frequency distribution in specific protein sequences from global statistics; (9) Preprocessing of sequence data for analysis in CMU/Cambridge Statistical Language Modeling (SLM) Toolkit [6]*. The functions of the toolkit were applied to protein sequences derived from whole-genome sequences of 44 different organisms. Amino acids were treated as words.

3 Results: comparative genome n-gram statistics.

(1) Probabilistic models can distinguish organisms – A simple Markovian unigram (context independent amino acid) model from the proteins of *Aeropyrum pernix* was trained. When training and test set were from the same organism, a perplexity (a variation on cross-entropy) of 16.6 was observed, whereas data from other organisms varied from 16.8 to 21.9. Thus the differences between the 'sub-languages' of the different organisms are automatically detectable with even the simplest language model.

[1] School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, Email: {madhavi, judithks, roni, jgc, rr}@cs.cmu.edu

(2) Comparative Zipf-like analysis – We developed a modification of Zipf-like analysis that can reveal differences between word-usage in different organisms. N-grams were ranked from most frequent to least frequent for one organism, and were compared to the respective frequencies in all the other organisms. While there was striking variation in rank of certain n-grams in different organisms, the most rare n-grams were rare in all organisms.

(3) Amino acid neighbor preferences - When analyzing bigram statistics in more detail, we found differences in local preferences for amino acid neighbors even when these were 5 amino acids apart. Again, these patterns varied amongst organisms.

(4) Organism-specific usage of "phrases" in protein sequences - As we moved to larger contexts, organisms showed marked differences in the statistics of their n-gram distribution. For example, we found n-grams that are very frequent in some organisms yet rare (or completely absent in some cases) in others. These highly idiosyncratic n-grams suggest "phrases" that are preferably used in the particular organism.

(5) Rare peptide sequences correlate with folding initiators - In natural languages, frequent words carry little meaning, while rare words often allow identification of what a particular text is about. The distribution of n-gram frequencies from the human genome was determined for the human lysozyme sequence. The inverse of n-gram (i.e. trigram) frequencies, but not frequencies, correlated significantly with the location of two clusters of structure at the initial stages of protein folding of lysoyzme that have been determined experimentally [7]. The observation that inverse amino acid n-gram frequency is more informative than frequency is in direct analogy to human languages. Importantly, rare n-grams seem to point at the initiators of folding.

4 Summary and conclusions.

A toolkit for genome-wide statistical n-gram analysis and comparison across organisms was developed and its functions were applied to 44 different bacterial, archaeal and the human genome. Amino acid n-gram distribution was found to be characteristic of organisms, as evidenced by (1) the ability of simple Markovian unigram models to distinguish organisms, (2) the marked variation in n-gram distributions across organisms, (3) organism-specific preferences for amino acid distance bigrams and (4) identification of organism-specific phrases in protein sequences. Furthermore, rare peptide sequences were found to correlate with folding initiators. The results suggest that further detailed analysis of n-gram statistics of protein sequences from whole genomes will likely – in analogy to word n-gram analysis – result in powerful models for prediction, topic classification and information extraction of biological sequences.

5 References.

1. *Language Modeling of Biological Data Workshop*. ed. D. Searles. 2001, University of Pennsylvania.http://www.ircs.upenn.edu/modeling2001/modeling.shtml,
2. Weiner, P. *Linear pattern matching algorithms*. in *14th Annual Symposium on Switching and Automata Theory*. 1973. University of Iowa.
3. Manber, U. and G. Myers, *Suffix arrays: A New Method for On-Line String Searches.* SIAM Journal on Computing, 1993. **22**(5): p. 935-948.
4. Burkhardt, S., et al. *q-gram Based Database Searching Using a Suffix Array (QUASAR)*. in *Third Annual Int. Conference on Computational Molecular Biology, RECOMB'99.* 1999. Lyon, France.
5. Kasai, T., et al. *Linear-Time Longest-Common-Prefix computation in Suffix Arrays and Its applications*. in *Ann. Symp. on Combinatorial Pattern Matching CPM-2001*. 2001. Jerusalem, Israel.
6. Clarkson, P.R. and R. Rosenfeld, *Statistical language modeling using the CMU-Cambridge toolkit.* Proceedings ESCA Eurospeech, 1997.
7. Klein-Seetharaman, J., et al., *Long-range clusters within a non-native protein.* Science, 2001. in press**.**