Comparative n-gram analysis of whole-genome protein sequences

M. Ganapathiraju, D. Weisser, R. Rosenfeld, J. Carbonell, R. Reddy & J. Klein-Seetharaman
Carnegie Mellon University
Pittsburgh, PA15217
412 383 7325
{madhavi, dweisser, roni, jgc, rr, judithks}@cs.cmu.edu

ABSTRACT

A current barrier for successful rational drug design is the lack of understanding of the structure space provided by the proteins in a cell that is determined by their sequence space. The protein sequences capable of folding to functional three-dimensional shapes of the proteins are clearly different for different organisms, since sequences obtained from human proteins often fail to form correct three-dimensional structures in bacterial organisms. In analogy to the question "What kind of things do people say?" we therefore need to ask the question "What kind of amino acid sequences occur in the proteins of an organism?" An understanding of the sequence space occupied by proteins in different organisms would have important applications for "translation" of proteins from the language of one organism into that of another and design of drugs that target sequences that might be unique or preferred by pathogenic organisms over those in human hosts.

Here we describe the development of a biological language modeling toolkit (BLMT) for genome-wide statistical amino acid n-gram analysis and comparison across organisms (freely accessible at www.cs.cmu.edu/~blmt). 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 above random variation, and (3) identification of organism-specific phrases in protein sequences that are greater than an order of magnitude standard deviations away from the mean. These lines of evidence suggest that different organisms utilize different "vocabularies" and "phrases", an observation that may provide novel approaches to drug development by specifically targeting these phrases. 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.

Keywords

Biological language modeling toolkit, genome signatures.

Proceedings of HLT 2002, Second International Conference on Human Language Technology Research, M. Marcus, ed., Morgan Kaufmann, San Francisco, 2002.

1. INTRODUCTION

1.1 Opportunity for human language technologies in biological data analysis

Central to the understanding of complex biological systems are proteins. Their form and function is in principle encoded in characteristic amino acid sequences. 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. Large amounts of genomic and protein sequence data for homo sapiens and other organisms have recently become available, together with a growing body of protein structure and function data. The expected exponential increase in the amount of this data in the coming decade creates an opportunity for attacking the sequence-structure-function mapping problem with increasingly sophisticated data-driven methods. Such methods have proven immensely successful in the domain of natural language, and are directly responsible for the success of automatic speech recognition, document classification, information extraction, statistical machine translation and other challenging tasks over the past two decades.

1.2 Introduction: biological language

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, word n-gram analysis has found applications to biological sequences, using various types of "vocabulary", for example the nucleotides or the 61-codon types in the case of DNA (e.g. [2]), and the standard 20 amino acids or reduced 3-letter charge groups of the amino acids in the case of proteins (e.g. [3]). Thus, nearest-neighbor correlation analysis have revealed specific preferences for proximity of certain amino acids in protein sequences [4]. The results from n-gram analysis have been used in some cases to demonstrate that genome or protein sequences follow Zipf law [5-11]. However, since non-deterministic sequences also follow a power law (see e.g. [12]), detection of linguistic features in biological sequence data based on the distribution of n-grams is controversial and the extent to which amino acid sequences can be modeled stochastically is not clear.

The advent of whole-genome sequencing efforts provides a new opportunity to revisit n-gram statistics and Zipf-type analysis in greater detail. In particular, specific questions that can now be addressed are: How characteristic is the amino acid n-gram distribution for specific organisms? Do different organisms tend to

use different phrases? Previous determination of global statistics of entire genomes supports that there are genome-specific regularities in n-gram statistics. For example, species-specific regularity in composition ("unigram count") has been identified [13], the typical length of prokaryotic proteins is different from that in eukaryotes [14, 15] and differences in patterns in usage of secondary structure elements [16] by various genomes have been observed. Here we extend global genome sequence analysis to systematically compare n-gram statistics of protein sequences from a larger number of known genomes. 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 (BLMT)

Statistical analysis of biological sequence data requires n-gram string matching and string searches. Due to the large size of genomic data, the search for subsequences becomes a computationally challenging problem. Searching for a sub-string from large text data is a well-studied problem in computer science, with applications to diverse areas including data compression, network intrusion detection, information retrieval and word processing [17]. Data structures like suffix trees [18] and suffix arrays [19] have been used as preferred data structures for applications of this kind [19-21] and more recently also for biological data [22]. When suffix arrays are complemented with other data arrays, e.g. the Least Common Prefix (LCP) array [19] and/or Rank arrays [23], they provide additional functionality at reduced computational cost. Thus, it permits search of a sub-string of length P in a string of length N in O(P+log N) time, and requires O(N) space for construction, which is competitive with those of suffix trees [19]. Preprocessed suffix arrays can now be used to efficiently extract global n-gram statistics and compare it amongst various genomes. The method is illustrated in Figure 1 using the organism Aeropyrum pernix as an example (Table 1).

To extract n-gram statistical data from the genome suffices, we have assembled a tool-kit that combines the following functions:
(1) Counting protein number and length
(2) Counting n-grams and most frequent n-grams
(3) Counting n-grams of specified length
(4) Determining relative frequencies of specific n-grams across organisms
(5) Identifying longest repeating sequences
(6) Localization and co-localization of n-grams for grouping proteins
(7) N-gram neighbor (left and right) identification
(8) Distribution of n-gram frequencies in specific protein sequences from global statistics
(9) Preprocessing of sequence data to prepare for analysis in CMU/Cambridge Statistical Language Modeling (SLM) Toolkit [24].

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. The numbers of proteins varied from 484 (175,928 amino acids) in Mycoplasma genitalium to 25612 (18,283,879 amino acids) in Homo sapiens.

Table 1. Format of protein sequence input files

>gi|5103389|dbj|BAA78910.1| 241aa long hypothetical protein
MVDILSSLLL
>gi|5103390|dbj|BAA78911.1| 112aa long hypothetical protein
MDPADKLMK
>gi|5103391|dbj|BAA78912.1| 100aa long hypothetical protein
MQA

3. RESULTS: COMPARATIVE GENOME N-GRAM STATISTICS

3.1 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. This observation is purely based on the large differences in unigram distributions (described in Section 3.2 and Figure 2 below) and is independent of the organism that is used to train the model.

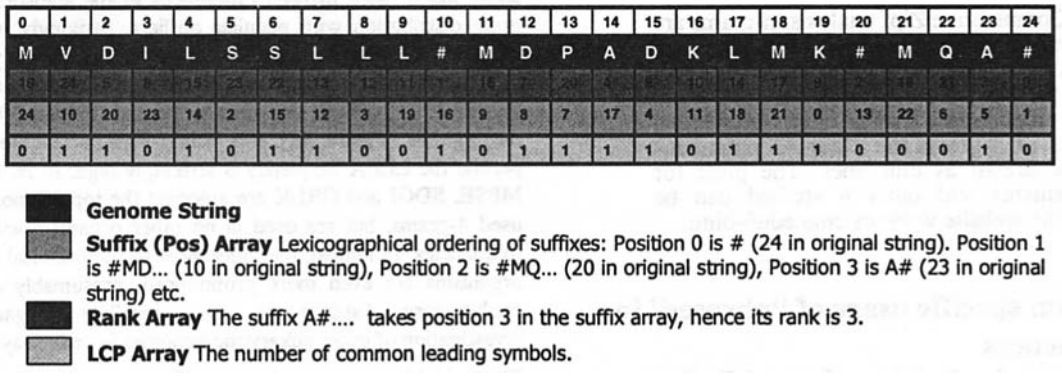

Figure1. Example for genome string organization in suffix arrays: *Aeropyrum pernix*.

3.2 Comparative Zipf-like analysis

We developed a modification of Zipf-like analysis that can reveal differences between word-usage in different organisms. First, the amino acid n-grams of a given length are sorted in descending order by frequency for the organism of choice. An example using the simplest case, n=1, is shown in Figure 2 for two organisms Aeropyrum pernix and Neisseria meningitidis to illustrate the principle. The frequencies of the sorted n-grams are shown in bold red. Thin lines indicate the respective frequencies of n-grams with given rank in Aeropyrum pernix (Figure 2A) or Neisseria meningitides (Figure 2B) in all the other organisms. The same plots for the other 42 organisms studied for n=1 and also for other n (n<5) can be viewed at www.cs.cmu.edu/~blmt. While there is striking variation in rank of certain n-grams in different organisms, the most rare n-grams in one organism are overall rare in all organisms. Specific differences in n-grams other than unigrams are explored in more detail below (Section 3.3).

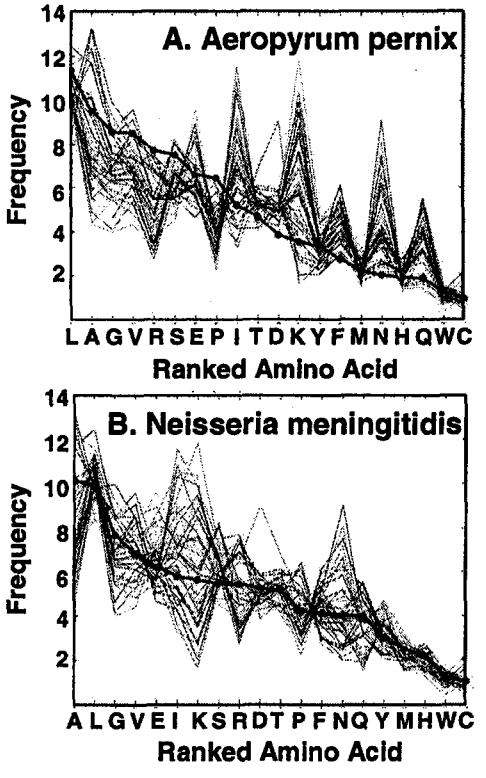

Figure 2. Comparative Zipf analysis: n-grams are ranked according to frequency for one organism (shown as bold, red line). Shown here are two examples for n=1, Aeropyrum pernix (A.) and Neisseria menigitidis (B.). The respective frequencies of n-grams the other 43 organisms studied are drawn as thin lines. The plots for other organisms and other n studied can be viewed at the website www.cs.cmu.edu/~blmt.

3.3 Organism-specific usage of "phrases" in protein sequences

The Zipf-like analysis described above (Section 3.2) allows us to quantify the differences in specific n-gram frequencies across

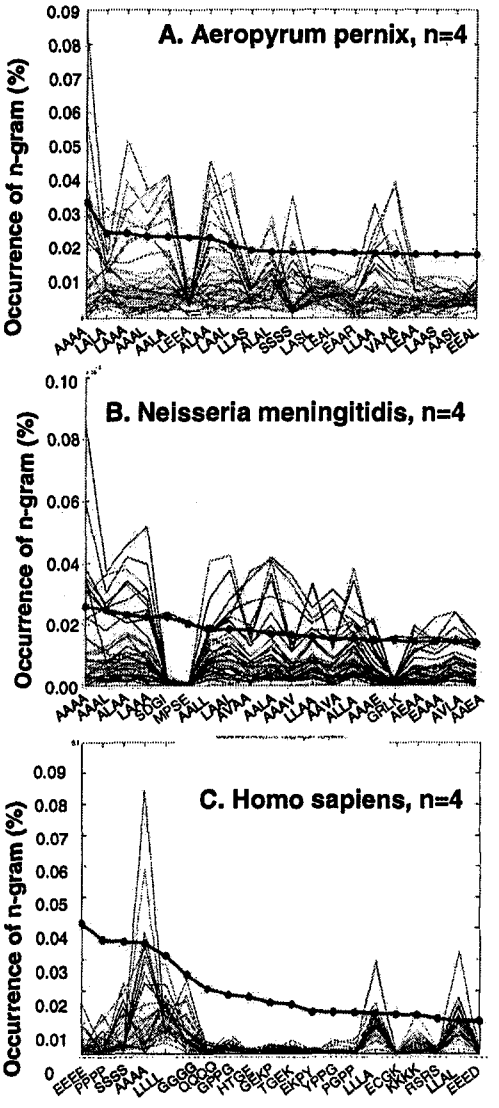

Figure 3. Comparative Zipf analysis: Top 20 most frequently used 4-grams in Aeropyrum pernix (A), Neisseria meningitidis (B) and Homo sapiens (C). Line colors as in Figure 2.

organisms. In particular, as we move to larger contexts, organisms show much more marked differences in the statistics of their n-gram distribution with peculiar outliers. Strikingly, we found n-grams that are very frequent in some organisms yet rare (or completely absent in some cases) in others. Examples are shown in Figure 3 for n=4 in Aeropyrum pernix (Figure 3A), Neisseria meningitidis (Figure 3B) and Homo sapiens (Figure 3C). In A. pernix, the LEEA frequency is strikingly high. In N. meningitidis, MPSE, SDGI and GRLK are amongst the top 20 most frequently used 4-grams, but are used in no other organism with such high frequencies. In human, the differences to the bacterial and archaeal organisms are even more pronounced, presumably due to their evolutionary distance to the unicellular organisms. The investigation of other eukaryotic genomes is underway.

These highly idiosyncratic n-grams suggest "phrases" that are preferably used in the particular organism. The observation of

organism-specific phrases is not unique to extremophile or other specialized organisms. Instead, idiosyncratic phrases appear in all the organisms (also see Section 3.4 below), and the results for other organisms (including very common and ubiquitous bacteria such as Escherichi coli) can be viewed at www.cs.cmu.edu/~blmt. Importantly, these phrases can be organism-specific.

3.4 Phrases are not due to random variation

To test if the observation of idiosyncratic n-grams could be explained by chance sampling, we generated two sets of 20 artificial genomes by Monte Carlo simulation using the unigram frequencies of Neisseria meningitidis and Aeropyrum pernix, respectively. Figure 4 shows a Zipf-like comparison as described above for the natural genomes, for Neisseria meningitidis in comparison to the random genomes (A), for Aeropurum pernix in comparison to the random genomes (B) and for one of the random genomes in comparison to the other random genomes and the

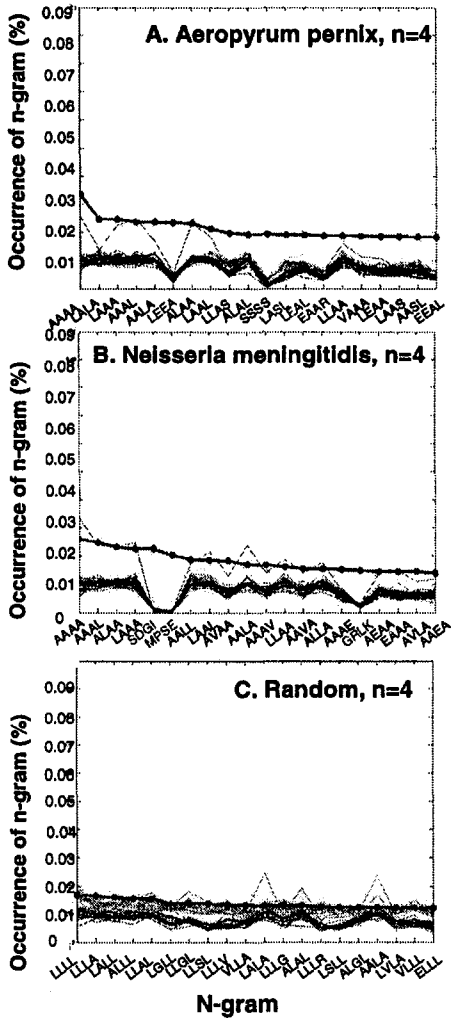

Figure 4. Comparative Zipf analysis of random genomes versus natural genomes: Top 20 most frequently used 4-grams in Aeropyrum pernix (A), Neisseria meningitidis (B) and a random genome (C). Line colors as in Figure 2. Note that both natural genomes strike out, not only the one according to which the n-grams were ranked.

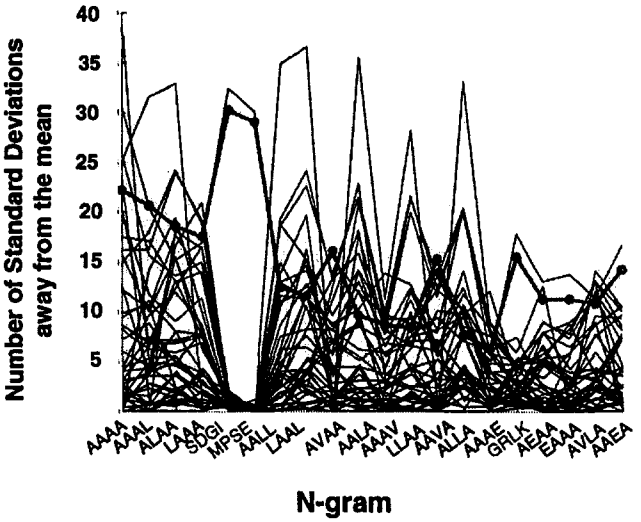

Figure 5. Distance from mean values based on unigram distributions in Neisseria meningitidis. Values are plotted as multiples of standard deviation from mean. The unigram distribution was as in Figure 2A.

Neisseria and Aeropyrum genomes (C). As one can see, in both natural genomes the frequencies are well above the baseline variation due to chance sampling.

3.5 Phrase frequencies can be very distant from mean values

To further strengthen the notion that the phrases are not due to random variation, we calculated the distance of 4-gram frequencies in multiples of standard deviations for the top 20 4-grams in Neisseria meningitidis. The result is shown in Figure 5. The phrases SDGI and MPSE are approximately 30 standard deviations away from the means based on unigram distributions. In contrast, all of the other organism, except for a different strain of Neisseria meningitidis, show only very small standard deviations from mean values based on their own unigram frequencies. GRLK is also more frequent than would be expected based on independent unigram probabilities, although not to the same degree as SDGI and MPSE. The large deviation from mean values clearly shows that phrases are not only organism-specific in absolute terms but are also quantifiably distant from the values predicted by independent unigram frequencies of the same organism.

3.6 How many phrases are there in an organism?

The previous section has shown that there is a correlation between deviation from mean values within the same organism and difference in frequency of certain n-grams in comparison with other organisms. The next step is to identify all the phrases in an organism. Towards this goal, we have quantified the number of n-grams as a function of standard deviation from mean values. The result is shown for one organism (Escherichia coli) in Figure 6, for n=2 (Figure 6A), n=3 (Figure 6B) and n=4 (Figure 6C). Especially 3-gram and 4-gram values are heavily tailed. It is this long tail which gives rise to the large deviations observed in Figure 5.

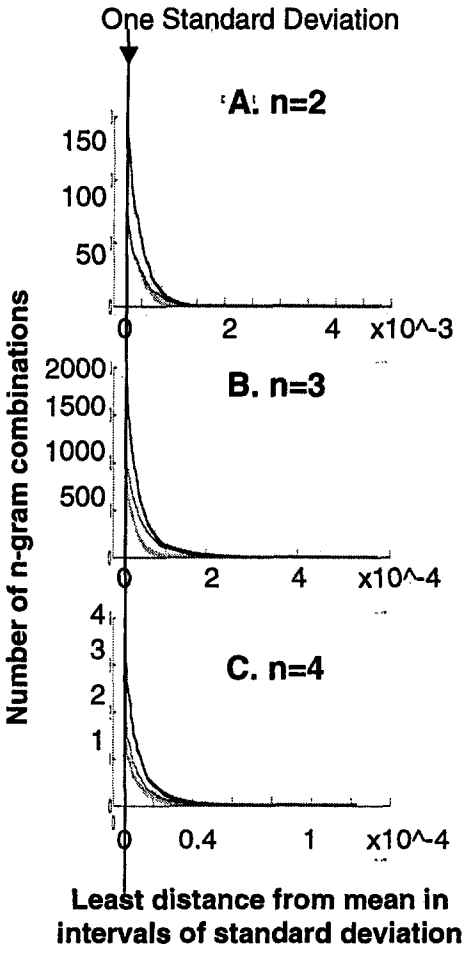

Figure 6. Number of n-grams in dependence of distance from mean value for Escherichia coli. Solid lines, absolute values; dotted lines, negative values (underrepresented n-grams); dashed lines, positive values (overrepresented n-grams).

4. CONCLUSIONS AND FUTURE WORK

Using n-gram statistical analysis of whole-genome protein sequences we have shown that there are organism-specific phrases in direct analogy to human languages. Future work will aim at detailed identification of these phrases to map out the sequence space occupied by proteins in different genomes. We will test experimentally what is the structure space occupied by these sequences to map their biological significance.

5. ACKNOWLEDGMENTS

This work was supported by Information Technology Research Grant 0204078 from the National Science Foundation.

6. REFERENCES

[1] Language Modeling of Biological Data Workshop. ed. D. Searles. 2001, University of Pennsylvania. http://www.ircs.upenn.edu/modeling2001/modeling.shtml,

[2] Burge, C., A.M. Campbell, and S. Karlin, Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci U S A, 1992. 89(4): p. 1358-62.

[3] Erhan, S., T. Marzolf, and L. Cohen, Amino-acid neighborhood relationships in proteins. Breakdown of amino-acid sequences into overlapping doublets, triplets and quadruplets. Int J Biomed Comput, 1980. 11(1): p. 67-75.

[4] Karlin, S., et al., Statistical methods and insights for protein and DNA sequences. Annu Rev Biophys Biophys Chem, 1991. 20: p. 175-203.

[5] Weiss, O., M.A. Jimenez-Montano, and H. Herzel, Information content of protein sequences. J Theor Biol, 2000. 206(3): p. 379-86.

[6] Li, W., Statistical properties of open reading frames in complete genome sequences. Comput Chem, 1999. 23(3-4): p. 283-301.

[7] Czirok, A., et al., Correlations in binary sequences and a generalized Zipf analysis. Physical Review. E. Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics, 1995. 52(1): p. 446-452.

[8] Israeloff, N.E., M. Kagalenko, and K. Chan, Can Zipf distinguish language from noise in noncoding DNA? Physical Review Letters, 1996. 76(11): p. 1976.

[9] Konopka, A.K. and C. Martindale, Noncoding DNA, Zipf's law, and language. Science, 1995. 268(5212): p. 789.

[10] Mantegna, R.N., et al., Linguistic features of noncoding DNA sequences. Phys Rev Lett, 1994. 73(23): p. 3169-72.

[11] Tsonis, A.A., J.B. Elsner, and P.A. Tsonis, Is DNA a language? J Theor Biol, 1997. 184(1): p. 25-9.

[12] Chatzidimitriou-Dreismann, C.A., R.M. Streffer, and D. Larhammar, Lack of biological significance in the 'linguistic features' of noncoding DNA--a quantitative analysis. Nucleic Acids Res, 1996. 24(9): p. 1676-81.

[13] Karlin, S., B.E. Blaisdell, and P. Bucher, Quantile distributions of amino acid usage in protein classes. Protein Eng, 1992. 5(8): p. 729-38.

[14] Venter, J.C., et al., The sequence of the human genome. Science, 2001. 291(5507): p. 1304-51.

[15] Lander, E.S., et al., Initial sequencing and analysis of the human genome. Nature, 2001. 409:860-921.

[16] Bradley, P., et al., BETAWRAP: successful prediction of parallel beta -helices from primary sequence reveals an association with many microbial pathogens. Proc Natl Acad Sci U S A, 2001. 98(26): p. 14819-24.

[17] Apostolico, A., The Myriad virtues of subword trees, Combinatorial Algorithms on Words. NATO ASI series in Computer and System Sciences, 1985. 12: p. 85-96.

[18] Weiner, P. Linear pattern matching algorithms. in 14th Annual Symposium on Switching and Automata Theory. 1973. University of Iowa.

[19] 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.

[20] Grossi, R. and J.S. Vitter. Compressed Suffix Arrays and Suffix Trees, with Applications to Text Indexing and String Matching. in Proceedings of the 32nd Annual ACM Symposium on Theory of Computing (STOC '00). 2000. Portland, OR.

[21] Larsson, N.J. Extended application of suffix trees to data compression. in IEEE Data Compression Conference. 1996.

[22] Burkhardt, S., et al. q-gram Based Database Searching Using a Suffix Array (QUASAR). in Third Annual Intl. Conference on Computational Molecular Biology, RECOMB'99. 1999. Lyon, France.

[23] Kasai, T., et al. Linear-Time Longest-Common-Prefix computation in Suffix Arrays and Its applications. in Annual Symposium on Combinatorial Pattern Matching CPM-2001. 2001. Jerusalem, Israel.

[24] Clarkson, P.R. and R. Rosenfeld, Statistical language modeling using the CMU-Cambridge toolkit. Proceedings ESCA Eurospeech, 1997.