

# BLMT

## Statistical Sequence Analysis Using *N*-Grams

Madhavi Ganapathiraju,<sup>1</sup> Vijayalaxmi Manoharan<sup>2</sup> and Judith Klein-Seetharaman<sup>1,2</sup>

1 Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

2 Department of Pharmacology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA

### Abstract

**Abstract:** Statistical analysis of amino acid and nucleotide sequences, especially sequence alignment, is one of the most commonly performed tasks in modern molecular biology. However, for many tasks in bioinformatics, the requirement for the features in an alignment to be consecutive is restrictive and '*n*-grams' (aka *k*-tuples) have been used as features instead. *N*-grams are usually short nucleotide or amino acid sequences of length *n*, but the unit for a gram may be chosen arbitrarily. The *n*-gram concept is borrowed from language technologies where *n*-grams of words form the fundamental units in statistical language models. Despite the demonstrated utility of *n*-gram statistics for the biology domain, there is currently no publicly accessible generic tool for the efficient calculation of such statistics. Most sequence analysis tools will disregard matches because of the lack of statistical significance in finding short sequences. This article presents the integrated Biological Language Modeling Toolkit (BLMT) that allows efficient calculation of *n*-gram statistics for arbitrary sequence datasets. **Availability:** BLMT can be downloaded from <http://www.cs.cmu.edu/~blmt/source> and installed for standalone use on any Unix platform or Unix shell emulation such as Cygwin on the Windows® platform. Specific tools and usage details are described in a 'readme' file. The *n*-gram computations carried out by the BLMT are part of a broader set of tools borrowed from language technologies and modified for statistical analysis of biological sequences; these are available at <http://flan.blm.cs.cmu.edu/>.

**Contact:** Judith Klein-Seetharaman ([judithks@cs.cmu.edu](mailto:judithks@cs.cmu.edu))

Many of the hallmarks of statistical analysis of nucleotide and protein sequences are similar to those of human languages: (i) large data bodies need to be analysed in both cases; (ii) the fundamental units of human languages include higher order structures, paralleled by domains, subunits or functionally linked proteins; and (iii) computer-based derivation of meaning from text is analogous to the prediction of structure and function from primary sequence data. Therefore, the analogy has led to the wide application of methods used in language technologies to the study of biological sequences.<sup>[1-3]</sup> For example, hidden Markov models (HMMs), support vector machines (SVMs) and neural networks are standard methods used in bioinformatics, and even formal language theory has found applications in biology.<sup>[1-8]</sup> Recently, probabilistic language models have been used to improve protein domain boundary detection,<sup>[9]</sup> predict transmembrane helix

boundaries<sup>[10]</sup> and in genome comparison.<sup>[11]</sup> Furthermore, latent semantic analysis, a technique used in text processing, has been used for secondary structure prediction.<sup>[12]</sup> In addition, methods used for topic segmentation of text or radio speech have been applied to the membrane protein boundary prediction problem, in particular using Yule's Q-statistic<sup>[13]</sup> and mutual information.<sup>[14]</sup> Finally, the application of conditional random fields, a technique also borrowed from language technologies, to secondary structure prediction has recently been reported.<sup>[15]</sup>

Language technologies often use word *n*-grams as features. *N*-grams refer to sequential occurrences of words in a text. In biological sequences, the best equivalent of human words is not known. Thus, *n*-grams usually describe short sequences of nucleo-

tides or amino acids of length  $n$ .<sup>1</sup> Other ‘vocabulary’ that has been used includes the 61-codon types or reduced amino acid alphabets.<sup>[16-19]</sup> Probably the most widely known application of  $n$ -grams is their use in the BLAST<sup>®</sup> algorithm, where they enhance computational efficiency in sequence searching in the initial step.<sup>[20]</sup> However,  $n$ -grams have also proven useful in a number of other bioinformatics areas. Global distributions of  $n$ -grams across different datasets of biological sequences have been shown to follow Zipf’s law.<sup>[21]</sup> Zipf’s law states that the frequency of a word is related to its rank by a power law,<sup>[22,23]</sup> and the observation of Zipf behaviour in biological sequences was originally used to infer ‘linguistic’ properties of such data.<sup>[21]</sup> However, it was later shown that randomised biological sequences follow similar trends;<sup>[24,25]</sup> it was also pointed out that any number of random processes can display Zipf-like behaviour<sup>[26]</sup> and that Zipf’s law cannot distinguish language from power-law noise.<sup>[27]</sup> Furthermore, depending on the dataset studied,  $n$ -gram distributions in biological sequences are better approximated by an exponential distribution function rather than Zipf’s law.<sup>[28]</sup> Thus, although the Zipf law cannot be used as a linguistic test,  $n$ -gram distributions in biological datasets do show significant nonrandomness.<sup>[28,29]</sup> Furthermore, a recently developed modification of the Zipf-like analysis of  $n$ -gram frequencies that allows comparison of individual  $n$ -grams rather than the global distributions (see the Applications section) has found recent application in identification of genome signatures.<sup>[11]</sup> It has also been shown that  $n$ -grams are useful features in the prediction tasks of protein family classification<sup>[10,30,31]</sup> and transmembrane helix boundary detection.<sup>[10,13,14]</sup> Finally, the matching of peptide sequences to their respective structures in the ROSETTA protein tertiary structure prediction algorithm<sup>[32]</sup> and the statistical analysis of known protein structures in terms of their distribution with sequences of a given length<sup>[33]</sup> can both be viewed as applications of  $n$ -gram technology.

This article describes a publicly accessible toolkit, the Biological Language Modeling Toolkit (BLMT), that integrates various functionalities related to statistical  $n$ -gram analysis. BLMT functions as a foundation over which various different analysis tools have been and may be built that take advantage of lexicographical pre-processing of the sequence data. Unlike many other integrated tools and platforms for sequence analysis such as the Systems Biology Workbench,<sup>[34]</sup> BLMT is not only a platform that enables

use of a collection of specific applications. Instead, it is a scalable framework over which new applications may be built that take advantage of the time-efficient pattern searching capability provided by the underlying data structures. In this article, we describe four examples of applications that have benefited or can profit from the  $n$ -gram analysis capabilities of BLMT.

## Suffix Array Data Structure

BLMT allows for statistical  $n$ -gram analysis of large biological sequence datasets. For this purpose, BLMT generates a suffix array structure, as originally described by Manber and Myers,<sup>[35]</sup> from the data provided by the user. A large number of genomic and proteomic datasets are also stored at the URL <http://flan.blm.cs.cmu.edu>. Suffix trees and arrays are powerful structures optimised for large data and fast processing requirements and have been used in the bioinformatics domain for applications in whole-genome sequence alignment,<sup>[36]</sup> motif identification<sup>[37,38]</sup> and protein family classification.<sup>[39,40]</sup> The efficiency of our suffix array is further improved by accompanying data structures, including the longest common prefix (LCP) and the rank array described in Kasai et al.<sup>[41]</sup> The construction of these data structures is shown in figure 1.

For a given sequence of characters, a suffix at position  $i$  is the subsequence beginning at position  $i$  and extending to the end. In figure 1, the suffix at position 5 is shown in pink colour. A suffix array is the arrangement of all the possible suffixes of the input string in lexicographical order. Thus, if a pattern repeats at multiple positions in the input string, the suffixes beginning at these positions all appear consecutively in the suffix array since they all begin with that same pattern. To store the sorted order of the suffixes, only their beginning *positions* are entered in the suffix array data structure, as shown in the 4th row in figure 1 (headed ‘Suffix array’). The LCP array stores the length of the common *prefix* substring between a suffix and its preceding entry in the sorted suffix array. The LCP array can be constructed in linear time from the sorted suffix array and can be used in conjunction with various repetitive-pattern identification algorithms. The rank array stores – for each position in the input string – the position of its suffix in the suffix array. That is, if a suffix  $S_j$  appears at the  $j$ -th position in the lexicographical order of the suffix array, then  $\text{rank}[i] = j$ .

**1** Other notations used instead of  $n$ -grams are  $k$ -grams,  $k$ - or  $n$ -tuples and  $k$ - or  $n$ -mers.

| Position                               | 0  | 1   | 2  | 3  | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  | 13  | 14  | 15  | 16  | 17  | 18  | 19  | 20  | 21  | 22  | 23  | 24  |
|--|----|-----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Rank array                             | 19 | 24  | 5  | 8  | 15  | 23  | 22  | 13  | 12  | 11  | 1   | 16  | 7   | 20  | 4   | 6   | 10  | 14  | 17  | 9   | 2   | 18  | 21  | 3   | 0   |
| Sequence                               | M  | V   | D  | I  | L   | S   | S   | L   | L   | L   | #   | M   | D   | P   | A   | D   | K   | L   | M   | K   | #   | M   | Q   | A   | #   |
| Suffix array                           | 24 | 10  | 20 | 23 | 14  | 2   | 15  | 12  | 3   | 19  | 16  | 9   | 8   | 7   | 17  | 4   | 11  | 18  | 21  | 0   | 13  | 22  | 6   | 5   | 1   |
| LCP array                              | 0  | 1   | 2  | 0  | 1   | 0   | 1   | 1   | 0   | 0   | 1   | 0   | 1   | 2   | 1   | 1   | 0   | 1   | 1   | 1   | 0   | 0   | 0   | 1   | 0   |
| Suffixes sorted in lexicographic order | #  | #   | #  | A  | A   | D   | D   | D   | I   | K   | K   | L   | L   | L   | L   | L   | M   | M   | M   | M   | P   | Q   | S   | S   | V   |
|  |    | M   | M  | #  | D   | I   | K   | P   | L   | #   | L   | #   | L   | L   | M   | S   | D   | K   | Q   | V   | A   | A   | L   | S   | D   |
|  |    | D   | Q  |    | K   | L   | L   | A   | S   | M   | M   | M   | #   | L   | K   | S   | P   | #   | A   | D   | D   | #   | L   | L   | I   |
|  |    | P   | A  |    | L   | S   | M   | D   | S   | Q   | K   | D   | M   | #   | #   | L   | A   | M   | #   | I   | K   |     | L   | L   | L   |
|  |    | ... | #  |    | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

**Fig. 1.** Example of a suffix array and its longest common prefix (LCP) array and rank array for the sequence MVDILSSLL#MDPADKLMK#MQA#. An example suffix S5 (row 3, shown in pink) is the suffix beginning at position 5 ( $S_5$ ). When the suffixes are sorted in lexicographic order, the suffix  $S_5$  appears as the 23rd suffix. Thus, the rank of  $S_5$  is 23. Similarly, the suffix array carries the value '5' at position 23, indicating that the 23rd suffix is  $S_5$ . Suffixes in lexicographic order are shown in the lower half of the figure, arranged vertically. LCPs between adjacent suffixes are shown in same colour. The LCP array contains the lengths of these LCPs. For example, at position 13, the LCP with the previous suffix is 'LL', and its length is 2. Hence,  $LCP(13) = 2$ .

## Tools

### N-Gram Counts

The functionality of the BLMT includes  $n$ -gram counts from protein and nucleic acid sequences, where  $n$  is an arbitrary integer or range of integers. While locating sequence repeats in any two given sequences is a trivial task, it becomes computationally expensive when the search is to be performed on a database of the size of multiple genomes or on large  $n$ . BLMT uses the underlying suffix-array structure to retrieve repeating sequences of any length greater than a threshold set by the user or the longest repeating sequences efficiently. BLMT will also compute the co-occurrence counts of specific  $n$ -grams in subsets of the data, e.g. within individual proteins, and supports identification of  $n$ -gram neighbours (left and right). BLMT allows retrieval of proteins that contain common sequences longer than a threshold and annotation of  $n$ -gram counts along proteins. The  $n$ -gram counts can be sorted in various ways.

### Statistical Correlations

The  $n$ -gram counts are used to compute statistical correlations between amino acids or nucleotides using Yule's Q statistic.<sup>[42]</sup> This tool computes the Yule value between two amino acids separated by a specific distance (such as A\*\*C), as described in detail in Ganapathiraju et al.<sup>[13]</sup> BLMT also allows reading in Yule or other correlation values that have been pre-computed with other tools. It is possible to annotate a given protein sequence/nucleotide se-

quence with Yule values computed from a dataset. Yule values range from  $-1$  to  $+1$  reflecting a negative or positive influence, respectively, of the occurrence of two amino acids on each other. For a random distribution of amino acids, the Yule values are expected to be zero, and this is what was observed when Yule values of randomised protein sequences were computed.<sup>[13]</sup>

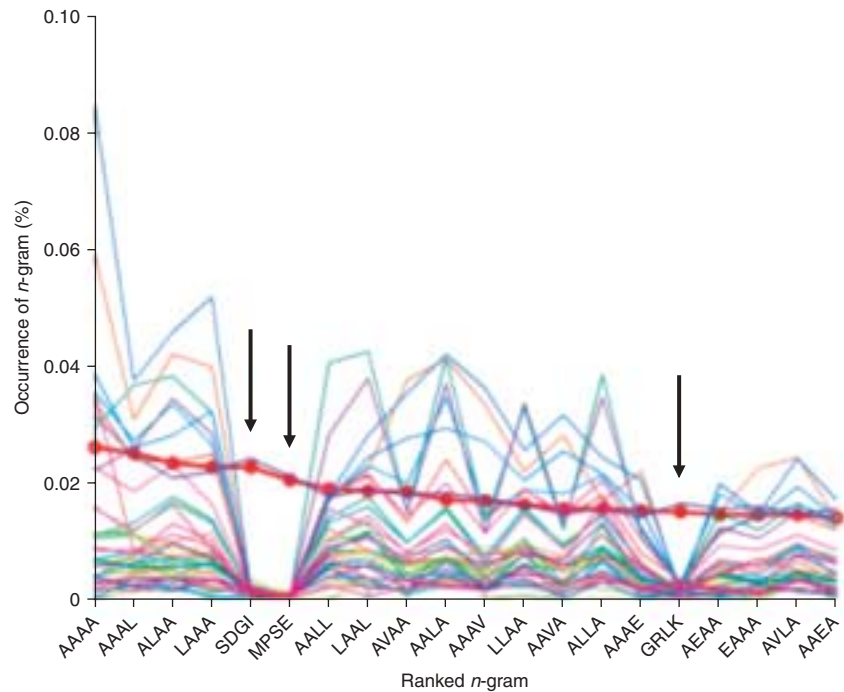
### Language Models

The computed  $n$ -grams form the input for  $n$ -gram language models  $P(w_i|w_1 \dots w_{i-1}) \approx P(w_i|w_{i-N+1} \dots w_{i-1})$ , where  $w_i$  is the probability (P) of observing the word  $i$ , which is computed using the Carnegie Mellon University/Cambridge Statistical Language Modeling (SLM) Toolkit.<sup>[43]</sup> The likelihood of observing a new sequence in comparison with a reference language model is reported. For  $n$ -grams that are absent in the reference model, the probabilities are estimated based on back-off models using lower values of  $n$ , up to  $n = 1$  (as described by Chen and Goodman<sup>[44]</sup>).

## Applications

### Dataset Comparisons: Identification of Genome Signature

One of the motivations for the development of the BLMT was to allow the formulation of new hypotheses using the analogy between language and biology. The first question we asked is if the sequences observed in proteins of different organisms are statistically similar or if organisms may be viewed as representations of different languages. We compared the  $n$ -gram frequencies of 44



**Fig. 2.** Distribution of amino acid  $n$ -grams with  $n = 4$  in *Neisseria meningitidis* in comparison with the distribution of the corresponding amino acids in 44 other organisms.<sup>[11]</sup>  $N$ -grams of *Neisseria* are plotted in descending order of their frequency in the genome (bold red line). The top 20 most frequent 4-grams in *Neisseria* are shown. Frequencies of corresponding  $n$ -grams from genomes of various other organisms are also shown (thin coloured lines). The second thin line closely following the bold red line corresponds to a different strain of *Neisseria meningitidis*. Arrows indicate the positions of 4-grams that are over-represented in *Neisseria* but are rare in other genomes.

different organisms using the  $n$ -gram comparison functions provided by BLMT and made the following observations:

1. A simple Markovian uni-gram (context-independent amino acid) model from the proteins of *Aeropyrum pernix* was trained. When the training set and test set contained different proteins but were selected from the same organism, a perplexity (a variation of cross-entropy) of 16.6 was observed, whereas data from other organisms varied from 16.8 to 21.9. Thus, even the simplest model can automatically detect the differences in the amino acid usage of different organisms.

2. We developed a modification of Zipf-like analysis that can reveal specific differences in  $n$ -grams in different organisms. First, the amino acid  $n$ -grams of a given length were sorted in descending order by frequency for the organism of choice. An example is shown in figure 2 for *Neisseria meningitidis* for  $n = 4$ . Remarkably, there are three  $n$ -grams (shown by red arrows in figure 2) that are among the top 20 most frequently occurring 4-grams in *Neisseria* but that are rare or absent in any of the other genomes. These highly idiosyncratic  $n$ -grams suggest motifs that are preferentially used in the particular organism. These motifs are highly statistically significant, as demonstrated by the fact that randomised sequences with an identical amino acid composition to the genomes

studied did not generate such motifs.<sup>[11]</sup> These motifs are not only statistically significant in the comparison across organisms but also within each organism itself. In *Neisseria* for example, the 4-grams SDGI and MPSE are highly over-represented compared with the frequencies expected based on the uni-gram distributions in *Neisseria*.<sup>[11]</sup> These differences are reflected in the observation that the observed frequencies of many  $n$ -grams are often more than ten standard deviations away from the mean.

3. While it is not known if these motifs correspond to similar or different substructures of proteins, we found that amino acid neighbour preferences are also different for different organisms, suggesting the possibility for underlying subtle changes in the mapping of sequences to structures of proteins. Figure 3 shows a plot of the preferences of amino acids with distances of up to five amino acids for two different organisms, *Aeropyrum pernix* and *Thermoplasma acidophilum*, which illustrates these differences.

#### Membrane Helix Boundary Prediction

Computational methods for protein secondary structure prediction are based on amino acid preferences in secondary structure elements.<sup>[45-47]</sup> Similarly, in the special case of transmembrane

helices, amino acid preferences also play a predominant role because of the importance of hydrophobicity for the transmembrane segments and the positive inside-out rule for the extramembranous parts of the proteins.<sup>[48]</sup> While the use of amino acid preferences can give an approximate boundary of transmembrane helices, some applications require knowledge of the exact positions of these boundaries. We used several  $n$ -gram approaches to address this question.

The problem is similar to topic segmentation in natural language, and we applied Yule's measure of association<sup>[42]</sup> to the membrane protein boundary prediction task based on its use in natural language processing.<sup>[49]</sup> Given a text with  $n$  different words, an  $n \times n$  table of Yule values for every pair of words is computed. The distribution of Yule values in the table differs for different categories of text, indicating the positions of the boundaries. In a model application to the G-protein coupled receptor (GPCR) family of membrane proteins, we found that Yule values can differentiate between transmembrane helices and loops connecting the helices.<sup>[13]</sup> Similar observations were made using other measures correlating  $n$ -grams, in particular the mutual information method.<sup>[14]</sup> Interestingly, minima in the mutual information correlate well with the transmembrane boundaries without using any sort of training data, hydrophobicity scales or other type of knowledge. This result suggests that  $n$ -gram correlations are able to capture structural features of proteins at the sequence level and can be used to discover these features automatically.<sup>[14]</sup>

We also adopted a language modelling approach. This method builds a language model for each 'topic' representing transmembrane helices and loops. Their performance in predicting the

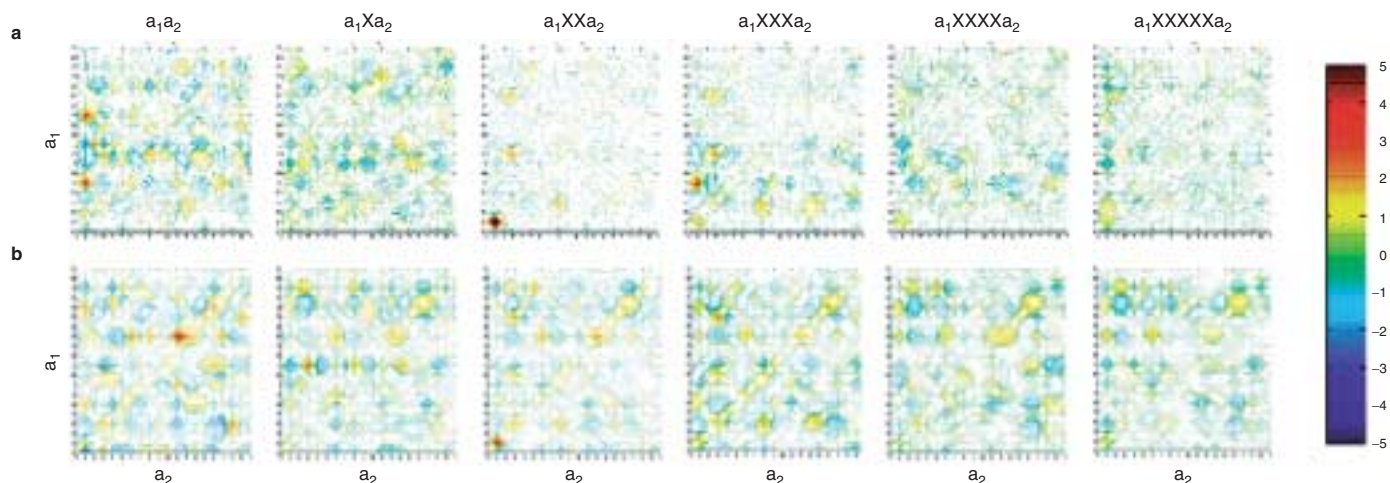
current amino acid is then compared, to determine whether a boundary occurs at the current position. The language models make use of only  $n$ -grams probabilities but, surprisingly, still produced promising results.<sup>[10,13,14]</sup>

### Protein Family Classification

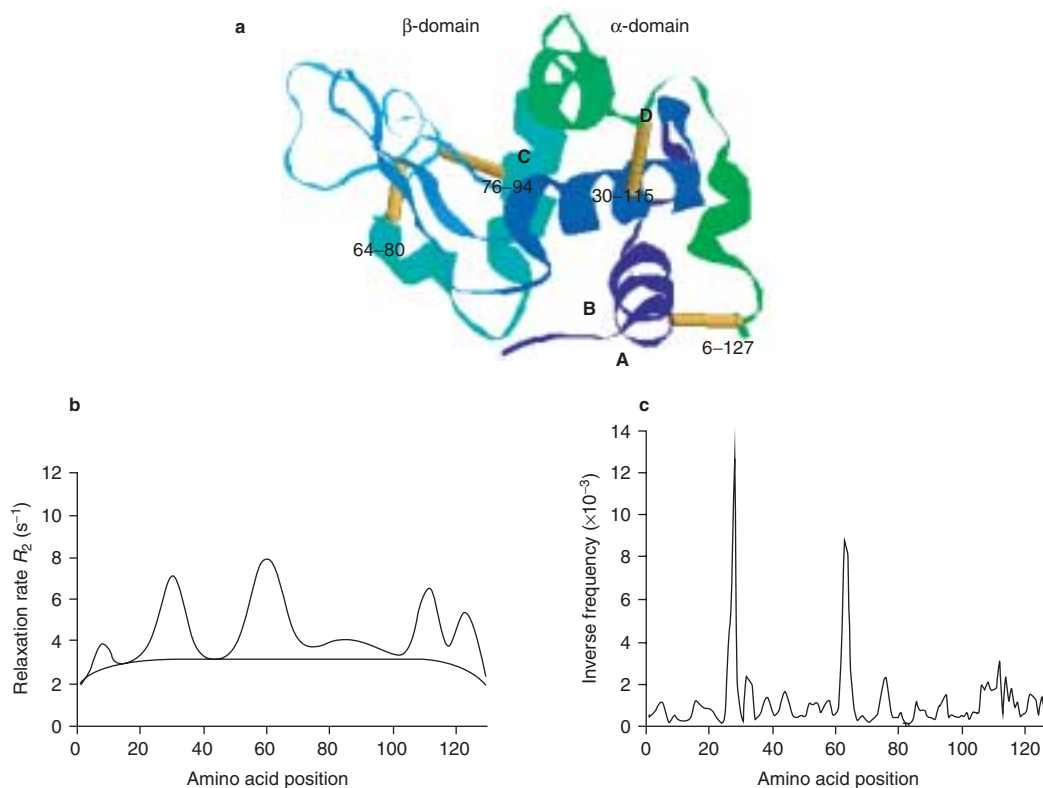
GPCRs are a superfamily of proteins and are particularly difficult to classify into families because of the extreme diversity among its members. A comparison of BLAST®, k-Nearest Neighbor (k-NN), HMMs and SVMs with alignment-based features has suggested that classifiers at the complexity of SVMs are needed to attain high accuracy.<sup>[50]</sup> However, we were able to show that the simple decision tree and naive Bayes classifiers in conjunction with chi-square feature selection on counts of  $n$ -grams perform extremely well, and the naive Bayes classifier even outperforms the SVM significantly.<sup>[10]</sup> We also generalised the utility of  $n$ -grams for high-accuracy classification of other protein families using the naive Bayes approach.<sup>[10,31]</sup> In line with these observations, Wu and co-workers have observed that neural networks perform well with  $n$ -gram features in the protein family classification task.<sup>[30]</sup>

### Protein Folding and Tertiary Structure Prediction

Prediction of folding initiation sites in proteins is a formidable task that requires novel approaches. We investigated if inverse frequencies might correlate with experimentally determined folding initiation sites in the protein-folding model system, lysozyme. Our hypothesis was based on the observation that, in natural languages, rare words carry the meaning of a text. Shown in figure



**Fig. 3.** Preferences between neighbouring amino acids in whole genomes. Positive and negative preferences between neighbouring amino acids separated by a distance of 0–5 residues are shown for *Aeropyrum pernix* (a) and *Thermoplasma acidophilum* (b). Red corresponds to positive preference and blue corresponds to negative preference (see colour bar).



**Fig. 4.** Location of folding initiation sites in the model protein, lysozyme. (a) 3-dimensional structure; (b) gaussian distributions fitted to experimentally observed transverse relaxation rates in the unfolded protein.<sup>[51]</sup> Large values above the flattish line (random cat model) indicate the presence of residual structure; and (c) Inverse trigram frequency in human lysozyme.

4 are inverse trigram frequencies plotted along the lysozyme sequence. Indeed, we observed a correlation between the locations of rare trigrams and the location of residual structure in the unfolded protein, as evidenced by maxima in relaxation rates measured in nuclear magnetic resonance (NMR) spectroscopic experiments. The statistical significance of this observation remains to be established by extension to other proteins, but lysozyme is the only protein for which the locations of folding initiation sites are known. However, the steady growth in the size of the Protein Data Bank will allow a systematic comparison between the sequences of  $n$ -grams and the number of structures that each  $n$ -gram can occur in. Such statistics are already beginning to be reported in the I-sites library<sup>[32]</sup> and in the analysis of sequences encoding certain types of structures.<sup>[33]</sup> Thus, it is expected that  $n$ -gram analysis may significantly contribute to the protein tertiary-structure prediction problem in the future.

## Conclusions and Future Work

Rapidly accumulating sequence data and accompanying structural and functional data provide new opportunities for statistical sequence analysis with the aim of identifying correlations between

these data. New datasets need to be analysed and global statistics computed. Currently, the popular methods for sequence analysis rely on alignment. However, alignment-based programs have limitations. They make the assumption that contiguity is conserved, which is not always the case. This has stirred interest in methods that do not rely on alignment for a number of bioinformatics tasks, in particular the use of  $n$ -grams. In this article, we presented a generic tool for the efficient calculation of  $n$ -gram statistics and provided several examples of the utility of this tool (BLMT) in various areas of bioinformatics. The BLMT provides a basis layer using powerful data structures optimised for fast search and retrieval of  $n$ -grams, computation of various  $n$ -gram statistics and comparison of such statistics across multiple datasets. We reviewed the interesting differences observed in comparing datasets at varying hierarchies, from describing whole-genome protein sequences to distinguishing soluble from transmembrane helices and the ability to detect transmembrane helix boundaries. In all cases, the observations made were significantly different from the corresponding randomised datasets. The next step is to quantify the utility of  $n$ -grams for the generation of predictive and, therefore, applicable models. It has already been shown that  $n$ -gram

features have unprecedented utility for protein sequence classification, to the extent that simple classifiers such as the naive Bayes outperform complex classifiers such as SVMs if the right features are selected. In principle, any type of prediction may utilise  $n$ -grams as features, either alone or in conjunction with other features. One as yet little explored area where  $n$ -grams may provide new predictive capabilities is the area of tertiary structure and folding pathway prediction. While a thorough and quantitative estimation of the degree of correlation of  $n$ -gram features with the structural and dynamic properties of proteins including folding initiation sites is still outstanding, the results of the four application areas described in this article suggest that  $n$ -gram features may be worthwhile to be explored as input in these and other predictive tasks.

## Acknowledgements

This research was supported by National Science Foundation grants NSF0225656 and NSF0225636, and the Sofya Kovalevskaya Program of the Alexander von Humboldt-Foundation/Zukunftsinvestitionsprogramm der Bundesregierung Deutschland.

The authors have no conflicts of interest directly relevant to the content of this review.

## References

- Baldi P. *Bioinformatics*. Cambridge (MA): MIT Press, 1998
- Durbin R, Eddy S, Krogh A, et al. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge: Cambridge University Press, 1998
- Searls DB. The language of genes. *Nature* 2002; 420 (6912): 211-7
- Baxevas AD, Ouellette BFF. *Bioinformatics: a practical guide to the analysis of genes and proteins*. New York: Wiley-Interscience, 1998
- Bolshoy A, Shapiro K, Trifonov EN, et al. Enhancement of the nucleosomal pattern in sequences of lower complexity. *Nucleic Acids Res* 1997; 25 (16): 3248-54
- Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 1997; 268: 78-94
- Gibas C, Jambeck P. *Developing bioinformatics computer skills*. Sebastopol (CA): O'Reilly & Associates, 2001
- Troyanskaya OG, Arbell O, Koren Y, et al. Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. *Bioinformatics* 2002; 18: 679-88
- Coin L, Bateman A, Durbin R. Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proc Natl Acad Sci U S A* 2003; 100: 4516-20
- Cheng BYM, Carbonell J, Klein-Seetharaman J. Protein classification based on text document classification techniques. *Proteins* 2004. In press
- Ganapathiraju M, Weisser D, Rosenfeld R, et al. Comparative  $n$ -gram analysis of whole-genome sequences. *Human Language Technologies Conference (HLT2002)*; 2002 Mar 24-27; San Diego (CA).
- Ganapathiraju M, Klein-Seetharaman J, Balakrishnan N, et al. Characterization of protein secondary structure using latent semantic analysis. *IEEE Signal Processing Magazine* 2004; 21 (3): 78-87
- Ganapathiraju M, Weisser D, Klein-Seetharaman J. Yule value tables from protein datasets. *SCI2004: 8th World Multi-Conference on Systemics, Cybernetics and Informatics*; 2004 Jul 18-21; Orlando (FL).
- Weisser D, Klein-Seetharaman J. Identification of fundamental building blocks in protein sequences using statistical association measures. *ACM Symposium on Applied Computing*; 2004 Mar 14-17; Nicosia, Cyprus. 154-61
- Liu Y, Carbonell J, Klein-Seetharaman J, Gopalakrishnan V. Comparison of probabilistic combination methods for protein secondary structure prediction. *Bioinformatics*. Epub 2004 Jun 24
- Erhan S, Marzolf T, Cohen L. Amino-acid neighborhood relationships in proteins: breakdown of amino-acid sequences into overlapping doublets, triplets and quadruplets. *Int J Biomed Comput* 1980; 11 (1): 67-75
- Karlin S, Bucher P, Brendel V, et al. Statistical methods and insights for protein and DNA sequences. *Annu Rev Biophys Biophys Chem* 1991; 20: 175-203
- Karlin S, Blaisdell BE, Bucher P. Quantile distributions of amino acid usage in protein classes. *Protein Eng* 1992; 5 (8): 729-38
- Karlin S, Burge C. Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc Natl Acad Sci U S A* 1996; 93 (4): 1560-5
- Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol* 1990; 215: 403-10
- Mantegna RN, Buldyrev SV, Goldberger AL, et al. Linguistic features of noncoding DNA sequences. *Phys Rev Lett* 1994; 73 (23): 3169-72
- Zipf GK. *Selective studies and the principle of relative frequency in language*. Cambridge (MA): Harvard University Press, 1932
- Miller GA, Newman EB. Tests of a statistical explanation of the rank-frequency relation for words in written English. *Am J Psychol* 1958; 71: 209-18
- Chatzidimitriou-Dreismann CA, Streffer RM, Larhammar D. Lack of biological significance in the 'linguistic features' of noncoding DNA: a quantitative analysis. *Nucleic Acids Res* 1996; 24 (9): 1676-81
- Tsonis AA, Elsner JB, Tsonis PA. Is DNA a language? *J Theor Biol* 1997; 184 (1): 25-9
- Niyogi P, Berwick RC. A note on Zipf's law, natural languages, and noncoding DNA regions. Cambridge (MA): Massachusetts Institute of Technology, Cambridge Artificial Intelligence Lab, 1995. Report no.: A024892
- Israeloff NE, Kagalenko M, Chan K. Can Zipf distinguish language from noise in noncoding DNA? [letter]. *Phys Rev Lett* 1996; 76 (11): 1976
- Li W. Statistical properties of open reading frames in complete contained genome sequences. *Comput Chem* 1999; 23 (3-4): 283-301
- Strait BJ, Dewey TG. The Shannon information entropy of protein sequences. *Biophys J* 1996; 71 (1): 148-55
- Wu C, Whitson G, McLarty J, et al. Protein classification artificial neural system. *Protein Sci* 1992; 1: 667-77
- Vries JK, Munshi R, Tobi D, et al. A sequence alignment-independent method for protein classification. *Appl Bioinformatics* 2004; 3 (2-3): 61-72
- Simons KT, Bonneau R, Ruczinski I, et al. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* 1999; Suppl. 3: 171-6
- Kuznetsov IB, Rackovsky S. On the properties and sequence context of structurally ambivalent fragments in proteins. *Protein Sci* 2003; 12: 2420-33
- Hucka M, Finney A, Sauro HM, et al. The ERATO Systems Biology Workbench: enabling interaction and exchange between software tools for computational biology. *Pac Symp Biocomput* 2002; 7: 450-61
- Manber U, Myers G. Suffix arrays: a new method for on-line string searches. *SIAM J Comput* 1993; 22 (5): 935-48
- Delcher AL, Kasif S, Fleischman RD, et al. Alignment of whole genomes. *Nucleic Acids Res* 1999; 27: 2369-76
- Sadakane K, Shibuya T. Indexing huge genome sequences for solving various problems. *Genome Inform Ser Workshop Genome Inform* 2001; 12: 175-83
- Mandel-Gutfreund Y, Gregoret LM. On the significance of alternating patterns of polar and non-polar residues in beta-strands. *J Mol Biol* 2002; 323: 453-61
- Dorohonceanu B, Nevill-Manning CG. Accelerating protein classification using suffix trees. *Proc Int Conf Intell Syst Mol Biol* 2000; 8: 128-33
- Bejerano G, Yona G. Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics* 2001; 17: 23-43
- Kasai T, Lee G, Arimura H. Linear-time longest-common-prefix computation in suffix arrays and its applications. *12th Annual Symposium on Combinatorial*

- Pattern Matching: CPM-2001; 2001 Jul 1-4; Jerusalem. Heidelberg: Springer-Verlag, 2001: 181-92
42. Bishop YMM, Fienberg SE, Holland PW. Discrete multivariate analysis. Cambridge (MA): MIT Press, 1975
  43. Clarkson PR, Rosenfeld R. Statistical language modeling using the CMU-Cambridge toolkit. Proceedings ESCA Eurospeech; 1997 Sep 23-25; Rhodes, Greece.
  44. Chen SF, Goodman J. An empirical study of smoothing techniques for language modeling. Proceedings of the 34th Conference of the Association for Computational Linguistics (ACL96); 1996 Jun 23-28; Santa Cruz (CA).
  45. Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 1978; 47: 45-148
  46. Richardson JS, Richardson DC. Amino acid preferences for specific locations at the ends of alpha helices. Science 1988; 240: 1648-52
  47. Cai YD, Liu XJ, Chou KC, et al. Prediction of protein secondary structure content by artificial neural network. J Comput Chem 2003; 24: 727-31
  48. Chen CP, Rost B. State-of-the-art in membrane protein prediction. Appl Bioinformatics 2002; 1 (1): 21-35
  49. Cai C, Rosenfeld R, Wasserman L. Exponential language models, logistic regression, and semantic coherence. Proceedings of the NIST/DARPA Speech Transcription Workshop; 2000 May 16-19; Adelphi (MD).
  50. Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. Bioinformatics 2002; 18: 147-59
  51. Klein-Seetharaman J, Oikawa M, Grimshaw SB, et al. Long-range interactions within a nonnative protein. Science 2002; 295 (5560): 1719-22

---

Correspondence and offprints: Dr *Judith Klein-Seetharaman*, Department of Pharmacology, University of Pittsburgh School of Medicine, 200 Lothrop St, Pittsburgh, PA 15261, USA.  
E-mail: [judithks@cs.cmu.edu](mailto:judithks@cs.cmu.edu)