biogram: a toolkit for biological n-gram analysis

Michał Burdukiewicz[1], Piotr Sobczyk[2], Paweł Mackiewicz[1] and Małgorzata Kotulska[3]
*michalburdukiewicz@gmail.com

[1]University of Wrocław, Department of Genomics
[2]Wrocław University of Technology, Faculty of Pure and Applied Mathematics
[3]Wrocław University of Technology, Department of Biomedical Engineering

Introduction

N-grams (k-tuples) are vectors of n characters derived from input sequence(s). They may form continuous sub-sequences or be discontinuous. Important n-gram parameter is its position. Instead of just counting n-grams, one may want to count how many n-grams occur at a given position in multiple (e.g. related) sequences.
Originally developed for natural language processing, n-grams are also used in genomics (Fang et al., 2011), transcriptomics (Wang et al., 2014) and proteomics (Guo et al., 2014).

	P1	P2	P3	P4	P5	P6
S1	C	C	C	T	C	C
S2	A	G	T	T	T	C
S3	G	A	G	G	C	T

Sample sequences. S - sequence, P - postion.

	A	C	G	T
S1	0	5	0	1
S2	1	1	1	3
S3	1	1	3	1

Unigram counts.

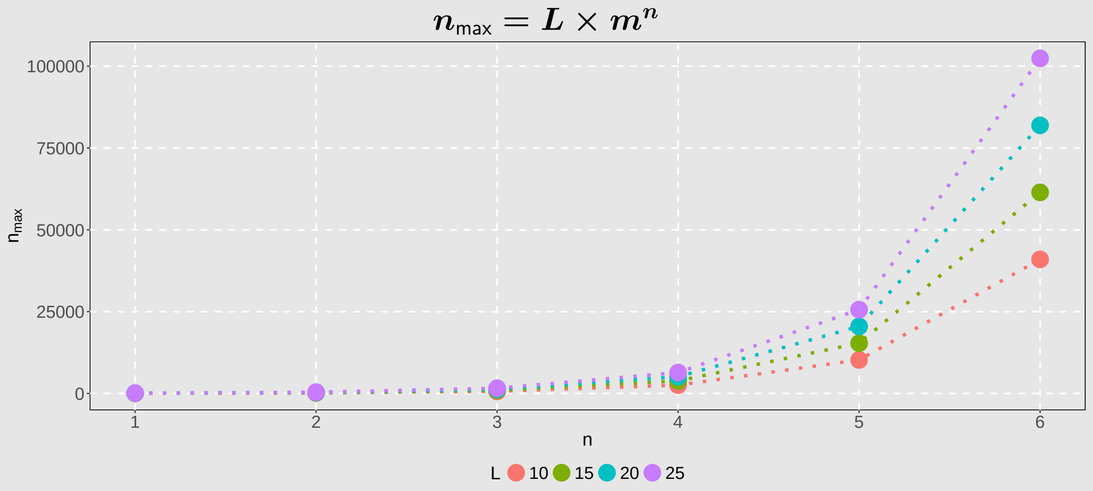
	P1_A	P2_A	P3_A	P4_A	P5_A	P6_A	P1_C	P2_C	P3_C	P4_C	P5_C	P6_C	P1_G
S1	0	0	0	0	0	0	1	1	1	0	1	1	0
S2	1	0	0	0	0	0	0	0	0	0	0	1	0
S3	0	1	0	0	0	0	0	0	0	0	1	0	1

A fraction of possible unigrams with position information.

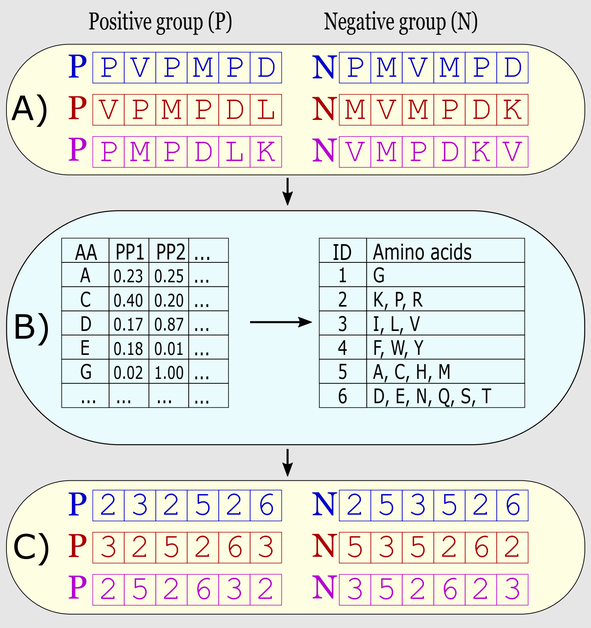
Curse of dimensionality

Even when we limit ourselves to only continuous positioned n-grams build on $\boldsymbol{m}$ possible characters, feature space growths rapidly with the number of elements in n-gram ($\boldsymbol{n}$) and the length of the sequence ($\boldsymbol{L}$).
The number of possible positioned n-grams:

$$\boldsymbol{n}_{\text{max}} = \boldsymbol{L} \times \boldsymbol{m}^{\boldsymbol{n}}$$



Reducing number of n-grams



A) Input data: peptides with a known status (e.g. amyloid/nonamyloid).
B) Creation of an encoding using a combination of physicochemical properties (PP).
C) Reduction of the amino acid alphabet according to an encoding. The number of possible n-grams is reduced, because $\boldsymbol{m}$ is smaller (e.g. in this case $\boldsymbol{m}$ is reduced from 20 to 6).

Bibliography

Fama, C., Dennison, S. R., Quintas, A., and Phoenix, D. A. (2015). Prediction of Peptide and Protein Propensity for Amyloid Formation. *PLOS ONE*, 10(8):e0134679.

Fang, Y.-C., Lai, P.-T., Dai, H.-J., and Hsu, W.-L. (2011). Meinfotext 2.0: gene methylation and cancer relation extraction from biomedical literature. *BMC Bioinformatics*, 12(1):471.

Garbuzynskiy, S. O., Lobanov, M. Y., and Galzitskaya, O. V. (2010). FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics (Oxford, England)*, 26(3):326–332.

Guo, S.-H., Deng, E.-Z., Xu, L.-Q., Ding, H., Lin, H., Chen, W., and Chou, K.-C. (2014). inuc-pseknc: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics*, 30(11):1522–1529.

Lehmann, E. (1986). *Testing statistical hypotheses*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley.

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8:785–786.

Walsh, I., Seno, F., Tosatto, S. C. E., and Trovato, A. (2014). PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research*, 42(W1):W301–W307.

Wang, Y., Liu, L., Chen, L., Chen, T., and Sun, F. (2014). Comparison of metatranscriptomic samples based on ¡italic¿k-¡/italic¿tuple frequencies. *PLoS ONE*, 9(1):e84348.

Selection of important n-grams

Model and statistic independent permutation tests can be used to filter features obtained through counting n-grams.
During a permutation test class labels are randomly exchanged during computation of a significance statistic. p-values are defined as:

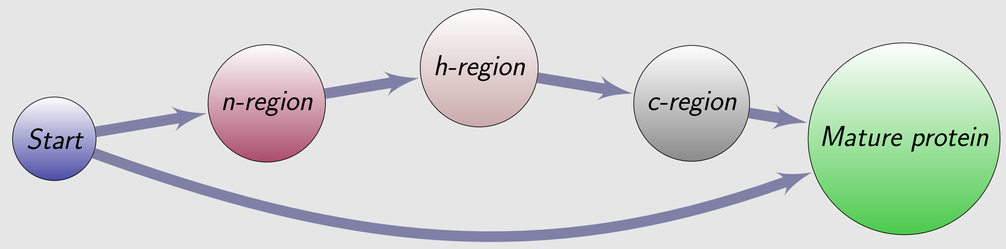
$$\text{p-value} = \frac{\boldsymbol{N}_{\boldsymbol{T_P} > \boldsymbol{T_R}}}{\boldsymbol{N}}$$

where $\boldsymbol{N}_{\boldsymbol{T_P} > \boldsymbol{T_R}}$ is number of times when $\boldsymbol{T_P}$ (permuted test statistic) was more extreme than $\boldsymbol{T_R}$ (test statistic for non-permuted data).
Permutation tests are computationally expensive (especially considering precise estimation of small p-values, because the number of permutations is inversely proportional to the interval between p-values).
Quick **P**ermutation **T**est (QuiPT) thanks to the unique parameterization replaces a permutation test with the exact two-sided Fisher's test speeding it up and reducing the computation cost (Lehmann, 1986).

signalHsmm - prediction of signal peptides

Signal peptides are n-terminal guiding sequences constitution of three specific regions: n-, h- and c-region. Using the n-gram approach we created *signalHsmm*, a software for prediction of signal peptides.

signalHsmm has two models representing respectively proteins with and without signal peptides. The probabilities of both fits and predicted cleavage site constitute the software output.



signalHsmm benchmark

	Sensitivity	Specificity	MCC	AUC
signalP 4.1 (no tm) (Petersen et al., 2011)	0.8235	**0.9100**	0.6872	0.8667
signalP 4.1 (tm) (Petersen et al., 2011)	0.6471	0.9431	0.6196	0.7951
signalHsmm	**0.9804**	0.8720	**0.7409**	**0.9262**
signalHsmm (raw aa)	0.8431	0.9005	0.6853	0.8718

Comparison of performance measures for different classifiers according to singal peptide-containing proteins from members of *Plasmodiidae*.

Thanks to the usage of reduced amino acid alphabet, *signalHsmm* better recognizes signal peptides belonging to *Plasmodiidae* which are characterized by atypical amino acid composition.

AmyloGram

Amyloids are proteins associated with the number of clinical disorders (e.g., Alzheimer's, Creutzfeldt-Jakobs and Huntingtons diseases). We created *AmyloGram*, n-based predictor of amyloidogenicity using decision rules extracted by random forests.

Classifier	AUC	MCC	Sensitivity	Specificity
AmyloGram	**0.8972**	**0.6307**	0.8658	0.7889
PASTA (Walsh et al., 2014)	0.8550	0.4291	0.3826	0.9519
FoldAmyloid (Garbuzynskiy et al., 2010)	0.7351	0.4526	0.7517	0.7185
APPNN (Fama et al., 2015)	0.8343	0.5823	**0.8859**	0.7222

Results of benchmark on *pep424* data set provided by creators of PASTA2. The peptides from this data set were not included in the training data set of *AmyloGram*.

Summary and availability

The n-gram analysis creates versatile classifiers able to extract more universal decision rules (e.g. *signalHsmm*, which is able to also predict signal peptides in atypical organisms) or better detect specific proteins. Nonetheless, despite computational quickness provided by the QuiPT method, curse of dimensionality limits n-gram methods to the analysis of shorter sequences.
Our software is avaible as web-servers:

signalHsmm web-server: `smorfland.uni.wroc.pl/signalHsmm`.
AmyloGram web-server: `smorfland.uni.wroc.pl/amylogram`.