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	G	C	C	T	A	A
S2	T	G	C	G	G	A
S3	T	T	G	T	C	G

Sample sequences. S - sequence, P - postion.

	A	C	G	T
S1	2	2	1	1
S2	1	1	3	1
S3	0	1	2	3

Unigram counts.

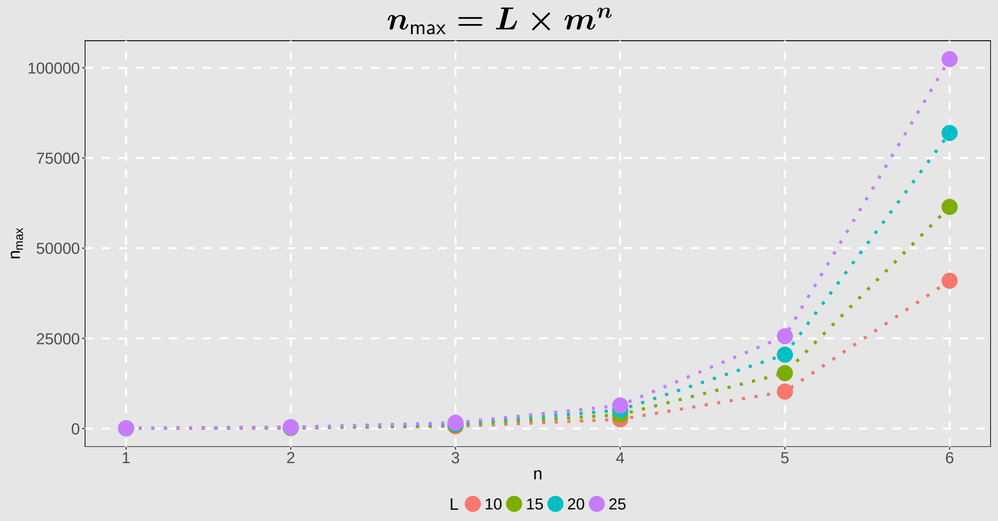
	P1_A	P2_A	P3_A	P4_A	P5_A	P6_A	P1_C	P2_C	P3_C	P4_C	P5_C	P6_C	P1_G
S1	0	0	0	0	1	1	0	1	1	0	0	0	1
S2	0	0	0	0	0	1	0	0	1	0	0	0	0
S3	0	0	0	0	0	0	0	0	0	0	1	0	0

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:

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



Feature selecting permutation tests

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).

QuiPT concept

In each permutation, for every observation, there are four possible results:

$$\boldsymbol{P}(\boldsymbol{Target}, \boldsymbol{Feature}) = (1,1)) = \boldsymbol{p} \cdot \boldsymbol{q}$$

$$\boldsymbol{P}(\boldsymbol{Target}, \boldsymbol{Feature}) = (1,0)) = \boldsymbol{p} \cdot (1 - \boldsymbol{q})$$

$$\boldsymbol{P}(\boldsymbol{Target}, \boldsymbol{Feature}) = (0,1)) = (1 - \boldsymbol{p}) \cdot \boldsymbol{q}$$

$$\boldsymbol{P}(\boldsymbol{Target}, \boldsymbol{Feature}) = (0,0)) = (1 - \boldsymbol{p}) \cdot (1 - \boldsymbol{q})$$

Where $\boldsymbol{p}$ and $\boldsymbol{q}$ are fractions of positive observations in target and feature respectively. An another view at permutation test is therefore that we get a contingency table, which is to be tested for independence. Computing probability of a such table with two constraints, $\boldsymbol{n}_{1,\cdot} = \boldsymbol{n}_{1,1} + \boldsymbol{n}_{1,0}$ and $\boldsymbol{n}_{\cdot,1} = \boldsymbol{n}_{1,1} + \boldsymbol{n}_{0,1}$, and conditioning on $\boldsymbol{n}_{1,1}$, leads to hypergeometric distribution. $\boldsymbol{n}_{i,j}$ denotes number of observations for which $(\boldsymbol{Target}, \boldsymbol{Feature}) = (i,j)$
Thanks to this parametrization we replace a permutation test with the exact two-sided Fisher's test (Lehmann, 1986).

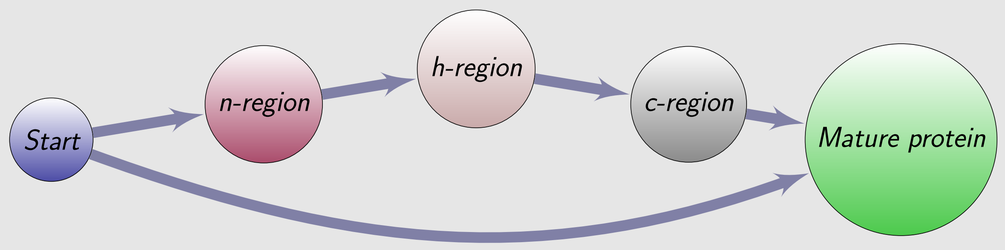
Signal peptides

Secretory signal peptides:

- are short (20-30 residues) N-terminal amino acid sequences,
- direct a protein to the endomembrane system and next to the extracellular localization,
- possess three distinct domains with variable length and characteristic amino acid composition (Hegde and Bernstein, 2006).
- are universal enough to direct properly proteins in different secretory systems; artifically introduced bacterial signal peptides can guide proteins in mammals (Nagano and Masuda, 2014) and plants (Moeller et al., 2009),
- tag among others hormons, immune system proteins, structural proteins, and metabolic enzymes.

Signal peptide prediction

During the test phase, each protein is fitted to two HSMMs representing respectively proteins with and without signal peptides. The probabilities of both fits and predicted cleavage site constitute the software output.



Summary and availability

AmyloGram is a model-independent predictor of amylogenicity. Instead, it provides insight on the structural features present in the hot-spots. Moreover, AmyloGram recognises amylogenic sequences better than existing predictors.
AmyloGram web-server: `smorfland.uni.wroc.pl/amylogram`.

Bibliography

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.

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.

Hegde, R. S. and Bernstein, H. D. (2006). The surprising complexity of signal sequences. *Trends in Biochemical Sciences*, 31(10):563–571.

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

Moeller, L., Gan, Q., and Wang, K. (2009). A bacterial signal peptide is functional in plants and directs proteins to the secretory pathway. *Journal of Experimental Botany*, 60(12):3337–3352.

Nagano, R. and Masuda, K. (2014). Establishment of a signal peptide with cross-species compatibility for functional antibody expression in both escherichia coli and chinese hamster ovary cells. *Biochemical and Biophysical Research Communications*, 447(4):655 – 659.

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.