biogram: a toolkit for n-gram analysis

Michał Burdukiewicz[1], Piotr Sobczyk[2], Małgorzata Kotulska[3], Paweł Mackiewicz[1]

[1]University of Wrocław, Department of Genomics, Poland

[2]Wrocław University of Technology, Institute of Mathematics and Computer Science, Poland

[3]Wrocław University of Technology, Department of Biomedical Engineering, Poland

Outline

1. biogram
   - n-grams
   - Encoding of amino acids
   - Quick Permutation Test (QuiPT)
2. Case study 1: amyloid prediction
3. Case study 2: signal peptide prediction
4. Conclusion
5. Availability

Outline

1. biogram
   - n-grams
   - Encoding of amino acids
   - Quick Permutation Test (QuiPT)
2. Case study 1: amyloid prediction
3. Case study 2: signal peptide prediction
4. Conclusion
5. Availability

Aim: convert **bio**logical sequences to n-**grams**, continuous or discontinuous sub-sequences.

	X1	X2	X3	X4	X5	X6
1	a	c	t	c	a	a
2	g	g	g	c	a	c
3	t	g	t	c	g	t

Sample sequences.

a	c	g	t
3	2	0	1
1	2	3	0
0	1	2	3

Unigrams.

X1_a_0	X2_a_0	X3_a_0	X4_a_0	X5_a_0	X6_a_0	X1_c_0
1	0	0	0	1	1	0
0	0	0	0	1	0	0
0	0	0	0	0	0	0

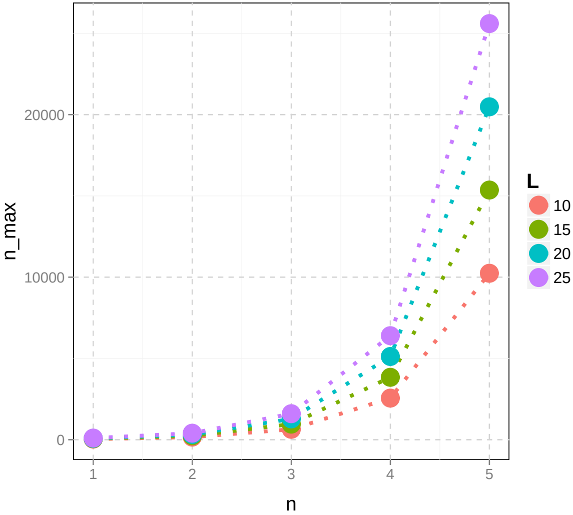
A fraction of possible unigrams with position information.

Positioned n-gram data is binary.

Number of possible positioned n-grams:

$$n_{max} = L \times m^n$$

- n_max: total number of n-grams.
- L: length of the sequence.
- m: number of unique symbols in the alphabet.
- n: length of the n-gram.


n_max
20000
10000
0
1
2
3
4
5
n
L
10
15
20
25

Encodings based on n-grams are cumbersome to use without the reduction of the dimensionality. Solutions:

- Reduce alphabet (in case of amino acid sequences).
- Filter features.

Existing encodings of amino acids:

- Orthogonal (20 bits, Ala = 10000000000000000000, Cys = 01000000000000000000, . . .).
- Orthogonal (5 bits, Ala = 00001, Cys = 00011, . . .).
- Exchange groups.

Drawbacks:

- Does not take into account relationships between amino acids (orthogonal encodings) or employs only selected relationships (exchange group).
- parse encoding enforces larger data sets, which hinders their management and analysis (Lin, May, & Taylor, 2002).

Amino acids may be assigned to groups based on their physicochemical similarity.
Every problem may have its own set of important physicochemical properties.

The encoding distance between **A** and **B** is defined as the minimum number of amino acids that have to be shifted between subgroups of encoding **A** to make it identical to **B** (order of subgroups in the encoding and amino acids in a group is unimportant).

Group	Elements
1	a, b, c
2	d, e

Encoding **A**.

Group	Elements
1	a, b
2	d, e
3	c

Encoding **B**.

The encoding distance between **A** and **B** is 1 (element *c* must be moved from Group 3 to Group 1).

1. Calculate test statistic for the given positioned n-gram and etiquettes (T_R).
2. Permute counts of n-grams and calculate permuted test statistic (T_P).
3. Repeat step 2. N times.
4. Calculate p-value using:

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

$N_{T_P > T_R}$ is number of times when T_P was bigger than T_R

Advantages:

- Model independent.

Advantages:

- Model independent.
- Statistic independent.

Drawbacks:

- Computationally expensive (number of cases, number of features).

Drawbacks:

- Computationally expensive (number of cases, number of features).
- Single feature analysis (no feature interaction).

Drawbacks:

- Computationally expensive (number of cases, number of features).
- Single feature analysis (no feature interaction).
- Unfeasible precise estimation of low p-values (the number of permutations is inversely proportional to the interval between p-values).

The binary positioned n-gram data tabulated by binary label can be easily described in 2d contingency table.

sequence ID	feature	target
1	1	0
2	1	0
3	0	0
4	1	1
5	0	1
...	...	...

Positioned n-grams with a label.

sequence ID	feature	target
1	1	0
2	1	0
3	0	0
4	1	1
5	0	1
...	...	...

Positioned n-grams with a label.

	target	feature
0	$n_{1,1}$	$n_{1,0}$
1	$n_{0,1}$	$n_{0,0}$

Contingency table.

Test statistics used by QuiPT (information gain, Kullback-Leibler divergence) measure inbalance of contingency tables.

The probability of certain contingency table is given as the conditional distribution, as impose restrictions on two parameters $n_{\cdot,1}$ and $n_{1,\cdot}$. The test statistic is computed for each possible value of $n_{1,1}$.

Advantages over permutation test

- Speed.

Advantages over permutation test

- Speed.
- Using the exact distribution of possible values of the criterion QuiPT yields precise small p-values without increasing the computation time.

Outline

1. biogram
   - n-grams
   - Encoding of amino acids
   - Quick Permutation Test (QuiPT)
2. Case study 1: amyloid prediction
3. Case study 2: signal peptide prediction
4. Conclusion
5. Availability

Amyloids:

- short proteins associated with the number of clinical disorders, for example Alzheimer's or Creutzfeldt-Jakob's diseases,

Amyloids:

- short proteins associated with the number of clinical disorders, for example Alzheimer's or Creutzfeldt-Jakob's diseases,
- create harmful zipper-like -structures through characteristic short subsequences of amino acids (hot-spots).

1. Nine scales representing properties important in the amylogenicity: hydrophobicity, size polarity and solvent accessibility from AAIndex database (Kawashima et al., 2008) were chosen. Additionally, two frequencies of forming contact sites (Wozniak & Kotulska, 2014) were added. All scales were normalized.

1. Nine scales representing properties important in the amylogenicity: hydrophobicity, size polarity and solvent accessibility from AAIndex database (Kawashima et al., 2008) were chosen. Additionally, two frequencies of forming contact sites (Wozniak & Kotulska, 2014) were added. All scales were normalized.
2. All combinations of characteristics (each time selecting only one scale per the property) were clustered using Euclidean distance and Ward's method.

1. Nine scales representing properties important in the amylogenicity: hydrophobicity, size polarity and solvent accessibility from AAIndex database (Kawashima et al., 2008) were chosen. Additionally, two frequencies of forming contact sites (Wozniak & Kotulska, 2014) were added. All scales were normalized.
2. All combinations of characteristics (each time selecting only one scale per the property) were clustered using Euclidean distance and Ward's method.
3. Each clustering was divided into 3 to 6 groups creating 144 encodings of amino acids.

1. Nine scales representing properties important in the amylogenicity: hydrophobicity, size polarity and solvent accessibility from AAIndex database (Kawashima et al., 2008) were chosen. Additionally, two frequencies of forming contact sites (Wozniak & Kotulska, 2014) were added. All scales were normalized.
2. All combinations of characteristics (each time selecting only one scale per the property) were clustered using Euclidean distance and Ward's method.
3. Each clustering was divided into 3 to 6 groups creating 144 encodings of amino acids.
4. Redundant 51 encodings (identical to other encodings) were removed.

1. Sequences shorter than 6 amino acids were discarded.

1. Sequences shorter than 6 amino acids were discarded.
2. From each sequence overlapping 6-grams were extracted. All n-grams were labelled as their sequence of the origin (e.g. all 6-grams extracted from amyloid sequence were labelled as positive).

1. Sequences shorter than 6 amino acids were discarded.
2. From each sequence overlapping 6-grams were extracted. All n-grams were labelled as their sequence of the origin (e.g. all 6-grams extracted from amyloid sequence were labelled as positive).
3. For each encoding features were filtered by the QuiPT and used to train the Random Forests (Liaw & Wiener, 2002). This procedure was performed independently on three training sets: a) 6 amino acids, b) 10 amino acids or shorter, c) 15 amino acids or shorter creating three classifiers.

1. Sequences shorter than 6 amino acids were discarded.
2. From each sequence overlapping 6-grams were extracted. All n-grams were labelled as their sequence of the origin (e.g. all 6-grams extracted from amyloid sequence were labelled as positive).
3. For each encoding features were filtered by the QuiPT and used to train the Random Forests (Liaw & Wiener, 2002). This procedure was performed independently on three training sets: a) 6 amino acids, b) 10 amino acids or shorter, c) 15 amino acids or shorter creating three classifiers.
4. All classifiers were evaluated in the 5-fold cross-validation eight times. The sequence was labelled as positive (amylogenic), if at least one 6-gram was assessed as amylogenic.

Training length	Number of groups	AUC	Specificity	Sensitivity
6	4	0.8183	0.9014	0.5038
<16	6	0.8320	0.5186	0.9195

Encodings with the best sensitivity and specificity.

The committee of the best specificity and best sensitivity classifiers has overall 0.8911 AUC, 0.7473 sensitivity and 0.8684 specificity.

The committee was compared to the two best-performing classifiers.

Benchmark of three best-performing classifiers on Pep424 data set (Walsh et al., 2014).

Name	AUC
PASTA 2.0 (Walsh et al., 2014)	0.8573
Committee	0.8390
FoldAmyloid (Garbuzynskiy et al., 2010)	0.8331

Outline

1. biogram
   - n-grams
   - Encoding of amino acids
   - Quick Permutation Test (QuiPT)
2. Case study 1: amyloid prediction
3. Case study 2: signal peptide prediction
4. Conclusion
5. Availability

Secretory signal peptides:

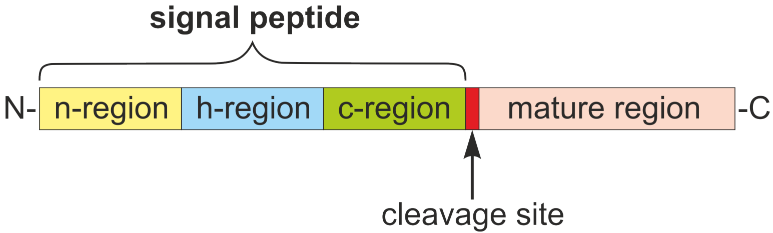
- are short (20-30 residues) N-terminal amino acid sequences,

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,

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 specific amino acid composition (Hegde & Bernstein, 2006).



Organization of signal peptide

Hidden semi-Markov models assumptions (Rabiner, 1989; Koski, 2001):

- the current region (state) of the sequence (process) depends on the previous region,

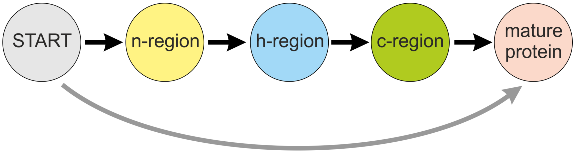
Hidden semi-Markov models assumptions (Rabiner, 1989; Koski, 2001):

- the current region (state) of the sequence (process) depends on the previous region,
- regions may be only indirectly determined using amino acid residues (observations),

Hidden semi-Markov models assumptions (Rabiner, 1989; Koski, 2001):

- the current region (state) of the sequence (process) depends on the previous region,
- regions may be only indirectly determined using amino acid residues (observations),
- probability of staying in a region is modeled by a probability distribution.

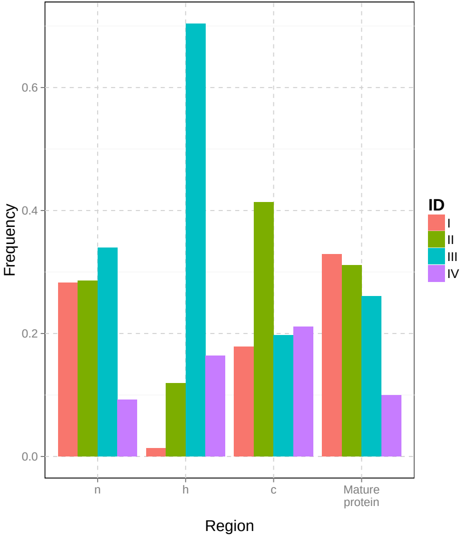
signalHsmm predictive model



During the test phase, each protein was fitted to two models. The outcome consists of probabilities that a particular residue belongs to a given model and predicted cleavage site.

The best AUC encoding.

ID	Amino acids
I	D, E, H, K, N, Q, R
II	G, P, S, T, Y
III	F, I, L, M, V, W
IV	A, C



Comparison of Area Under the Curve, H-measure and Matthews Correlation Coefficient for different classifiers considering only proteins belonging to Plasmodiidae.

Software name	AUC	Sensitivity	Specificity
signalP 4.1 (no tm) (Petersen et al., 2011)	0.8356	0.7745	0.8966
signalP 4.1 (tm) (Petersen et al., 2011)	0.7928	0.6471	0.9385
PrediSi (Hiller et al., 2004)	0.6597	0.3725	0.9469
Phobius (Käll et al., 2004)	0.7963	0.6765	0.9162
Philius (Reynolds et al., 2008)	0.7753	0.6176	0.9330
signalHsmm-2010	**0.9340**	**1.0000**	0.8436
signalHsmm-1989	0.9326	0.9510	**0.8631**

Outline

1. biogram
   - n-grams
   - Encoding of amino acids
   - Quick Permutation Test (QuiPT)
2. Case study 1: amyloid prediction
3. Case study 2: signal peptide prediction
4. Conclusion
5. Availability

biogram is a flexible toolkit for analysis of biological sequences. It reduces efficiently feature space by extracting important features and removing redundant information.

Outline

1. biogram
   - n-grams
   - Encoding of amino acids
   - Quick Permutation Test (QuiPT)
2. Case study 1: amyloid prediction
3. Case study 2: signal peptide prediction
4. Conclusion
5. Availability

biogram R package:
`http://cran.r-project.org/web/packages/biogram/`

biogram source: `https://github.com/michbur/biogram`

Garbuzynskiy, S. O., Lobanov, M. Y., & Galzitskaya, O. V. (2010, February). FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics (Oxford, England)*, *26*(3), 326–332. doi: 10.1093/bioinformatics/btp691

Hegde, R., & Bernstein, H. (2006). The surprising complexity of signal sequences. *Trends Biochem. Sci.*, *31*(10), 563–571. doi: 10.1016/j.tibs.2006.08.004

Hiller, K., Grote, A., Scheer, M., Münch, R., & Jahn, D. (2004). PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.*, *32*(suppl 2), W375–W379. Retrieved 2014-05-28, from `http://nar.oxfordjournals.org/content/32/suppl_2/W375` doi: 10.1093/nar/gkh378

Käll, L., Krogh, A., & Sonnhammer, E. (2004). A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, *338*(5), 1027–1036. doi: 10.1016/j.jmb.2004.03.016

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., & Kanehisa, M. (2008, January). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research*, *36*(suppl 1), D202–D205. Retrieved 2015-07-27, from `http://nar.oxfordjournals.org/content/36/suppl_1/D202` doi: 10.1093/nar/gkm998

Koski, T. (2001). *Hidden markov models for bioinformatics*. Springer.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from `http://CRAN.R-project.org/doc/Rnews/`

Lin, K., May, A. C., & Taylor, W. R. (2002). Amino acid encoding schemes from protein structure alignments: multi-dimensional vectors to describe residue types. *J Theor Biol*, *216*(3), 361-65.

Petersen, T., Brunak, S., Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, *8*(10), 785–786. doi: 10.1038/nmeth.1701

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. doi: 10.1109/5.18626

Reynolds, S., Käll, L., Riffle, M., Bilmes, J., & Noble, W. (2008). Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput. Biol.*, *4*(11), e1000213. doi: 10.1371/journal.pcbi.1000213

Walsh, I., Seno, F., Tosatto, S. C. E., & Trovato, A. (2014, May). PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research*, gku399. Retrieved 2015-09-15, from `http://nar.oxfordjournals.org/content/early/2014/05/21/nar.gku399` doi: 10.1093/nar/gku399

Wozniak, P. P., & Kotulska, M. (2014). Characteristics of protein residue-residue contacts and their application in contact prediction. *Journal of Molecular Modeling*, *20*(11). Retrieved 2015-07-27, from `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4221654/` doi: 10.1007/s00894-014-2497-9

Comparison of Area Under the Curve, Sensitivity, Specificity and Matthews Correlation Coefficient for different classifiers.

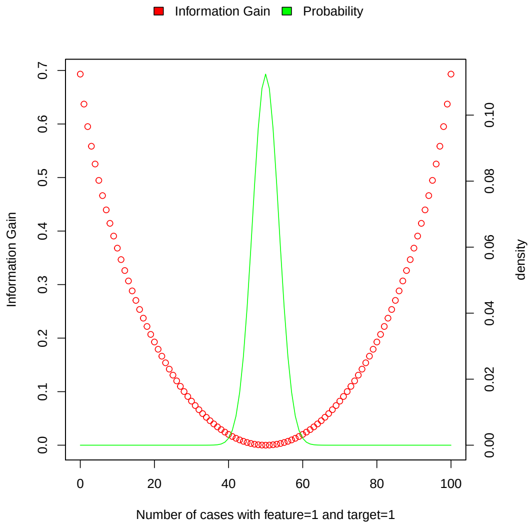
Software name	AUC	Sensitivity	Specificity
signalP 4.1 (no tm) (Petersen et al., 2011)	0.9416	**0.9720**	0.9112
signalP 4.1 (tm) (Petersen et al., 2011)	**0.9673**	0.9579	**0.9766**
PrediSi (Hiller et al., 2004)	0.8949	0.9065	0.8832
Phobius (Käll et al., 2004)	0.9509	0.9673	0.9346
Philius (Reynolds et al., 2008)	0.9369	0.9533	0.9206
signalHsmm-2010	0.9526	0.9533	0.8832
signalHsmm-1989	0.9562	0.9626	0.8972

If probability that target equals 1 is p and probability that feature equals 1 is q and feature and target are independent then each of them has the following probabilities

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

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

- $n_{1,1}$ is from range $[0, min(n_{\cdot,1}, n_{1,\cdot})]$.
- The probability of certain contingency table is given as the conditional distribution, as impose restrictions on two parameters $n_{\cdot,1}$ and $n_{1,\cdot}$.
- The test statistic is computed for each possible value of $n_{1,1}$.
- The distribution of test statistics under hypothesis that target and feature are independant is computed using values from 3.



	Target	Feature	Freq
1	0	0	50
2	1	0	50
3	0	1	50
4	1	1	50