Simplified alphabets in protein analysis

Michał Burdukiewicz

Department of Genomics, University of Wrocław

Outline

Simplified alphabets

Signal peptides

Methodology

Results

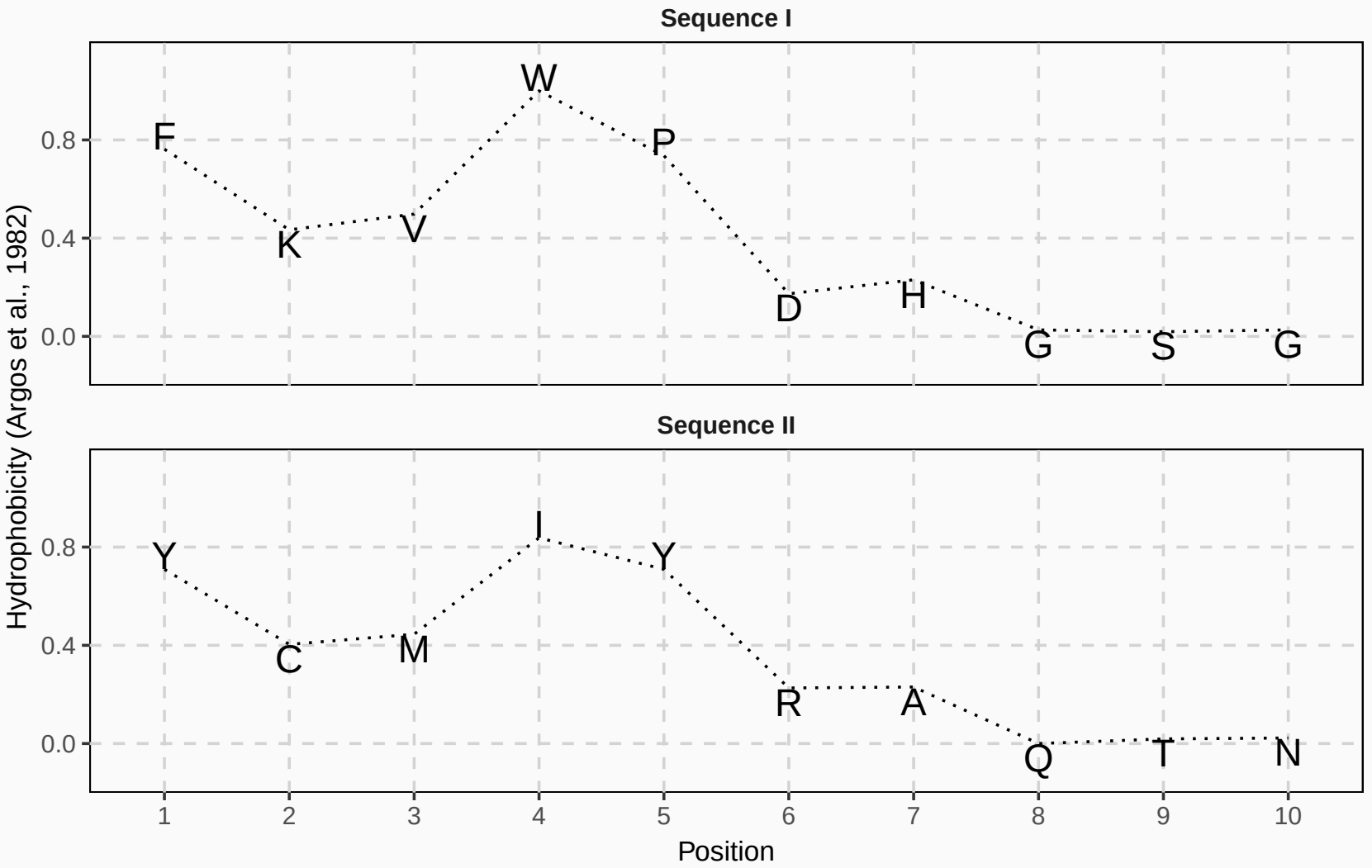
Simplified alphabets

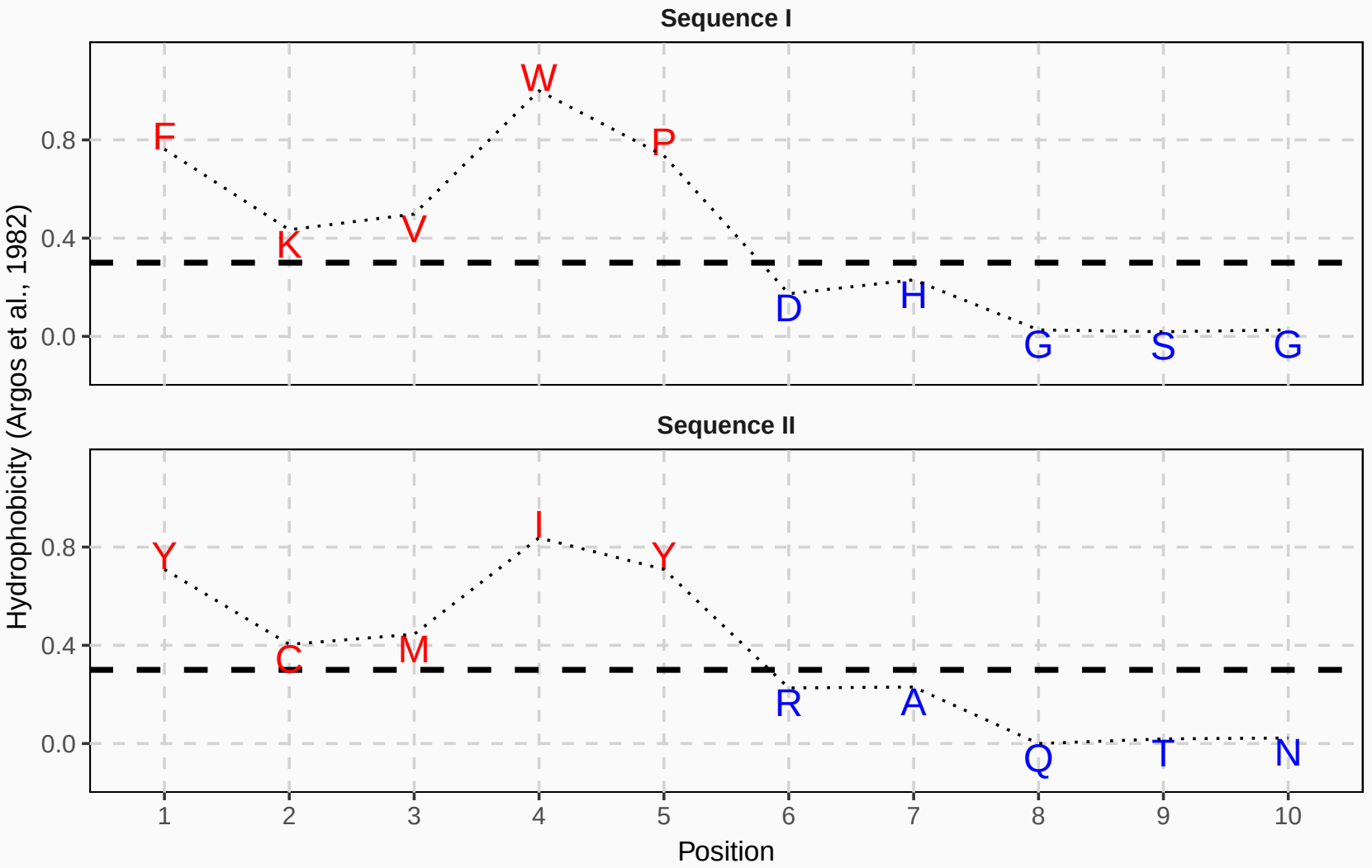
Simplified alphabets:

- are based on grouping amino acids with similar physicochemical properties,
- ease computational analysis of a sequence (Murphy et al., 2000),
- create more explicit models.

Two sequences that are drastically different considering their amino acids composition can have the same physicochemical properties.

Sequence I:	`FKVWPDHGSG`
Sequence II:	`YCMIYRAQTN`


Sequence I
Sequence II
Hydrophobicity (Argos et al., 1982)
0.0
0.4
0.8
F
K
V
W
P
D
H
G
S
G
Y
C
M
I
Y
R
A
Q
T
N
1
2
3
4
5
6
7
8
9
10
Position


Sequence I
Sequence II
Hydrophobicity (Argos et al., 1982)
0.0
0.4
0.8
F
K
V
W
P
D
H
G
S
G
Y
C
M
I
Y
R
A
Q
T
N
1
2
3
4
5
6
7
8
9
10
Position

Subgroup	Amino acid
1	C, I, L, K, M, F, P, W, Y, V
2	A, D, E, G, H, N, Q, R, S, T

Sequence I: `FKVWPDHGSG`

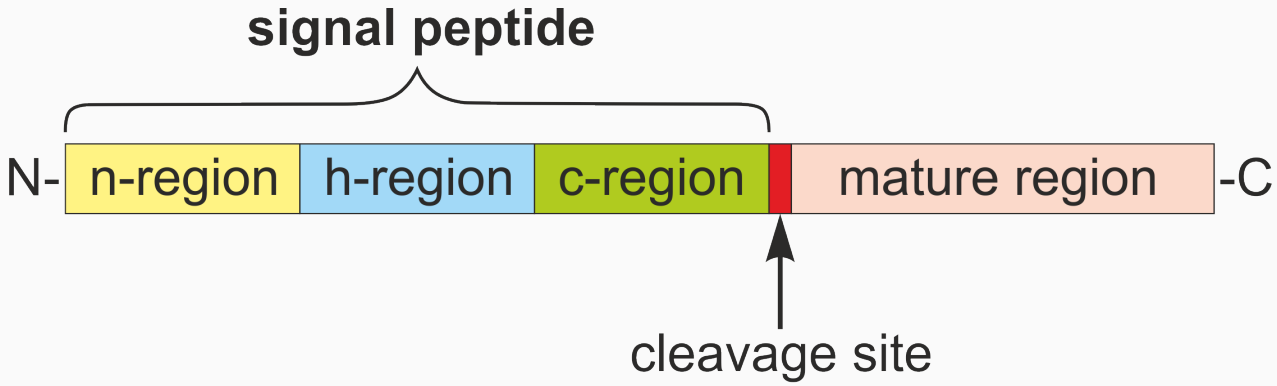
Sequence II: `YCMIYRAQTN`

Subgroup	Amino acid
1	C, I, L, K, M, F, P, W, Y, V
2	A, D, E, G, H, N, Q, R, S, T

Sequence I: FKVWPDHGSG

Sequence II: YCMIYRAQTN

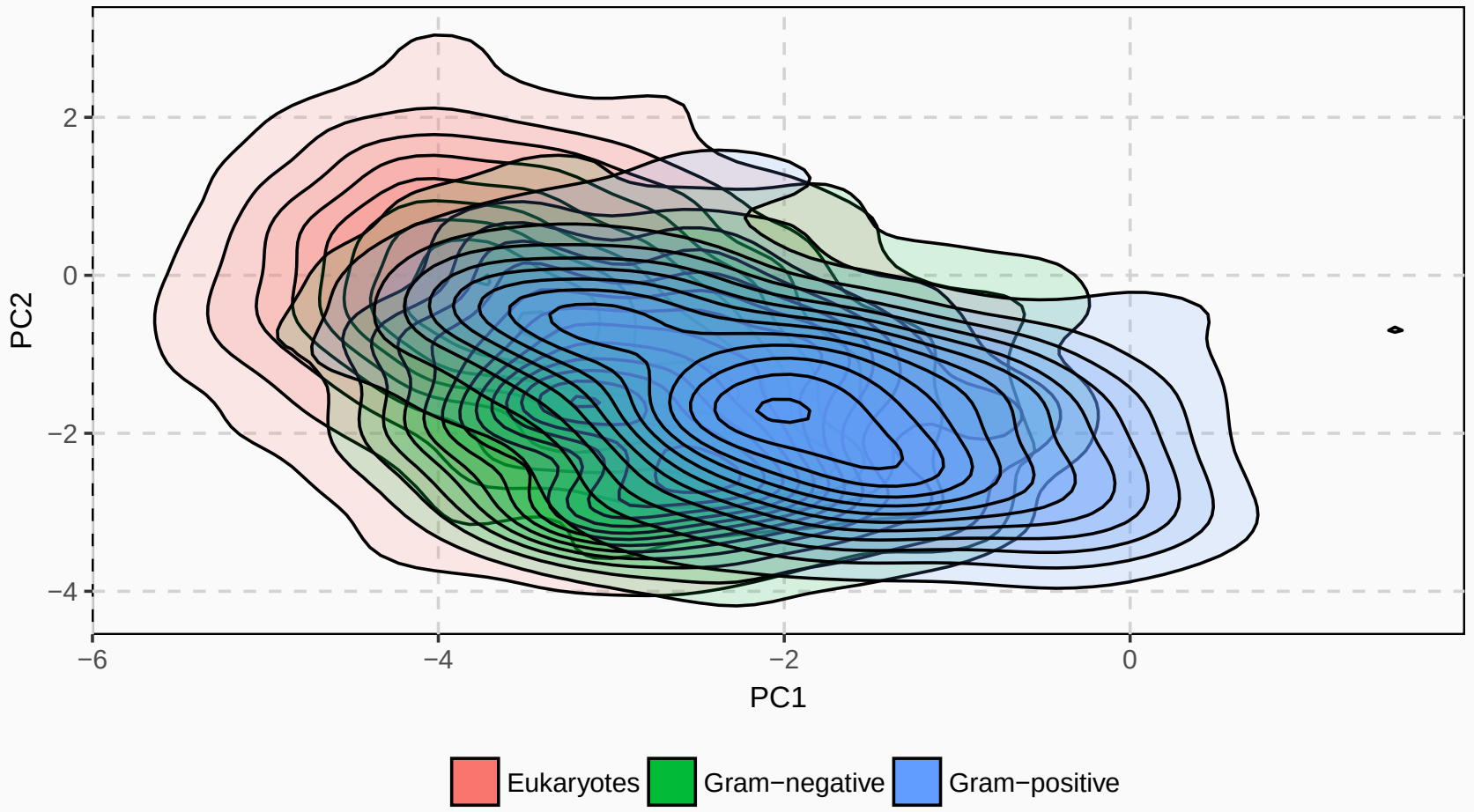
Signal peptides



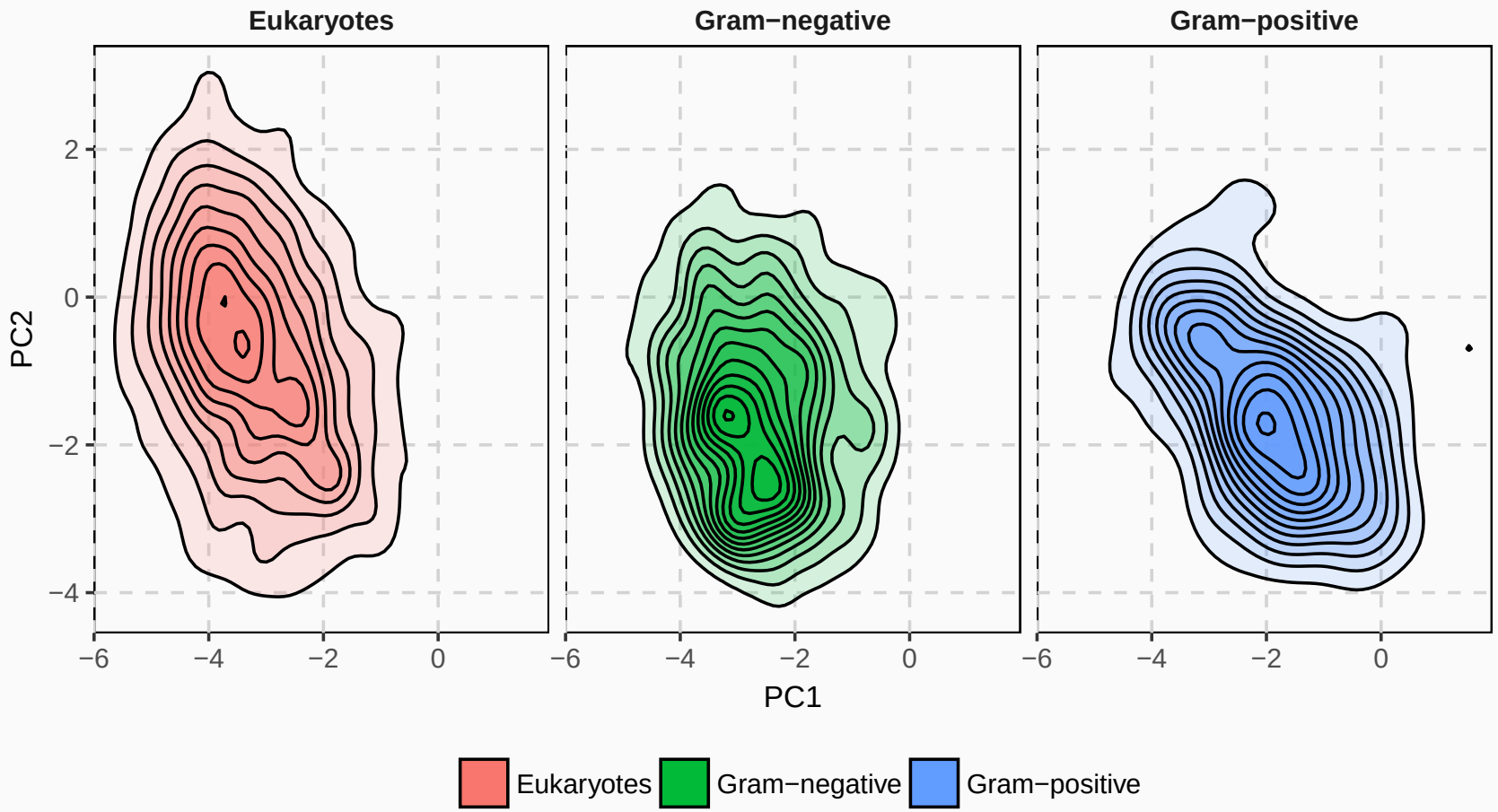
Signal peptides possess three distinct domains with variable length and characteristic amino acid composition (Hegde and Bernstein, 2006):

- n-region: mostly basic residues (Nielsen and Krogh, 1998),
- h-region: strongly hydrophobic residues (Nielsen and Krogh, 1998),
- c-region: a few polar, uncharged residues, similar to the mature protein.

Amino acid composition of signal peptides differ between eukaryotes and bacteria (gram positive and gram negative). Therefore, predictors of signal peptides are not universal and had to be taxon-specific.



PCA of amino acid frequency in signal peptides.



PCA of amino acid frequency in signal peptides.

Aim

Obtain a simplified alphabet to create a unified signal peptide predictor.

Methodology

The are several algorithm for the effective probing of the alphabet space. They were created mostly with the protein folding/alignment in mind, but can be altered to work in prediction models.

Algorithms:

- branch and bound (Cannata, 2002),
- genetic algorithm without mutation (Palensky, 2006),
- genetic algorithm with mutation (Lenckowski and Walczak, 2007).

Genetic algorithm: candidate solutions are individuals in a evolving population, where survival is depending on the quality of the solution.

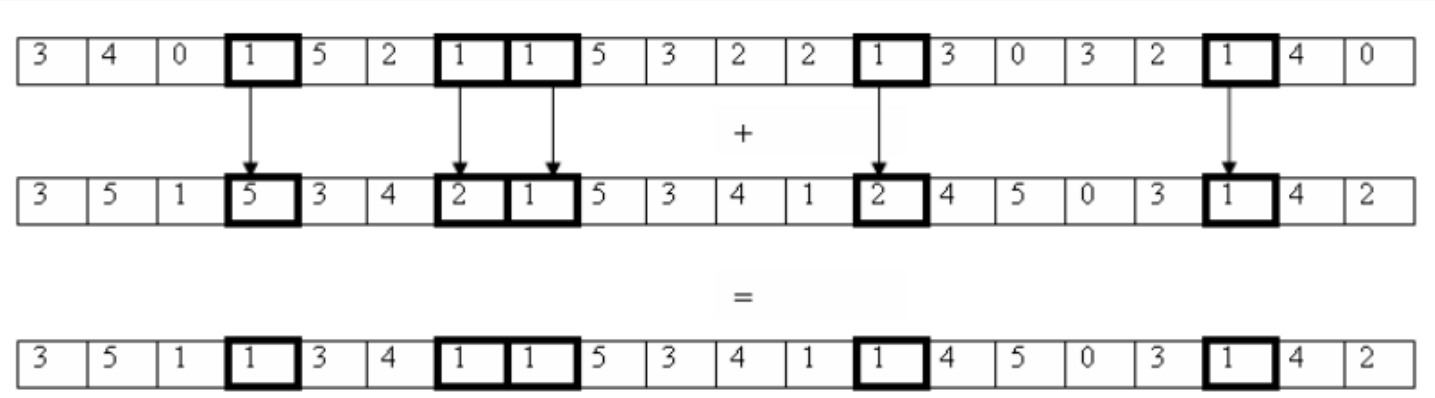
New individuals are created through cross-over and mutation of the fittest individuals.

Mutation operator

Randomly change assignments of amino acids.

Crossover operator

Assignments of amino acids from a single subgroup belonging to a chosen parent alphabet overwrite assignments of amino acids belonging to the other chosen parent alphabet.



Caveat

Keep the number of groups constant. If any group is lost, randomly revert chosen amino acids to the missing groups.

Fitness function

Fitness function: a mean value of multiple χ^2 statistics.

We compare six data sets: 1-grams from signal peptides and mature proteins belonging to gram negative bacteria, gram positive bacteria and eukaryotes. The fitness function for a single comparison f_{single} is the value of chi^2 statistic for following data (assuming the simplified alphabet of length 6):

	I	II	III	IV	V	VI
Taxon A	$x_{A,I}$	$x_{A,II}$	$x_{A,III}$	$x_{A,IV}$	$x_{A,V}$	$x_{A,VI}$
Taxon B	$x_{B,I}$	$x_{B,II}$	$x_{B,III}$	$x_{B,IV}$	$x_{B,V}$	$x_{B,VI}$

$x_{TAXON,N-GRAM}$ denotes the count of a specific 1-gram (a Latin number) in a specific taxon (A or B).

Fitness function

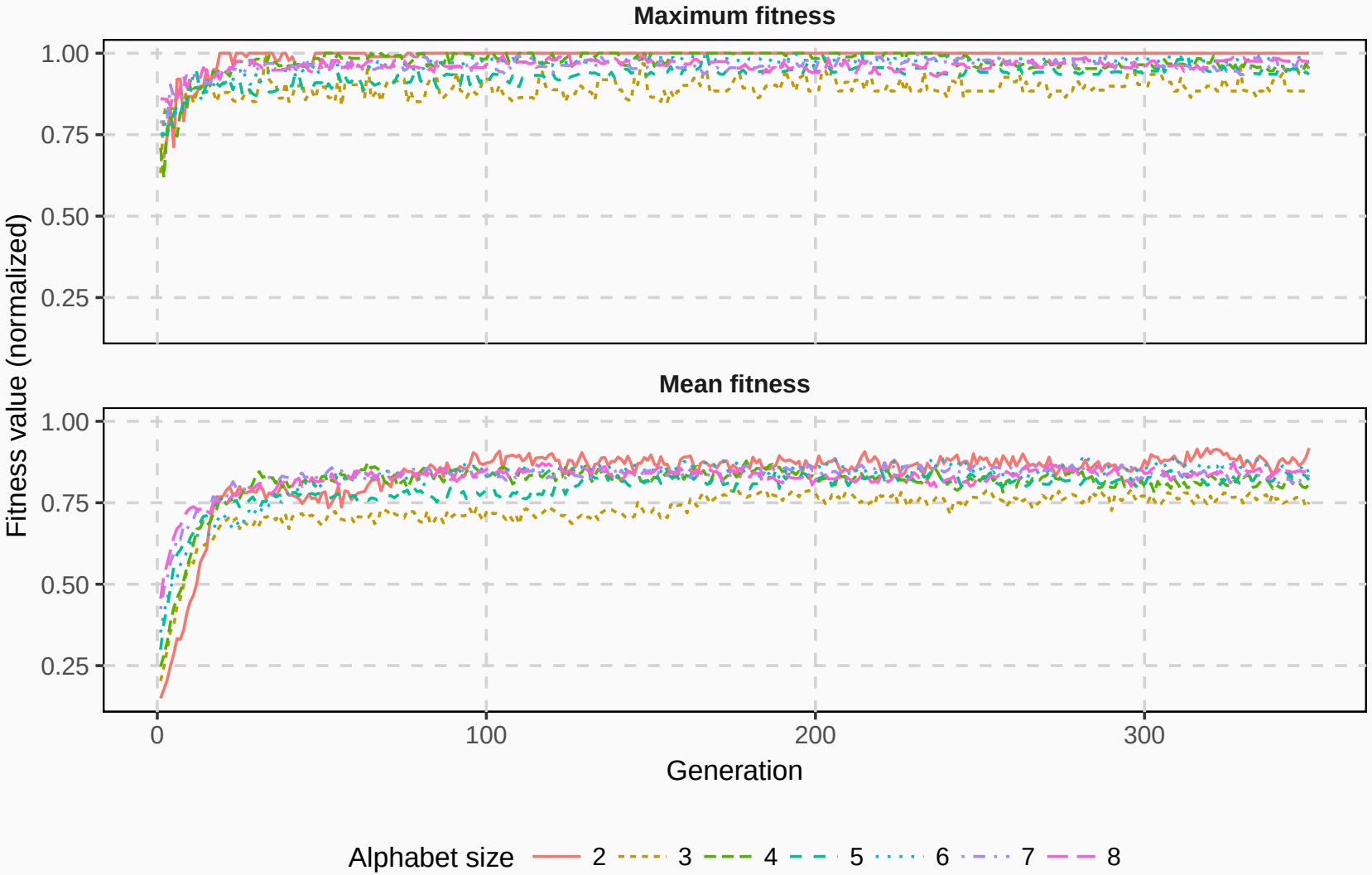
f_{final} is an **arithmetic mean** of multiple χ^2 statistics for all possible taxon comparisons:

Signal peptide origin	Mature peptide origin
gram negative bacteria	gram negative bacteria
gram positive bacteria	gram negative bacteria
eukaryotes	gram negative bacteria
gram negative bacteria	gram positive bacteria
gram positive bacteria	gram positive bacteria
eukaryotes	gram positive bacteria
gram negative bacteria	eukaryotes
gram positive bacteria	eukaryotes
eukaryotes	eukaryotes

Results

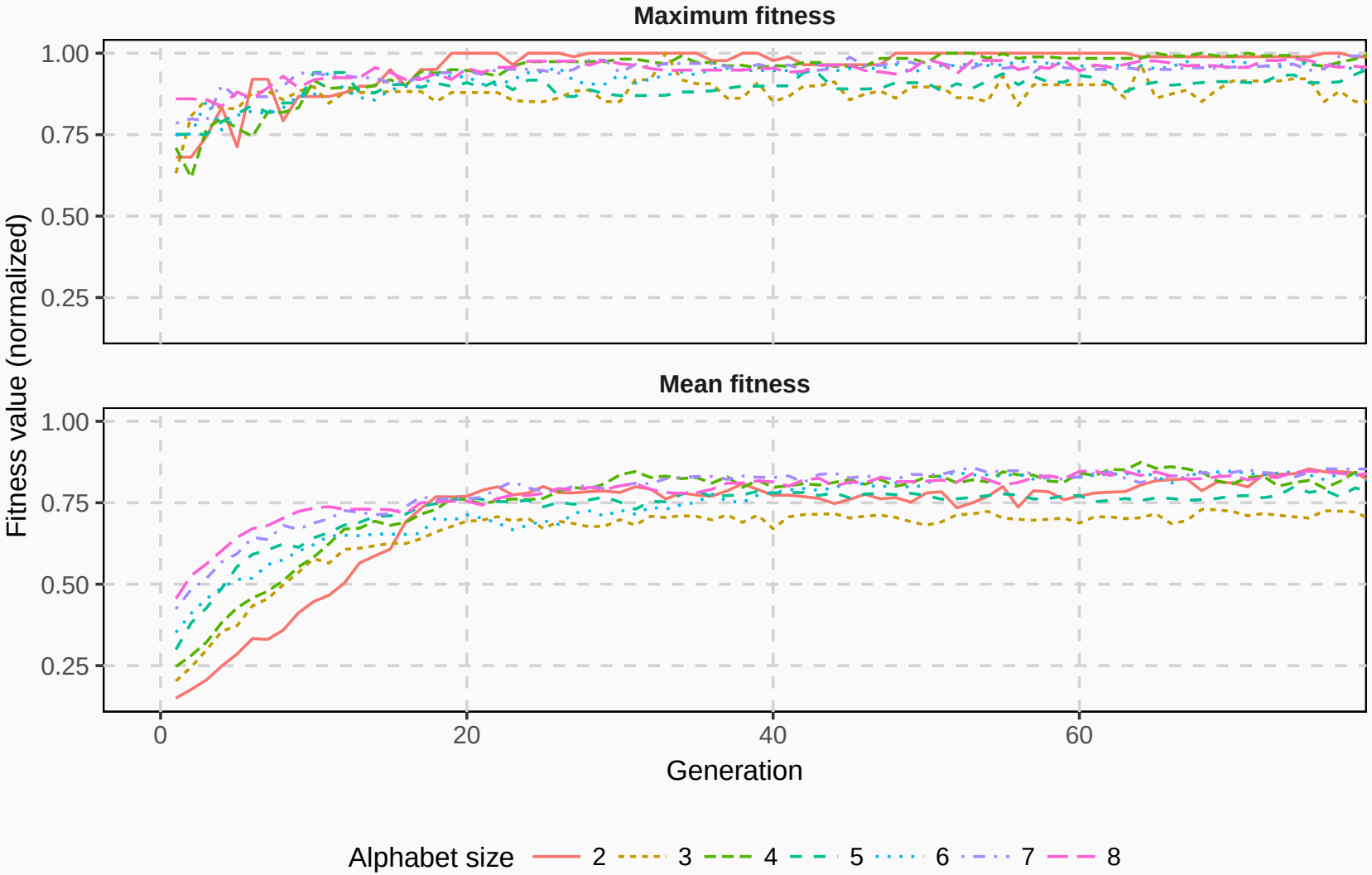
Fitness function

Fitness plot



Fitness function (close-up)

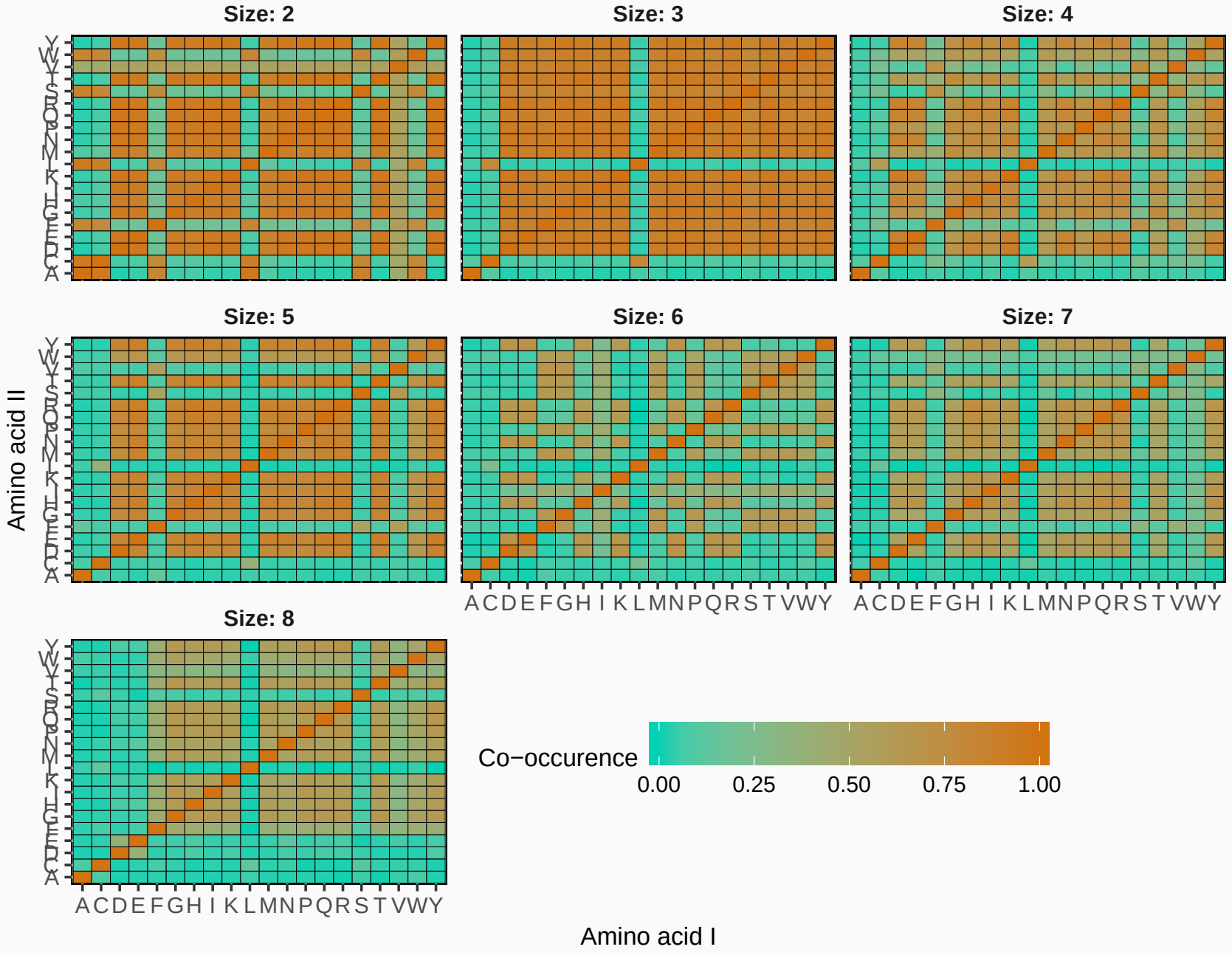
Fitness plot



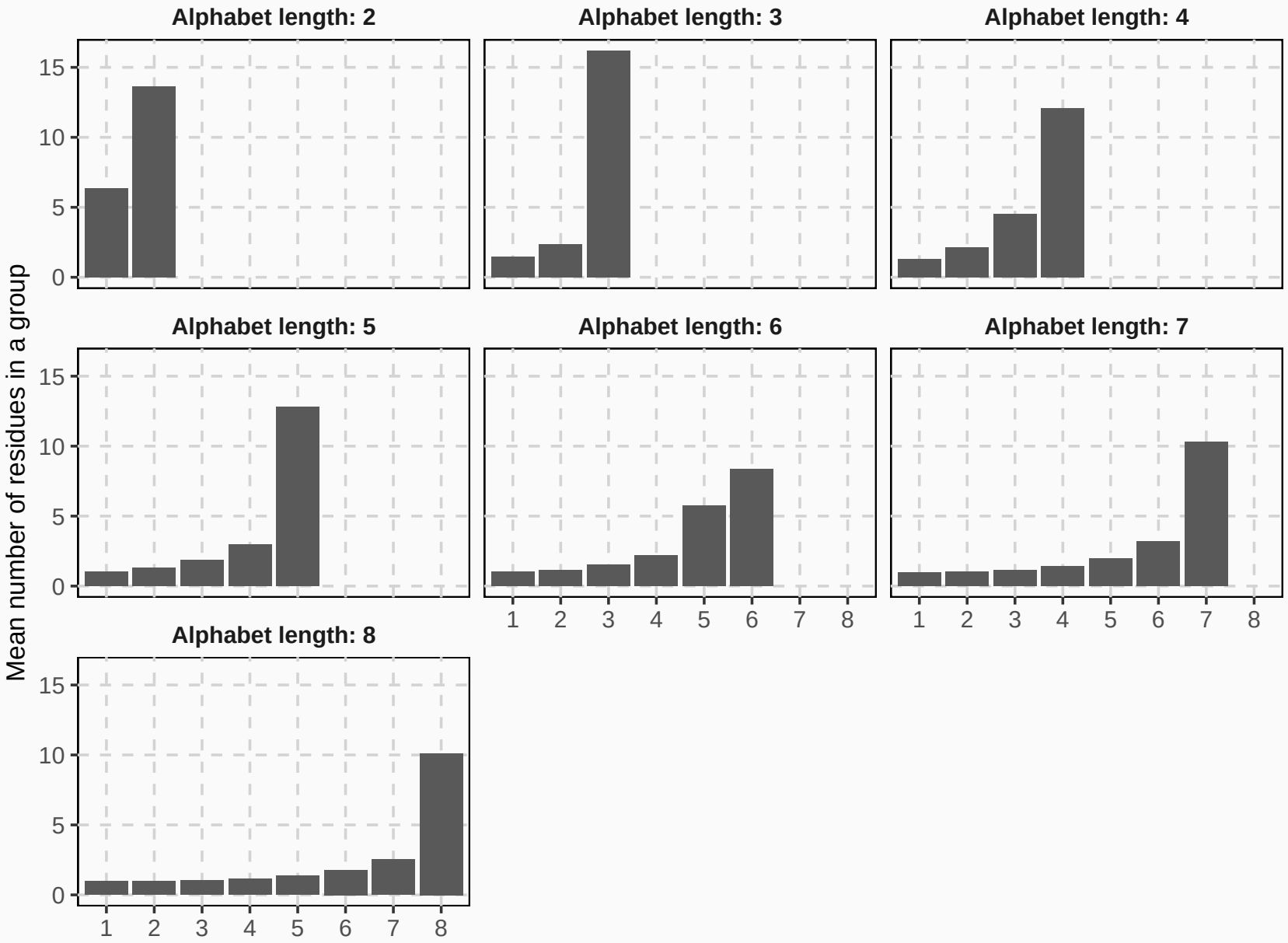
Best simplified alphabets

To be absolutely sure that we have only alphabets in the fitness plateau, I considered only alphabets from generations 200-350.

To find the most common groupings of amino acids, I computed the co-occurrence of amino acids in groups. The co-occurrence is defined as the fraction of best-fitness alphabets, where the amino acid I is in the same group as the amino acid II.


Size: 2
Size: 3
Size: 4
Size: 5
Size: 6
Size: 7
Size: 8
Amino acid II
Amino acid I
A C D E F G H I K L M N P Q R S T V W Y
Co−occurence
0.00
0.25
0.50
0.75
1.00

Preference for small groups



Preference for small groups

The prevalence of groups of a length 1 may stem from the algorithm itself. Since the Walczak's algorithm keeps the alphabet length constant, when due to the crossover or mutation, a single group is missing, **randomly chosen amino acid** is altered to be in this group.

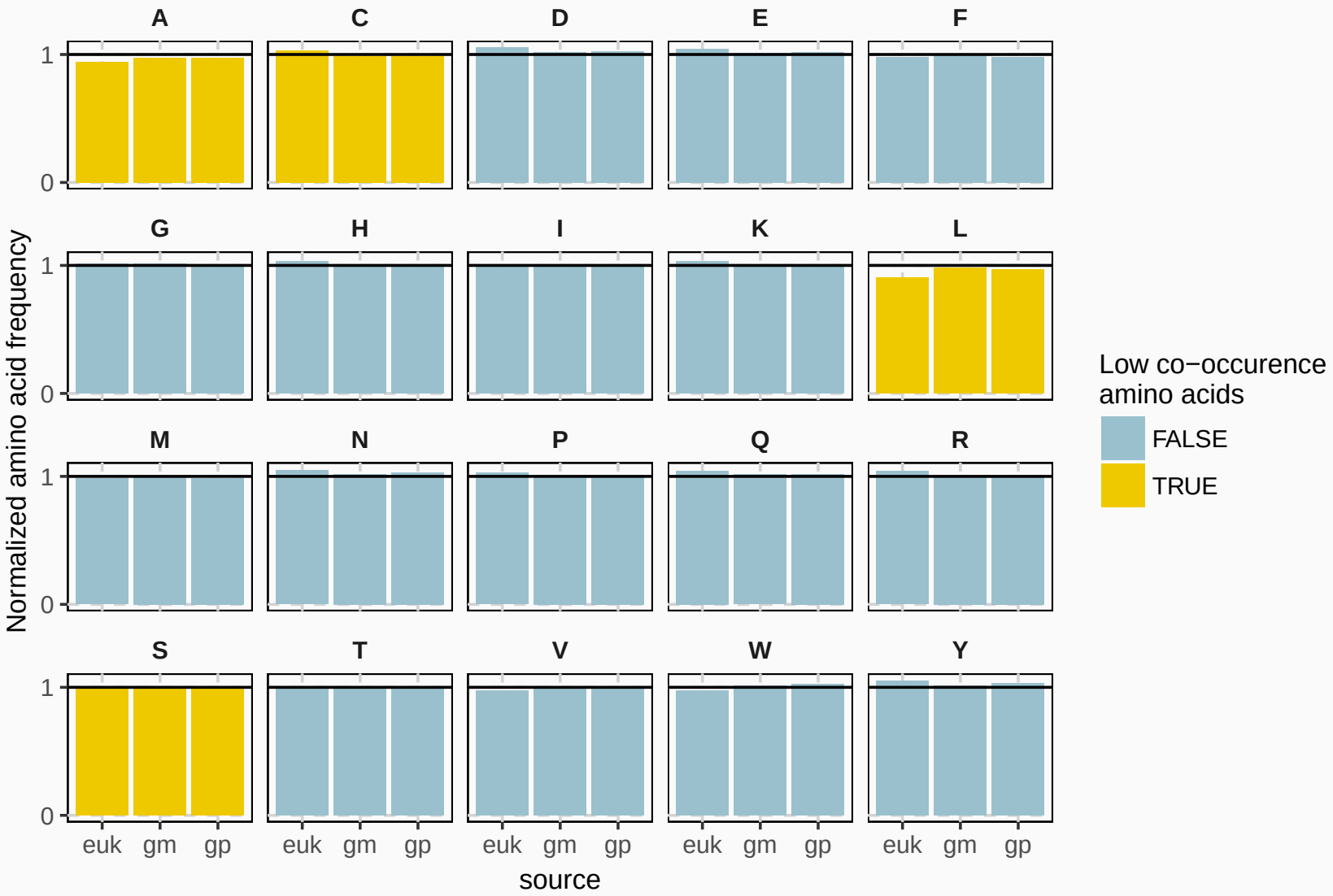
Frequency standardization

Normalization by the global amino acid frequency removes the taxon-specific amino acid bias.

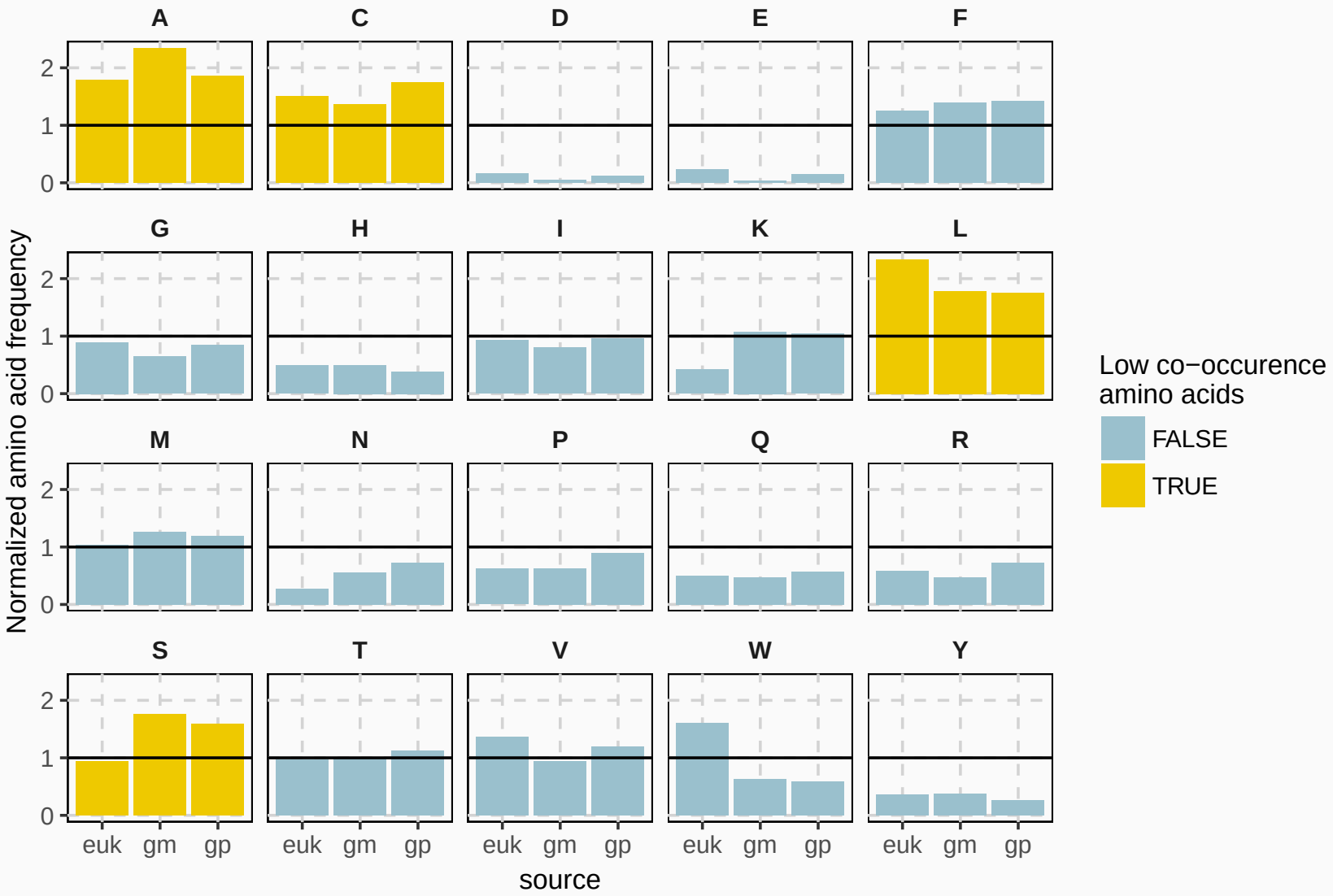
The global amino acid frequency is very similar to the amino acid frequency in mature proteins, because they are the majority in our data set.

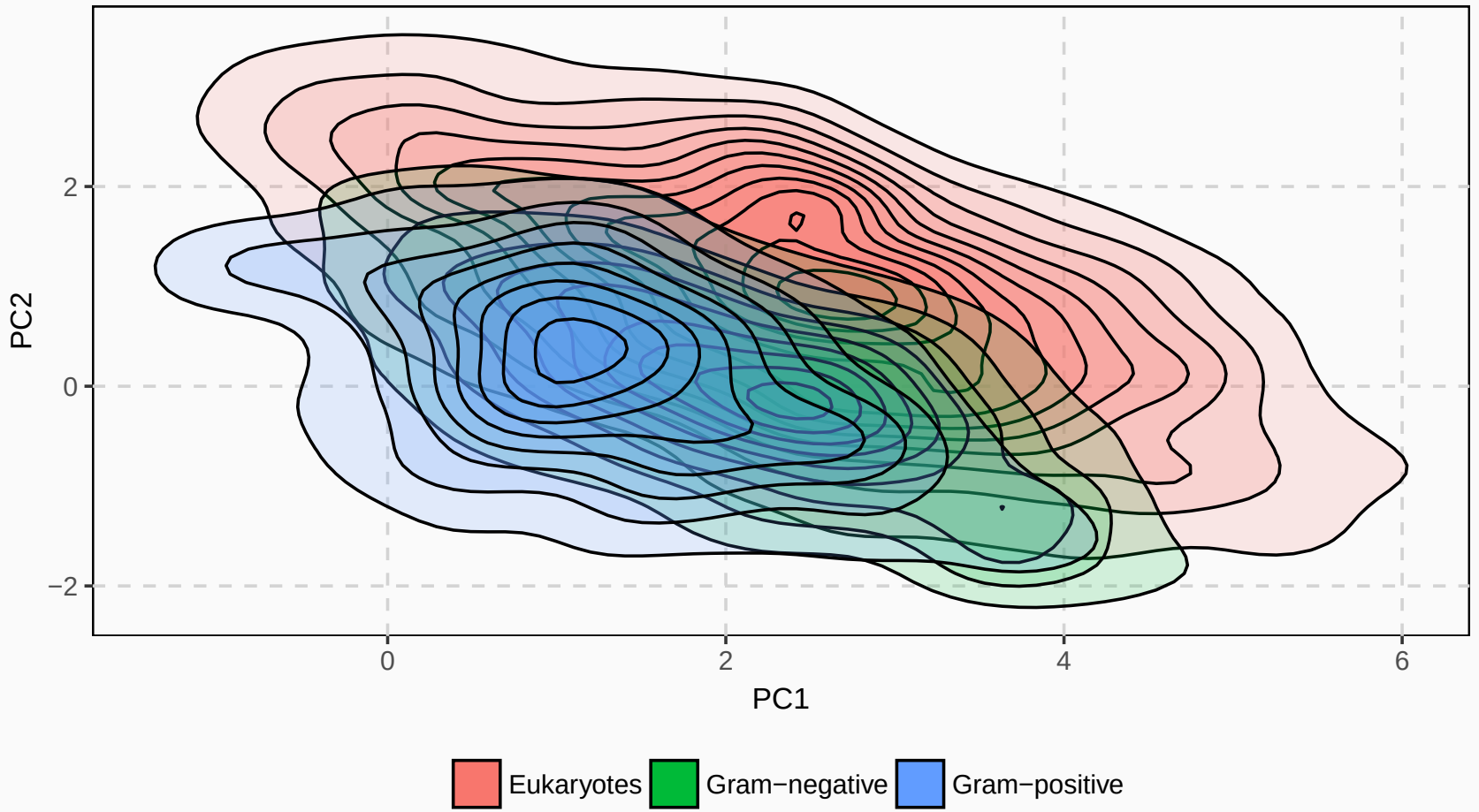
Only fours residues had constantly the lowest co-occurrence: A, C, L and S.

Mature protein

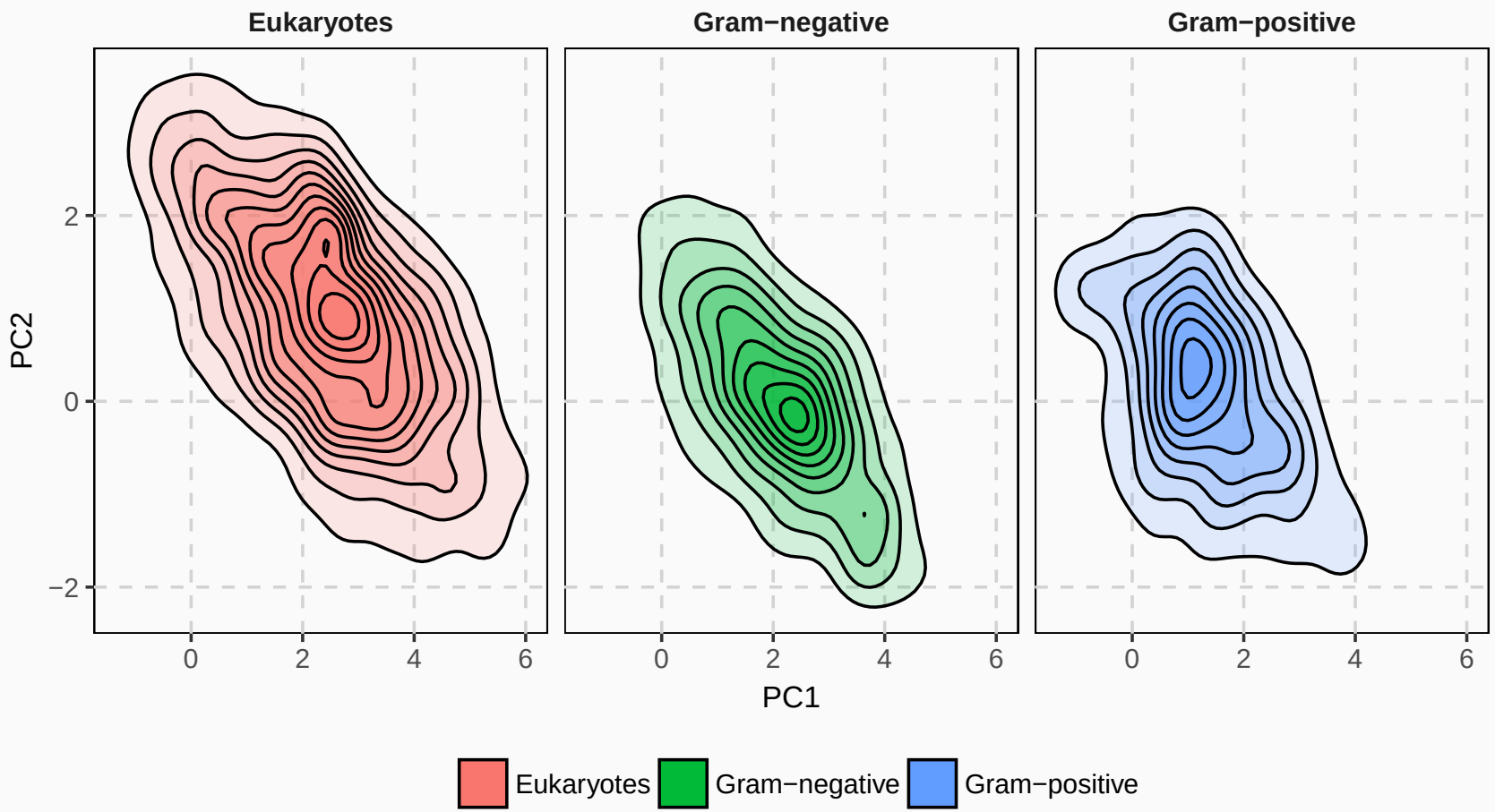


Signal peptide





PCA of amino acid frequency in signal peptides (simplified alphabet).



PCA of amino acid frequency in signal peptides (simplified alphabet).

Summary

- Mean χ^2 is not the appropriate fitness measure.
- Repeats of specific amino acids (especially leucine) are typical for signal peptides and have functions associated with protein localisation (Labaj et al., 2010; Mier et al., 2017).

Acknowledgments

National Science Center.

References

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

Labaj, P. P., Leparc, G. G., Bardet, A. F., Kreil, G., and Kreil, D. P. (2010). Single amino acid repeats in signal peptides. *The FEBS journal*, 277(15):3147–3157.

References II

Lenckowski, J. and Walczak, K. (2007). Simplifying Amino Acid Alphabets Using a Genetic Algorithm and Sequence Alignment. In *Evolutionary Computation,Machine Learning and Data Mining in Bioinformatics*, pages 122–131. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-540-71783-6_12.

Mier, P., Alanis-Lobato, G., and Andrade-Navarro, M. A. (2017). Context characterization of amino acid homorepeats using evolution, position, and order. *Proteins: Structure, Function, and Bioinformatics*, 85(4):709–719.

Murphy, L. R., Wallqvist, A., and Levy, R. M. (2000). Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering*, 13(3):149–152.

References III

Nielsen, H. and Krogh, A. (1998). Prediction of signal peptides and signal anchors by a hidden Markov model. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 6:122–130.