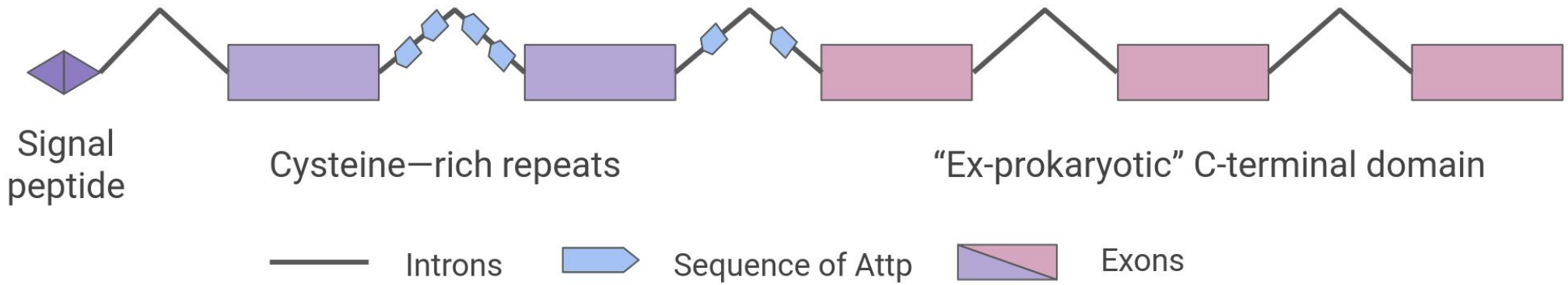
Potential bacteriophage recombination sites inside genes containing cysteine repeats

Sorokina N. [1] , Daugavet M. [2]
[1] Applied Genomics Laboratory, SCAMT Institute, ITMO University, Saint Petersburg, Russia
[2] Laboratory of Noncoding DNA, Institute of Cytology, Russian Academy of Sciences, Saint Petersburg, Russia

Background

Rusticalin-like gene structure in the ascidian *Ciona intestinalis*



1. The **rusticalin protein** has been identified.
2. The **N-terminal domain** contained **2 cysteine-rich repeats**.
3. The **C-terminal domain** had similarities with both **bacterial MD peptidases** and **bacteriophage A500 L-alanyl-D-glutamate peptidase**.
4. A **bacteriophage recombination site** - **Attp** was detected **adjacent to the cysteine repeats [1,2]**.

Hypothesis

Cysteine repeats associated with Attp sequences may be **markers** of putative horizontal transfer of **bacterial genes by bacteriophages to eukaryotes**.

[1] https://doi.org/10.1186/s12859-020-03599-y
[2] https://doi.org/10.1186/s13100-019-0146-7

Research Objective

Evaluate association of AttP sequences positions and cysteine repeats positions

Materials and Methods

1. **Identify cysteine repeats pattern**

$CX_{4-9}CX_{4-9}CX_{4-9}CX_{4-9}CC$,
where X - any amino acid except cysteine

2. **Create datasets of proteins and Attp**

- **Experimental dataset of proteins**

All proteins matching the pattern in the Uniprot Trembl database were selected

Genes, encoding proteins containing cysteine repeats
8944 sequences

- **Control dataset of proteins**

Proteins without cysteine repeats

Genes, encoding proteins without cysteine repeats
122 sequences

- **Dataset of Attp sequences**

Received earlier from NCBI

Attp of viruses/phages (NCBI)
186 sequences
Length from **10** till **491** nucleotides

3. **Find Attp sequences inside genes**

4. **Analyze obtained results**

Results and Conclusion

By searching for **Attp in the genes** of proteins **with and without cysteine repeats** without **mismatches and indels**

	Attp found	**Attp not found**	**Total**
With cysteine repeats	2947	5997	**8944**
Without cysteine repeats	17	105	**122**
Total	**2964**	**6102**	**9066**

2x2 Fisher's exact test conclude that the **two-tailed P value is less than 0.0001**
The **association between attp and cysteine repeats** is considered to be **statistically significant**.

1. **Potential recombination sites of bacteriophages** within genes containing **cysteine repeats were found**.
2. Compared with the control group, it can be concluded that **attp was more frequently found** in the dataset **containing cysteine repeats**.
3. Further, it is necessary to **search for bacterial domains in eukaryotic genes** containing recombination sites.
4. It is reasonable to continue investigating the **hypothesis of horizontal transfer associated with bacteriophages**.

Acknowledgements

This work was supported by the Russian Science Foundation (grant number 19–74-20102) and Ministry of Science and Higher Education of the Russian Federation (Agreement № 075-15-2021-1075, signed 28.09.2021

Thank you for attention !