Predicting elongation efficiency of gene translation for annotation of bacterial genomes: a case study for biosynthetic gene clusters of nonribosomal peptides

A.I. Klimenko, Yu.G. Matushkin, D.A. Afonnikov
Kurchatov Genomics Center,
Institute of Cytology and Genetics, ICG SB RAS
Novosibirsk, Russia
klimenko@bionet.nsc.ru

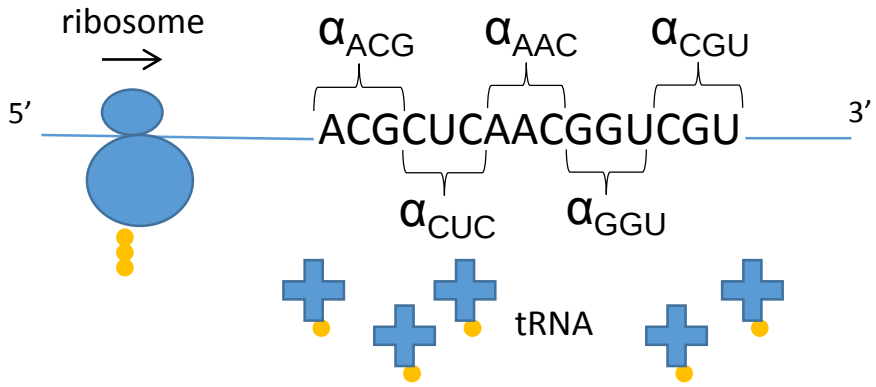
Predicting elongation efficiency of gene translation in bacteria

- The EloE software is a tool for gene ranking based on their putative translation elongation efficiency inferred from their nucleotide sequences taking into account such factors as codon composition, presence and stability of secondary structures in mRNA [Likhoshvai and Y. G. Matushkin, 2000].
- The obtained predicted values correlate with available experimental data on gene expression in different microorganisms [Sokolov et al., 2015].
- Thus, EloE is useful as a bioinformatic tool for genome annotation that enables a researcher with a capacity to infer *a priori* estimates of gene expression efficiency based on whole-genome nucleotide sequences only.

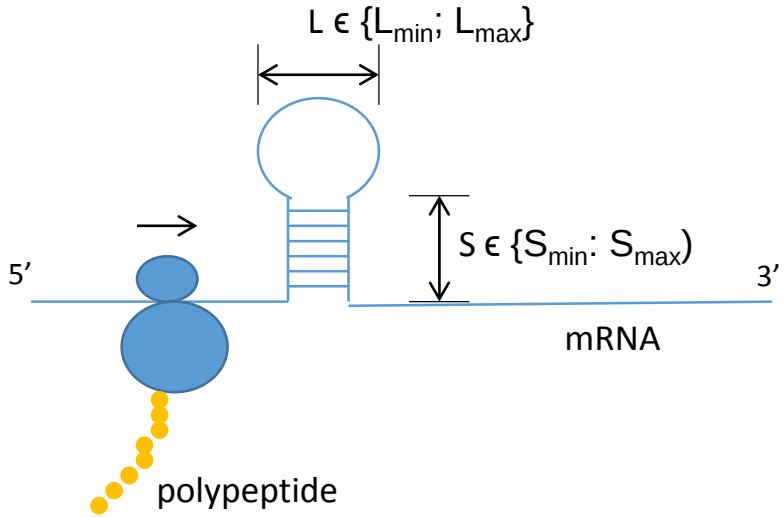
Index type	Codon usage	Secondary structure only	Secondary structure and energy
EEI1	+	-	-
EEI2	-	+	-
EEI3	-	-	+
EEI4	+	+	-
EEI5	+	-	+

EEI (Elongation Efficiency Index) consists of two parts:

1. Codon composition of mRNA



2. Number of local inverted repeats and 3. stability of corresponding secondary structures



Nonribosomal peptides

- Nonribosomal peptides (NRPs) constitute an important fraction of bacterial peptidomes acting as antibiotics, toxins, surfactants, siderophores, anti-tumor agents and immune response modifiers [Caboche et al., 2008].
- Biosynthesis of NRPs is dependent on particular enzymes - nonribosomal peptide synthetases (NRPSs), which are encoded by biosynthetic gene clusters (BGCs) in bacterial genomes [Süssmuth and Mainz, 2017].

We processed 12436 bacterial genomes (1052 genera and 34 phyla) presented in JGI GOLD with the genome project's status 'Complete and Published'.
The information on NRP biosynthetic gene clusters (BGCs) was obtained from ANTISMASH-DB: we analyzed 2249 bacterial containing 5676 NRPS clusters

JGI GOLD GENOMES ONLINE DATABASE

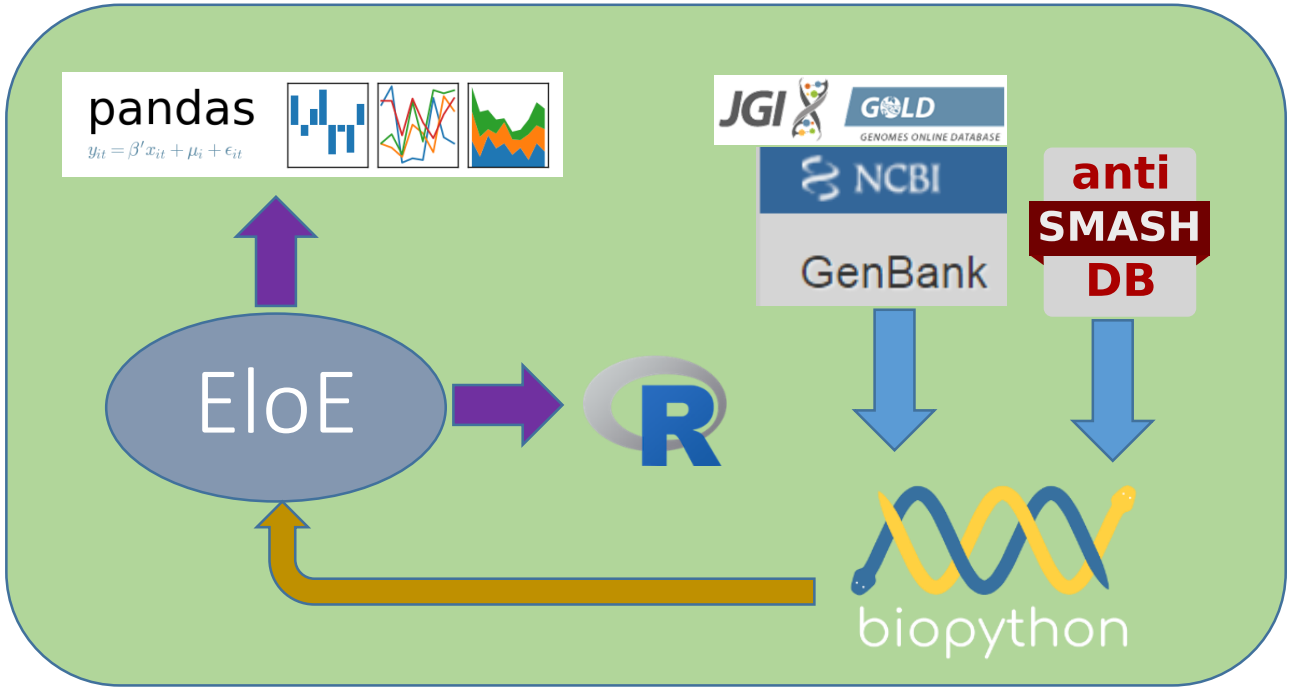
[Mukherjee et al., 2018]

anti SMASH DB

[Blin et al., 2017]

Pipeline of data analysis

- We have performed bioinformatic analysis of NRP biosynthetic gene clusters (BGCs) obtained from ANTISMASH-DB using whole-genome sequences of bacterial genomes that are available at NCBI Genbank.
- The analysis is based on the method predicting gene translation elongation efficiency that is implemented in the EloE software.
- Statistical and bioinformatic analysis scripts have been developed on Python using software library Biopython.

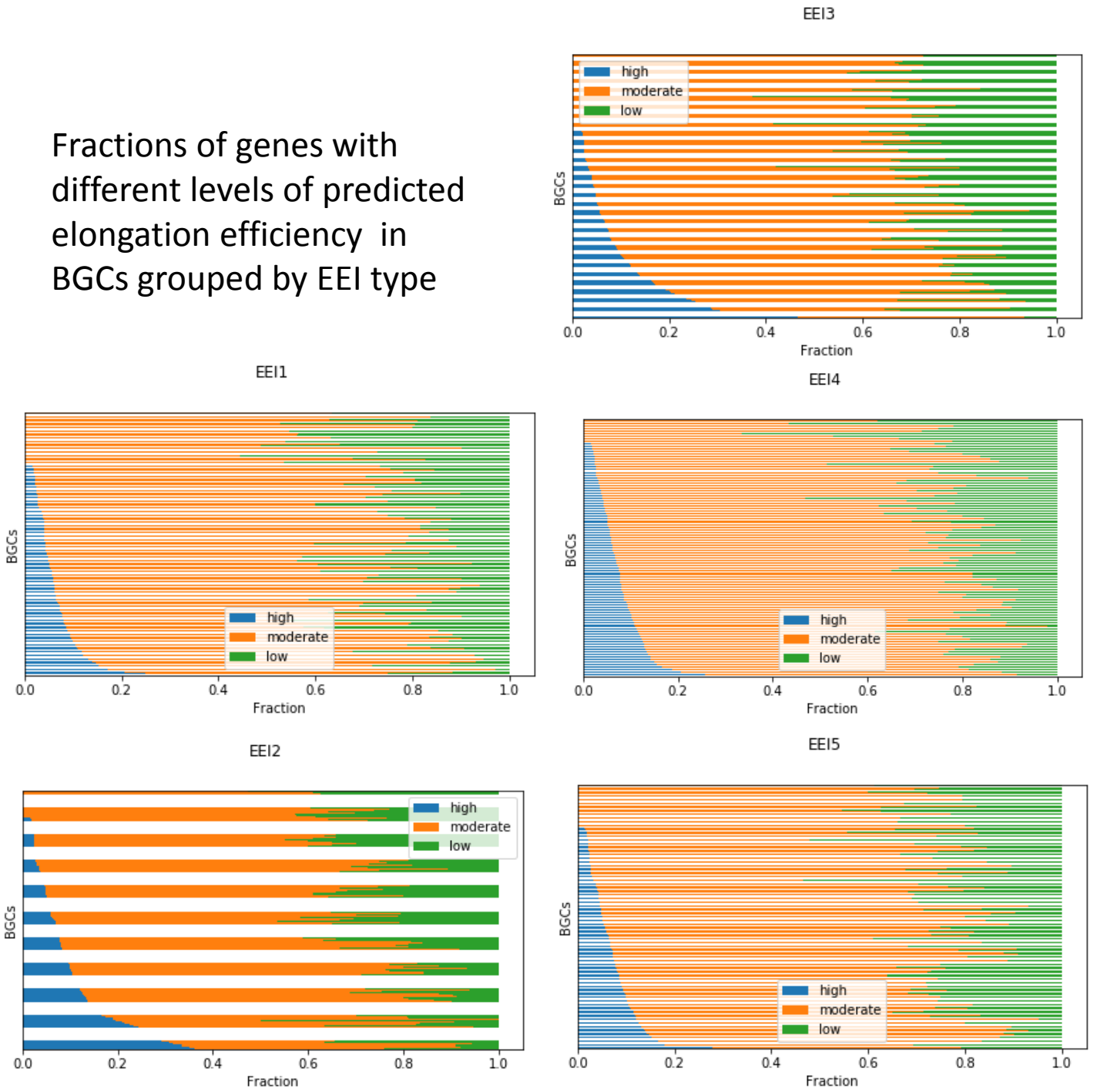


The principle scheme of the bioinformatic pipeline. The data on NRPS clusters as well as genome sequence data were downloaded via Biopython and served as an input for predicting elongation efficiency of gene translation in EloE. The results of the prediction are conveyed then to further statistical analysis using Python (Pandas library) and R (robCompositions library).

Results

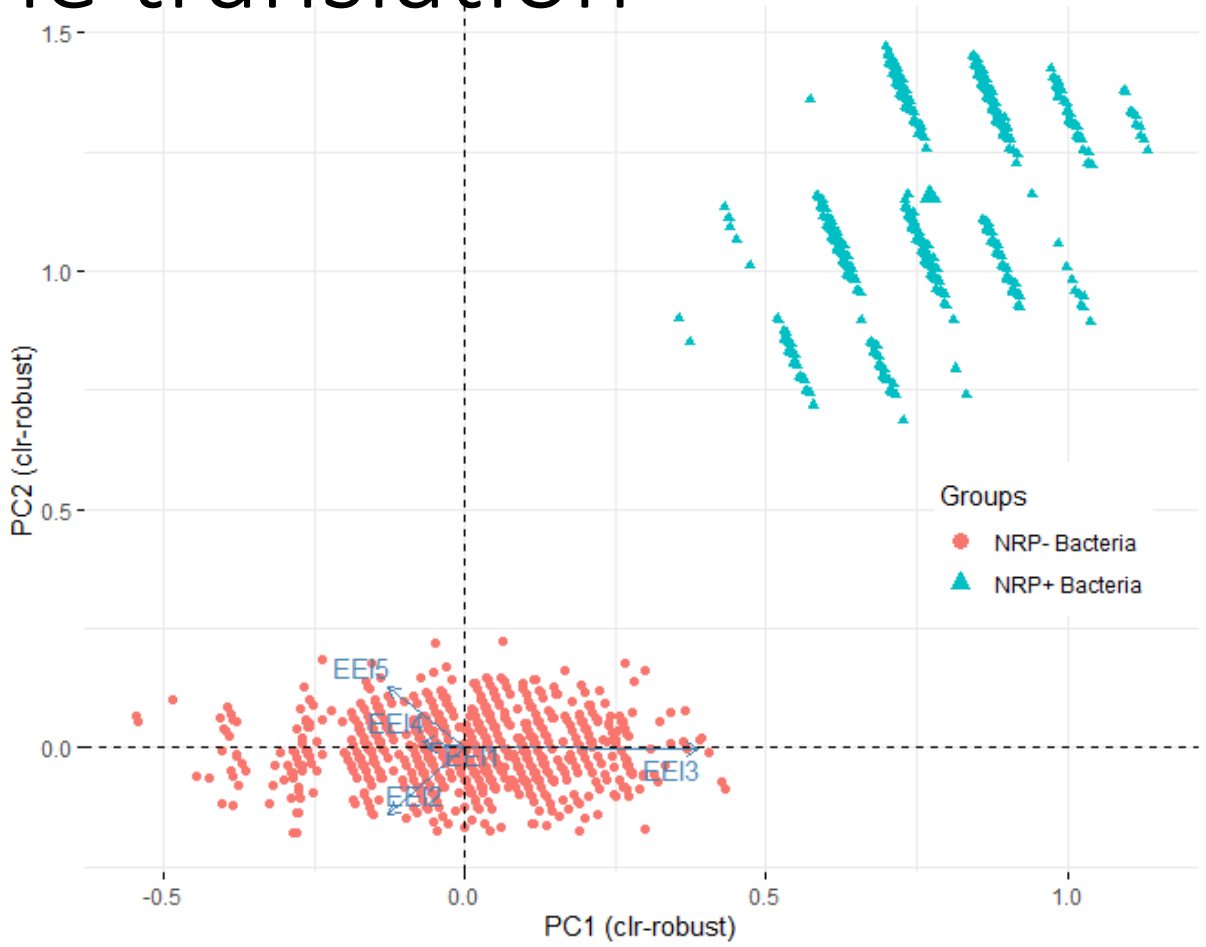
- The statistical analysis provided the information about distribution of nonribosomal peptide biosynthetic gene clusters in bacteria and their putative translation elongation efficiency.
- The taxa possessing genomes enriched by BGCs were revealed and it was shown that while only 6.6% NRPS genes fall into the category of high predicted efficiency of translation, there is a number of BGCs distinguished by both large absolute number of genes with high predicted efficiency of translation and their percentage.
- These BGCs belong to such genera as *Pseudomonas, Staphylococcus, Corynebacterium, Streptomyces, Amycolatopsis, Paenibacillus, Rhodococcus* and *Burkholderia*.

Fractions of genes with different levels of predicted elongation efficiency in BGCs grouped by EEI type



Results: NRP+ bacteria exhibit optimized elongation efficiency of gene translation

- The results of PCA for compositional data [Filzmoser, Hron, Reimann, 2007] show that groups of organisms distinguished by the trait of presence(NRP+)/absence(NRP-) of NRP form distinct clusters. Notably, the fractions of EEI2 (where the number of potential hairpins is minimized) and EEI5 (where both codon composition and number and stability of potential hairpins is minimized) type genomes explain the major part of the variance. The same fractions distinguish NRP+ and NRP- classes of microorganisms.
- We believe that it indicates that acquiring a capacity to synthesize nonribosomal peptides incurs certain maintenance energy costs and therefore NRP+ microorganisms are pressured towards more rigorous optimization of their genome structures .



The results of PCA for compositional data performed on resampled data of EEI type compositions. Each dot represents EEI type compositions for a random sample of genomes where for each genus one representative was chosen.

Conclusion

The performed bioinformatic analysis has provided the information about distribution of nonribosomal peptide biosynthetic gene clusters in bacteria and their putative translation elongation efficiency.

Thus, the analysis of translation elongation efficiency is useful as a high-throughput technique for gene ranking in bacterial genomes, which can be regarded as a rough estimate of expression level for various groups of genes of interest.