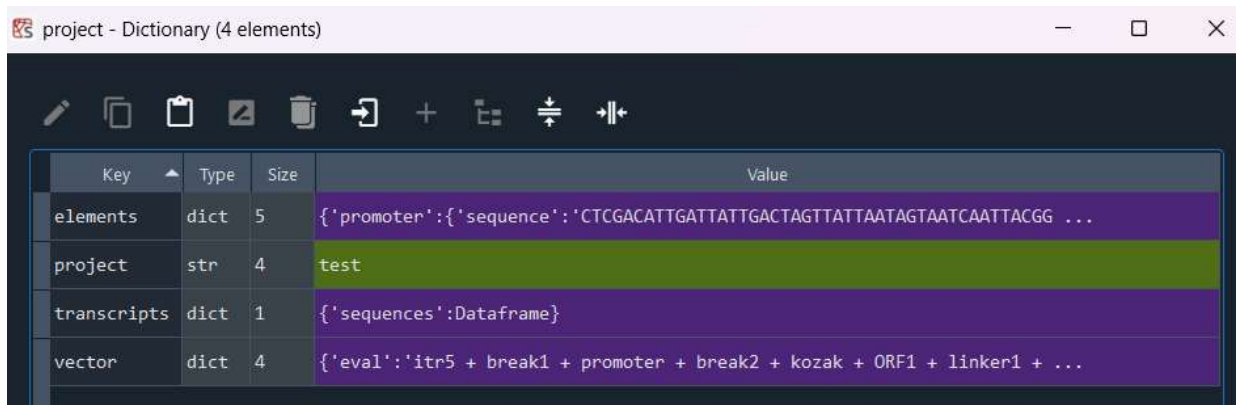


ABM – raport 16.05.2023

Raport zawiera podsumowanie informacji o strukturze danych oraz ich przetwarzaniu w tworzeniu wektora AAV z wykorzystaniem JBioSeqTools dla aplikacji webowej.

## Project – dictionary

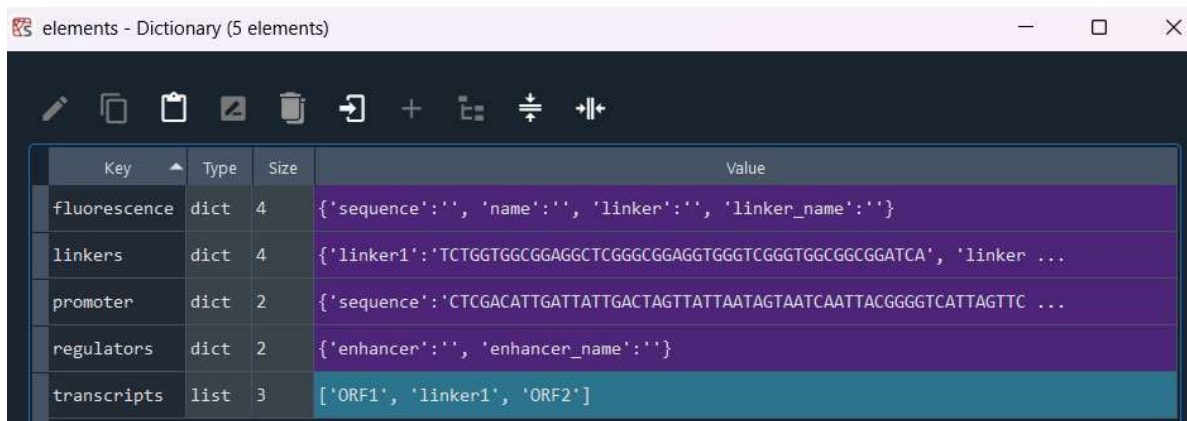


The screenshot shows a Jupyter Notebook window titled "project - Dictionary (4 elements)". It displays a dictionary with the following structure:

Key	Type	Size	Value
elements	dict	5	<code>{'promoter':{'sequence':'CTCGACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGG ...</code>
project	str	4	<code>test</code>
transcripts	dict	1	<code>{'sequences':Dataframe}</code>
vector	dict	4	<code>{'eval':'itr5 + break1 + promoter + break2 + kozak + ORF1 + linker1 + ...</code>

Elements – element wektora (chodzi o biologiczny aspekt AAV, nie wektor w rozumieniu komputerowym).

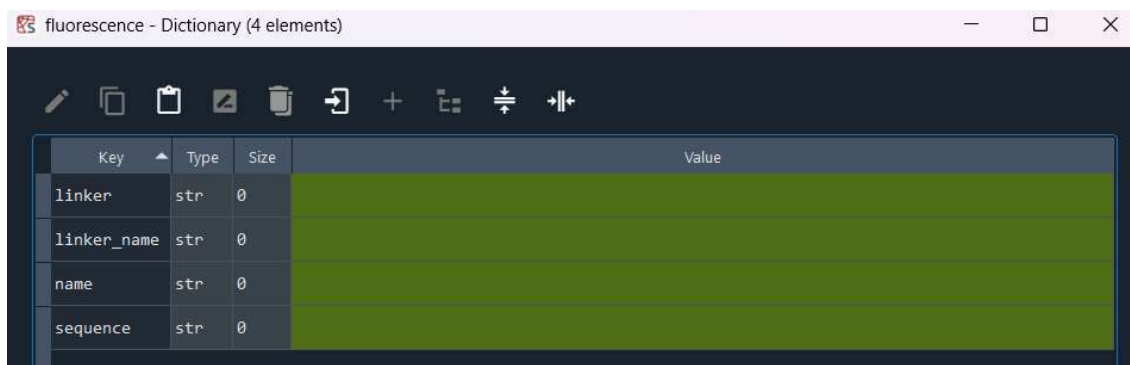
Elements – 2nd level dictionary:



The screenshot shows a Jupyter Notebook window titled "elements - Dictionary (5 elements)". It displays a dictionary with the following structure:

Key	Type	Size	Value
fluorescence	dict	4	<code>{'sequence':'', 'name':'', 'linker':'', 'linker_name':''}</code>
linkers	dict	4	<code>{'linker1':'TCTGGTGGCGGAGGCTCGGGCGGAGTGGGTCGGGTGGCGGCGGATCA', 'linker ...</code>
promoter	dict	2	<code>{'sequence':'CTCGACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTC ...</code>
regulators	dict	2	<code>{'enhancer':'', 'enhancer_name':''}</code>
transcripts	list	3	<code>['ORF1', 'linker1', 'ORF2']</code>

Fluorescence 3rd level dictionary – informacje o tagu fluorescencyjnym.



The screenshot shows a Jupyter Notebook window titled "fluorescence - Dictionary (4 elements)". It displays a dictionary with the following structure:

Key	Type	Size	Value
linker	str	0	
linker_name	str	0	
name	str	0	
sequence	str	0	

Linker – sekwencja linkera

Linker\_name – nazwa linkera

- Z linkerów znajdujących się w bazie danych

Sequence – sekwencja tagu fluorescencyjnego

Name – nazwa tagu

- Z tagów znajdujących się w bazie danych

!JEŻELI UŻYTKOWNIK NIE CHCE LINKERA W MIEJSCE WSZYSTKICH PODANYCH ZMIENNYCH TRZEBA WSTAWIĆ PUSTY STRING „”, TAK ROZPOZNAJĄ TO SKRYPTY.

Linkers 3rd level dictionary – zawiera informacje o linkerach pomiędzy transkryptami np. t1 + l1 t2, ect.

Key	Type	Size	Value
linker1	str	48	TCTGGTGGCGGAGGCTCGGGCGGAGGTGGGTGGGTGGCGGCGGATCA
linker1_name	str	7	3xGGGGS
linker2	str	0	
linker2_name	str	0	

Linker – sekwencja linkera

Linker\_name – nazwa linkera

- Z bazy danych, ilość linkerów zależy od ilości transkryptów -1 (n-1).

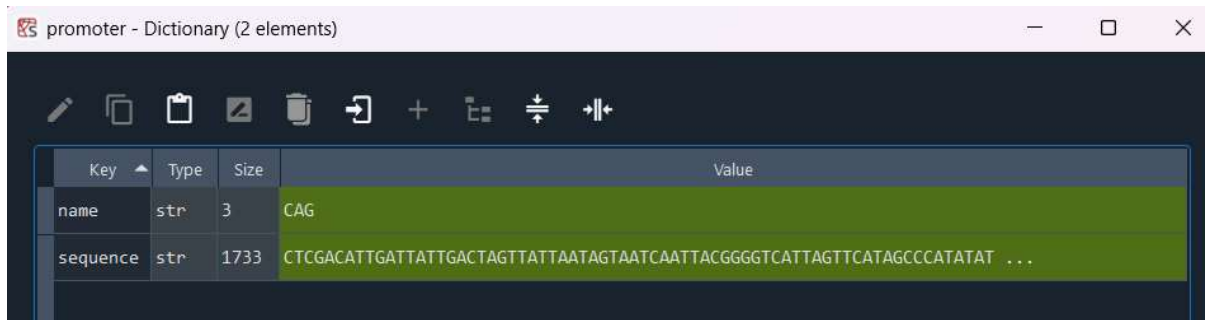
!JEŻELI UŻYTKOWNIK MA WIĘCEJ NIŻ JEDEN TRANSKRYPT I NIE CHCE LINKERA MIEDZY NIMI TO ANALOGICZNIE JAK WYŻEJ DLA KAŻDEGO LINKERA JAKI POWINIEN SIĘ POJAWIĆ ZGODNIE Z ZASADĄ n-1 DAJEMY „”.

linkers - Dictionary (2 elements)

Key	Type	Size	Value
linker1	str	0	
linker1_name	str	0	

!W PRZYPADKU GDY MAMY TYLKO JEDEN TRANSKRYPT / GEN WYBRANY TO NIE PYTAMY UŻYTKOWNIKA O LINKER BO NIE JEST POTRZEBNY I OD RAZU W LINKERS PODAJEMY linker1 oraz linker\_name1 z pustym STRINGIEM „” – tak odczytują skrypty.

## Promoter 3r level dictionary – informacje o promotorze



Key	Type	Size	Value
name	str	3	CAG
sequence	str	1733	CTCGACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATAT ...

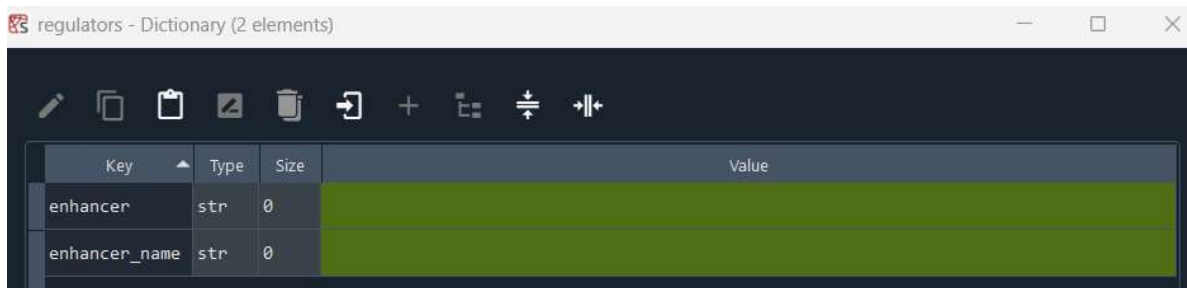
Name – nazwa promotora

Sequence – sekwencja promotora

- Zgodnie z tym co znajduje się w bazie danych (na początku znane promotory, które przesłałem, później wraz z single-cell promotory z EPD)

!UŻYTKOWNIK MUSI WYBRAĆ JAKĄŚ SEKWENCJĘ PROMOTORA LUB EWENTUALNIE WKLEIĆ CUSTOMOWĄ I NADAĆ NAZWĘ!

## Regulators 3rd level dictionary



Key	Type	Size	Value
enhancer	str	0	
enhancer_name	str	0	

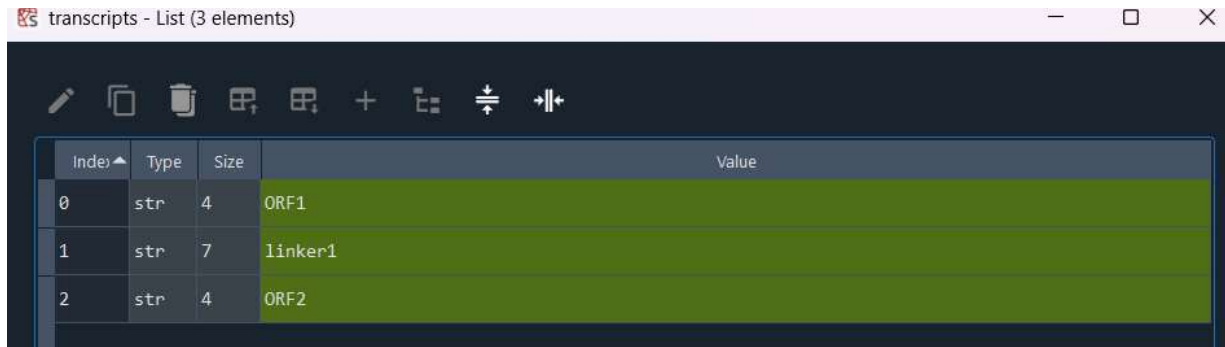
Enhancer – sekwencja regulatora

Enhancer\_name – nazwa

- Znajdujące się w bazie danych

!UŻYTKOWNIK NIE MUSI CHcieć SEKWENCJI REGULUJĄCEJ WIĘC ANALOGICZNIE JEŻELI BRAK WYBORU TO PRZEKAZUJEMY WARTOŚĆ „”. MOŻLIWE, ŻE UŻYTKOWNIK BĘDZIE CHIAŁ PODAĆ SEKWENCJĘ CUSTOMOWĄ TO WTEDY PODAJE SEKWENCJĘ I NADAJE NAZWĘ.

## Transcripts 3rd level dictionary – schemat kolejności transkrypt ~ linker

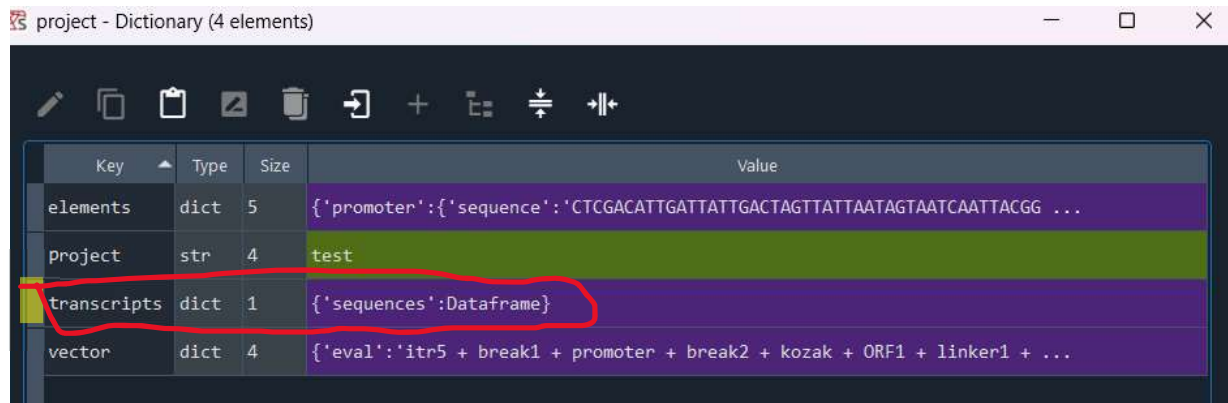


Index	Type	Size	Value
0	str	4	ORF1
1	str	7	linker1
2	str	4	ORF2

- Zależy od ilości transkryptów i linkerów (zasada n-1).

!NAZWY ORF1 ,2 ,3 SĄ NADAWANE W SKRYPTACH JAKO KOLEJNE TRANSKRYPTY (ORF – open reading frame – takie biologiczne stwierdzenie mówiące co jest czytane przez organizm żywy z kodu DNA/RNA). TYUTAJ ZAWARTA JEST TLKO INFORMACJA O KOLEJNOŚCI index 0,1,2 I ZAWSZE JEST transkrypt, linker, transkrypt linker. DANE O LINKERACH SĄ W LINKERS A DANE O SEKWENCJACH ORF ZNAJDUJĄ SIĘ W 2nd level dictionary [transcripts] > 3rd level data frame [sequences] – patrz niżej

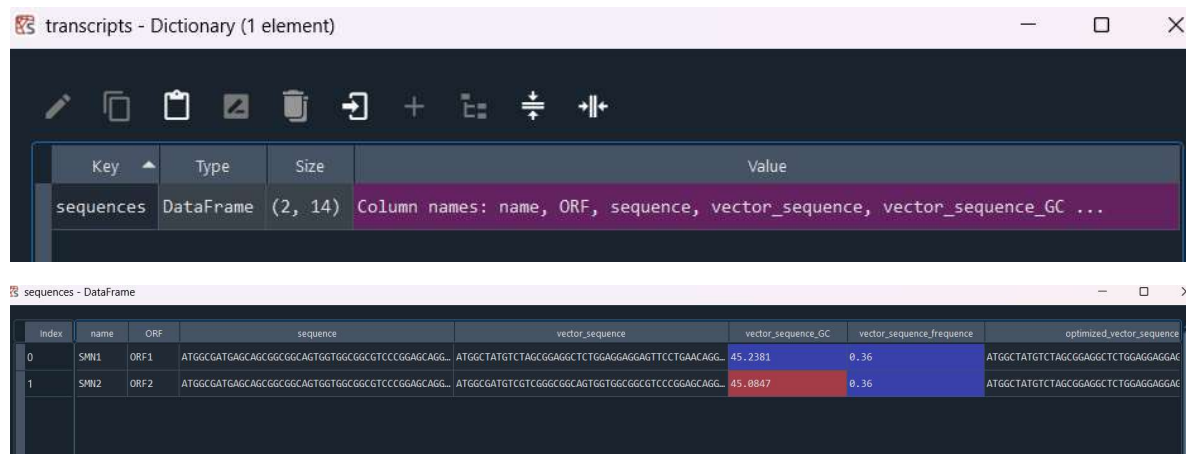
Transcripts 2nd level dictionary:



project - Dictionary (4 elements)

Key	Type	Size	Value
elements	dict	5	{'promoter':{'sequence':'CTCGACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGG ...
project	str	4	test
transcripts	dict	1	{'sequences':Dataframe}
vector	dict	4	{'eval':'itr5 + break1 + promoter + break2 + kozak + ORF1 + linker1 + ...

Sequences – data frame 3rd level – informacje o transkryptach



transcripts - Dictionary (1 element)

Key	Type	Size	Value
sequences	DataFrame	(2, 14)	Column names: name, ORF, sequence, vector_sequence, vector_sequence_GC ...

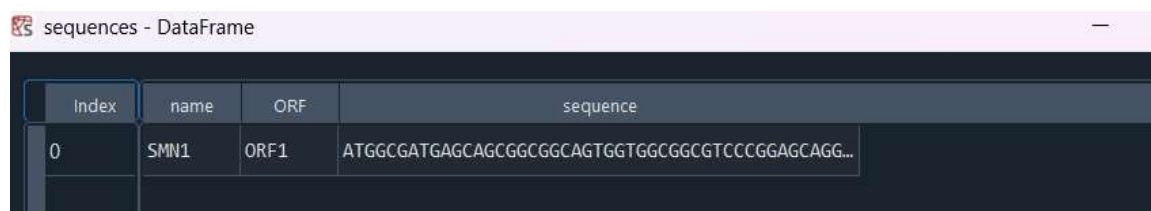
sequences - DataFrame

Index	name	ORF	sequence	vector_sequence	vector_sequence_GC	vector_sequence_frequency	optimized_vector_sequence
0	SMN1	ORF1	ATGGCGATGAGCAGCGGCGGAGTGGTGGCGGCGTCCCGGAGCAGG...	ATGGCTATGTCTAGCGGAGGCTCTGGAGGAGGATTCCTGAACAGG...	45.2381	0.36	ATGGCTATGTCTAGCGGAGGCTCTGGAGGAGGAG
1	SMN2	ORF2	ATGGCGATGAGCAGCGGCGGAGTGGTGGCGGCGTCCCGGAGCAGG...	ATGGCTATGTCTAGCGGAGGCTCTGGAGGAGGATTCCTGAACAGG...	45.0847	0.36	ATGGCTATGTCTAGCGGAGGCTCTGGAGGAGGAG

- Zawiera wszystkie informacje o sekwencji (nazwę, sekwencję, zawartość GC przed optymalizacją, po optymalizacji z usunięciem miejsc restrykcyjnych, ect.

!TUTAJ TRZEBA USTLIĆ JAK TO MA DZIAŁAĆ, BO SKRYPTY AKTUALNEI ODCZYTUJĄ TAKĄ FORMĘ!

Tak naprawdę na początek będzie nam potrzebne tylko poniższa wersja tabelki (data frame):



sequences - DataFrame

Index	name	ORF	sequence
0	SMN1	ORF1	ATGGCGATGAGCAGCGGCGGAGTGGTGGCGGCGTCCCGGAGCAGG...

Name – nazwa genu/transkryptu

ORF – który jest to ORF 1,2,3 ...

Sequence – sekwencja.

Pip do przetwarzania powyższych danych:

[https://github.com/jkubis96/IBioSeqTools/blob/dev\\_branch/vectro\\_build\\_app\\_pip.py](https://github.com/jkubis96/IBioSeqTools/blob/dev_branch/vectro_build_app_pip.py)

Rezultaty:

Do pokazania użytkownikowi [FRONTEND]

### Statystyki sekwencji:

[illegible]

Elementy wektora AAV:

index	element	sequence	start	end	length
0	itr5	CTGCGCGCTCGCTCGCTCACTGAGGCCGCCCGGGCAAAGCCCGGGC...	1	130	130
1	backbone_element	TCTAGACAACATTTGTATAGAAAAGTTG	131	157	27
2	promoter : CAG	CTCGACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGG...	158	1890	1733
3	backbone_element	CAAGTTTGTACAAAAAAGCAGGCT	1891	1914	24
4	kozak	GCCACC	1915	1920	6
5	ORF1 : SMN1	ATGGCTATGTCTAGCGGAGGCTCTGGAGGAGGAGTTCTCTGAACAGG...	1921	2805	885
6	backbone_element	ACCCAGCTTTCTTGTAAGTGGGAATTC	2806	2835	30
7	backbone_element	GAATTCCTAGAGCTCGCTGATCAGCCTCGA	2836	2865	30
8	bgh	CTGTGCCCTTAGTTGCCAGCCATCTGTTGTTGCCCTCCCCCGT...	2866	3073	208
9	backbone_element	GGGCCGC	3074	3080	7
10	itr3	CTGCGCGCTCGCTCGCTCACTGAGGCCGCCCGGGCAAAGCCCGGGC...	3081	3210	130
11	backbone_element	CTGCCCTCAGGGGCGCTGATGCGGTATTTCTCTTACGCATCTG...	3211	4137	927
12	amp	ATGAGTATTCAACATTTCCGTGTCGCCCTTATCCCTTTTTTGCGG...	4138	4998	861
13	backbone_element	CTGTGAGACCAAGTTTACTCATATATACTTTAGATTGATTTAAAA...	4999	5168	170
14	puc	TTGAGATCCTTTTTTCTGCGCGTAATCTGCTGCTTGCAAAACAAA...	5169	5757	589
15	backbone_element	AACGCCAGCAACGCGGCTTTTTACGGTTCTTGGCCTTTTGTGCG...	5758	5829	72

Sekwencja całego wektora w formacie FASTA:

>nazwa wektora

```
'CTGCGCGCTCGCTCGCTCACTGAGGCCGCCCGGGCAAAGCCCCGGGCGTCGGGCGACCTTTGGTCGCCCCGGCC
TCAGTGAGCGAGCGAGCGCGCAGAGAGGGAGTGGCCAACCTCCATCACTAGGGGTTCTTCTAGACAACTTTGT
ATAGAAAAGTTGCTCGACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCA
TATATGGAGTTCCGCGTTACATAACTTACGGTAAATGCCCCGCTGGCTGACCGCCCAACGACCCCCGCCATTG
ACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTAC
GGTAAACTGCCCCTTGGCAGTACATCAAGTGATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAA
ATGCCCCGCTGGCATTATGCCAGTACATGACCTTATGGGACTTTCTACTTGGCAGTACATCTACGTATTAGTCA
TCGCTATTACCATGGTCGAGGTGAGCCCCACGTTCTGCTTCACTCTCCCCATCTCCCCCCCCCTCCCCACCCCAATT
TTGTATTTATTTATTTTAAATTATTTTGTGACGATGGGGGCGGGGGGGGGGGGGGGGGCGCGCGCCAGGCG
GGGCGGGGCGGGGCGAGGGGCGGGGCGGGGCGAGGCGGAGAGGTGCGGCGGCGAGCCAATCAGAGCGGCG
CGCTCCGAAAGTTTCTTTTATGGCGAGGCGGCGGGCGGCGGCGGCCCTATAAAAAGCGAAGCGCGCGGGCGG
CGGGAGTCGCTGCGCGCTGCCTTCGCCCCGTGCCCGCTCCGCCGCCGCTCGCGCCGCCCGCCCGGCTCTGA
CTGACCGCGTTACTCCACAGGTGAGCGGGCGGGACGGCCCTTCTCCTCCGGGCTGTAATTAGCGCTTGTTTAA
ATGACGGCTTGTTTCTTTTCTGTGGCTGCGTGAAAGCCTTGAGGGGCTCCGGGAGGGGCCCTTGTGCGGGGGG
AGCGGCTCGGGGGGTGCGTGCGTGTGTGTGCGTGGGGAGCGCCGCGTGCGGCTCCGCGCTGCCCGGCGG
CTGTGAGCGCTGCGGGCGCGGCGCGGGGCTTTGTGCGCTCCGAGTGTGCGCGAGGGGAGCGCGGCCGGGG
GCGGTGCCCGCGGTGCGGGGGGGGCTGCGAGGGGAACAAAGGCTGCGTGCGGGGTGTGTGCGTGGGGGG
GTGAGCAGGGGGTGTGGGCGCGTGGTGGGCTGCAACCCCCCTGCACCCCCCTCCCGAGTTGCTGAGCAC
GGCCCGCTTCGGGTGCGGGGCTCCGTACGGGGCGTGCGCGGGGCTCGCCGTGCCGGGCGGGGGGTGGCG
GCAGGTGGGGGTGCCGGGCGGGGCGGGGCCGCTCGGGCCGGGAGGGCTCGGGGGAGGGGCGCGGCGG
CCCCCGGAGCGCCGGCGGCTGTGAGGCGCGGCGAGCCGAGCCATTGCCTTTTATGGTAATCGTGCGAGAGG
GCGCAGGGACTTCTTTTGTCCCAAATCTGTGCGGAGCCGAAATCTGGGAGGCGCCGCCGACCCCTCTAGCG
GGCGCGGGGGCGAAGCGGTGCGGCGCCGCGAGGAAGGAAATGGGCGGGGAGGGCCTTCGTGCGTCGCCGCG
CCGCCGTCCCCTTCTCCCTCTCCAGCCTCGGGGCTGTCCGCGGGGGGACGGCTGCCTTCGGGGGGGACGGGG
CAGGGCGGGGTTTCGGCTTCTGGCGTGTGACCGGCGGCTCTAGAGCCTCTGCTAACCATGTTTCATGCCCTTCTT
TTTCTACAGCTCCTGGGCAACGTGCTGTTATTGTGCTGTCTCATATTTTGGCAAAGAATTGCAAGTTTGTACA
AAAAAGCAGGCTGCCACCATGGCTATGTCTAGCGGAGGCTCTGGAGGAGGAGTTCCTGAACAGGAGGACTCTG
TGCTGTTCCGGAGGGGACAGGACAAAGCGATGACAGCGACATCTGGGACGACACCGCTCTGATTAAGGCCTA
CGACAAGGCCGTGGCCTCTTCAAGCACGCCCTGAAGAACGGCGACATCTGCGAGACCAGCGGAAAGCCTAAA
ACCACCCCTAAGAGAAAGCCTGCTAAAAAGAACAAGAGCCAGAAGAAGAACACCGCCGCCAGCCTGCAGCAG
TGGAAGGTGGGCGACAAGTGACGCGCCATTTGGAGCGAGGACGGATGTATCTACCCTGCCACAATGCCTCTAT
CGACTTCAAGCGGGAGACCTGCGTGGTGGTGATACCGGCTACGGCAACAGGGAAGAGCAGAACCTGAGCGA
CCTGCTGAGCCCTATTTGCGAGGTGGCCAATAACATCGAGCAGAACGCCAGGAGAACGAGAACGAGAGCCAG
GTGAGCACCGACGAGAGCGAGAACAGCCGAGCCCCGGCAATAAGAGCGACAACATCAAGCCCAAGAGCGCC
CCCTGGAACCTCTTCTGCCCCCTCTCCTCCTATGCCTGGACCCAGACTGGGACCCGAAAACCTGGACTGAAA
TTCAACGGCCCCCTCTCCTCCTCCTCCTCCTCCTCCTCCTCCTCCTCCTCCTCCTCCTCCTCCTCCTCCTG
ACCCCTATTATCCCCCTCTCCTCCTATCTGTCTGATTCTCTGGACGACGCCGATGCTTTGGGCTCTATGCTGAT
CTTTGGTACATGAGCGGCTACCACACCGGCTACTACATGGGCTTCCGGCAGAACCAGAAGGAGGGCCGGTGC
AGCCACTCTCTGAACCTGAACCCAGCTTTCTGTACAAAGTGGGAATTCGAATTCCTAGAGCTCGCTGATCAGCCT
CGACTGTGCCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCCTCCCCGTGCCCTTCCTTGACCCTGGAAGGTGCCA
CTCCCACTGTCCTTCTCTAATAAAATGAGGAAATGCATCGCATTGTCTGAGTAGGTGTCATTCTATTCTGGGGGG
TGGGGTGGGGCAGGACAGCAAGGGGGAGGATTGGGAAGAGAATAGCAGGCATGCTGGGGAGGGCCGCCTG
CGGCTCGCTCGCTCACTGAGGCCGCCCGGGCAAAGCCCCGGGCGTCGGGCGACCTTTGGTCGCCCCGCCCTCA
GTGAGCGAGCGAGCGCGCAGAGAGGGAGTGGCCAACCTCCATCACTAGGGGTTCTCTGCCTGCAGGGGCGCC
TGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTACACCGCATACGTCAAAGCAACCATAGTACGCGCCC
```



TGTAGCGGCGCATTAAAGCGCGGCGGGGGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCTTAG  
CGCCCGCTCCTTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCCGCCGCTTTCCCCGTCAAGCTCTAAATCGGGG  
GCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAACTTGATTTGGGTGATGGTTCACG  
TAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTT  
GTTCCAAACTGGAACAACACTCAACTCTATCTCGGGCTATTCTTTTGATTATAAGGGATTTTGCCGATTTTCGGTC  
TATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGTTTACAATTTTATG  
GTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGCCCCGACACCCGCCAACACCCGCTGACGC  
GCCCTGACGGGCTTGCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTC  
AGAGGTTTTACCGTCATCACCGAAACGCGCGAGACGAAAGGGCCTCGTGATACGCCTATTTTTATAGGTTAATG  
TCATGATAATAATGGTTTCTTAGACGTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATT  
TTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCTGATAAATGCTTCAATAATATTGAAAAAGGA  
AGAGTATGAGTATTCAACATTTCCGTGTCGCCCTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCAC  
CCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTACATCGAACTGGATC  
TCAACAGCGGTAAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGC  
TATGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTGCGCGCATACACTATTCTCAGAATG  
ACTTGGTTGAGTACTACCAAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCT  
GCCATAACCATGAGTGATAAACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGC  
TTTTTTGCACAACATGGGGGATCATGTAACCTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAA  
ACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACCTATTAAGTGGCGAACTACTT  
ACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGC  
CCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGAAGCCGCGGTATCATTGCAGCAC  
TGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGA  
AATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATATAC  
TTTAGATTGATTTAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAA  
ATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCT  
TTTTTTCTGCGCGTAATCTGCTGCTTGCAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAA  
GAGCTACCAACTCTTTTTCCGAAGGTAAGTGGCTTACGAGAGCGCAGATACCAAATACTGTTCTTCTAGTGTAG  
CCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAAGTGG  
CTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGG  
TCGGGCTGAACGGGGGGTTCGTGCACACAGCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTAC  
AGCGTGAGCTATGAGAAAGCGCCACGCTTCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGG  
GTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCTGTGCGGTTTC  
GCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAA  
CGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTCCTGCAGGCAG'

Wykres wektora:

