How to get coding sequences of proteins of interest from plants that have no genome data available

Kaia Kukk[1]

[1]Latvian Biomedical Research and Study Centre

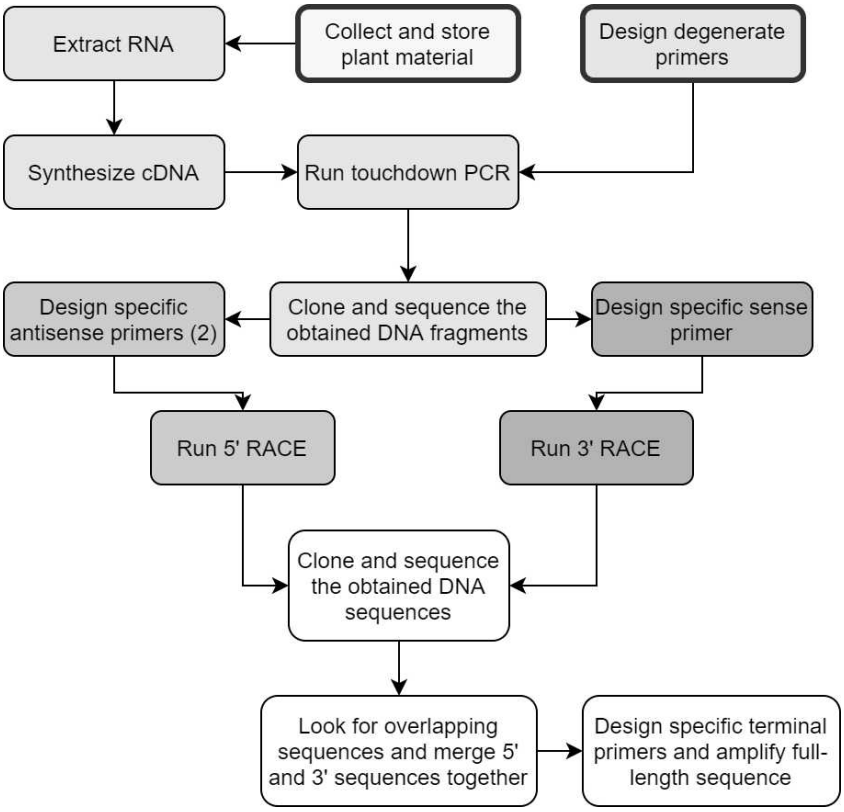
Apr 16, 2021

1 *Works for me* dx.doi.org/10.17504/protocols.io.btxwnppe

Kaia Kukk
Latvian Biomedical Research and Study Centre

ABSTRACT

The protocol describes obtaining coding DNA sequences of plant proteins when genomic data is not available, but there are sequences of orthologs. The protocol was created during implementing the project no 1.1.1.2/VIAA/2/18/286 "Optimization of novel plant derived enzyme expression in microorganisms for biotechnological application" funded by European Regional Development Fund. Only the first two steps of the protocol are specific to plant material, the rest is universal.



DOI

dx.doi.org/10.17504/protocols.io.btxwnppe

PROTOCOL CITATION

Kaia Kukk 2021. How to get coding sequences of proteins of interest from plants that have no genome data available. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.btxwnppe

KEYWORDS

5' RACE, 3' RACE, touchdown PCR, colony PCR, degenerate primer, RNA extraction, cDNA synthesis

CREATED

Apr 05, 2021

LAST MODIFIED

Apr 16, 2021

PROTOCOL INTEGER ID

48854

MATERIALS TEXT

Spectrum™ Plant Total RNA Kit **Sigma – Aldrich Catalog #STRN50**

Phusion High-Fidelity DNA Polymerase (2 U/µL) **Thermo Fisher Catalog #F530L**

GeneJET Gel Extraction and DNA Cleanup Micro Kit **Thermo Fisher Catalog #K0832**

FirstChoice™ RLM-RACE Kit **Thermo Fisher Catalog #AM1700**

GeneJET Plasmid Miniprep Kit **Thermo Fisher Catalog #K0503**

CloneJET PCR Cloning Kit **Thermo Fisher Catalog #K1232**

Liquid nitrogen **Contributed by users**

Taq DNA Polymerase, recombinant (5 U/µL) **Thermo Fisher Catalog #EP0401**

T4 DNA Ligase (5 U/µL) **Thermo Fisher Catalog #EL0011**

RevertAid H Minus First Strand cDNA Synthesis Kit **Thermo Fisher Catalog #K1631**

Conventional laboratory equipment and materials used for cloning such as agarose gel electrophoresis apparatus, competent *E. coli* DH5alpha cells, EtBr etc.

Collecting and storing plant material

1 Collect plant material, keep it fresh and freeze as soon as possible in liquid nitrogen. Store at **-70 °C** or lower.

Citation: Kaia Kukk (04/16/2021). How to get coding sequences of proteins of interest from plants that have no genome data available. https://dx.doi.org/10.17504/protocols.io.btxwnppe

RNA extraction and cDNA synthesis

2 Extract RNA. For simple tissues such as green leaves use Protocol 1 described in:

> Luis Oñate-Sánchez & Jesús Vicente-Carbajosa (2008). DNA-free RNA isolation protocols for Arabidopsis thaliana, including seeds and siliques. BMC Research Notes.
> http://doi:10.1186/1756-0500-1-93

Anthocyanin and/or polysaccharide rich tissues are more complex subjects. Use Spectrum Plant Total RNA Kit for such material.

RNA quality was the highest when freshly picked plant material was frozen in liquid nitrogen and directly subjected to RNA extraction (Figure 1). Native (not denaturing) agarose gel electrophoresis and OD260/OD280 (~2.0) were enough to assess the quality of RNA. We used Spectrum Plant Total RNA Kit according to manufacturer's instructions for extracting RNA from bilberry berries and flowers of cornflower and got enough RNA for downstream experiments.

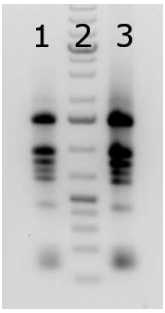


Figure 1. Native agarose gel electrophoresis of RNA extracted from freshly picked leaf of sunflower. 1. Sample 1: 600 ng/µl of RNA, OD260/OD280 = 2.04; 2. GeneRuler 1 kb Plus DNA Ladder, 3. Sample 2: 870 ng/µl of RNA, OD260/OD280 = 1.71. Colors reversed. RNA was extracted using Protocol 1 by Oñate-Sánchez & Vicente-Carbajosa, 2008.

3 Synthesize cDNA using RevertAid H Minus First Strand cDNA Synthesis Kit.

Degenerate primer design

4 Search for available protein sequences of other similar species from databases like NCBI (Protein or Nucleotide from drop-down menu) and UniProt. If available, save both nucleotide and amino acid sequences, and reference numbers. Check alternative protein names from KEGG or BRENDA.

5 Align amino acid sequences, using MegAlign program from DNAStar, COBALT: Alignment Tool, Clustal Omega (fasta file format) or some other multiple sequence alignment tool. Look for conserved regions and among them regions that contain amino acid residues that are encoded by smaller number of codons (Table 1, for example methionine, tryptophan, asparagine, aspartate, cysteine, glutamine, glutamate, histidine, lysine, phenylalanine, tyrosine). The explanations for non-A-T-G-C nucleotides are presented in Table 2.

Citation: Kaia Kukk (04/16/2021). How to get coding sequences of proteins of interest from plants that have no genome data available. https://dx.doi.org/10.17504/protocols.io.btxwnppe

Table 1. Amino acids and the respective codons (inspired by Wikipedia, CC BY-SA)

A	B	C	D	E	F
Amino acid	**DNA codons**	**Compressed**	**Amino acid**	**DNA codons**	**Compressed**
Ala / A	GCT, GCC, GCA, GCG	GCN	Leu / L	TTA, TTG, CTT, CTC, CTA, CTG	YTR, CTN
Arg / R	CGT, CGC, CGA, CGG, AGA, AGG	CGN, AGR	Lys / K	AAA, AAG	AAR
Asn / N	AAT, AAC	AAY	Met / M	ATG	
Asp / D	GAT, GAC	GAY	Phe / F	TTT, TTC	TTY
Cys / C	TGT, TGC	TGY	Pro / P	CCT, CCC, CCA, CCG	CCN
Gln / Q	CAA, CAG	CAR	Ser / S	TCT, TCC, TCA, TCG, AGT, AGC	TCN, AGY
Glu / E	GAA, GAG	GAR	Thr / T	ACT, ACC, ACA, ACG	CAN
Gly / G	GGT, GGC, GGA, GGG	GGN	Trp / W	TGG	
His / H	CAT, CAC	CAY	Tyr / Y	TAT, TAC	TAY
Ile / I	ATT, ATC, ATA	ATH	Val / V	GTT, GTC, GTA, GTG	GTN
START	ATG	STOP	TAA, TGA, TAG	TRA, TAG	

Table 2. Symbols of nucleotides (inspired by Wikipedia, CC BY-SA)

Symbol	Description	Bases represented					Symbol	Description	Bases represented				
A	adenine	A					**K**	keto			G	T	
C	cytosine		C				**R**	purine	A		G		2
G	guanine			G		1	**Y**	pyrimidine		C		T	
T	thymine				T		**B**	not A		C	G	T	
U	uracil				U		**D**	not C	A		G	T	3
W	weak	A			T		**H**	not G	A	C		T	
S	strong		C	G		2	**V**	not T	A	C	G		
M	amino	A	C				**N**	any base	A	C	G	T	4

A degenerate primer is defined as a mix of oligonucleotide sequences in which some positions contain a number of possible bases, giving a set of specific primers with similar sequences that cover all possible nucleotide combinations for a given protein sequence. For example, the degenerate primer AAD GCN TGY GTN GAH ATG GAY CAR AA is a mix of 3 × 4 × 2 × 4 × 3 × 2 × 2 = 1152 primers and it corresponds to amino acid sequence N/K-ACV-D/E-MDQ. The number 1152 is also called degeneracy of the primer.

6 Try to find sequence that has melting temperature (Tm) of at least 54 °C and low degeneracy. Calculate both the lowest and highest potential Tm. Design the primers such a way that the amplified sequence will be approximately 200-500 bp long (both forward and reverse primer is needed). The lower the degeneracy the longer can the distance between the primers be.

We used Tm calculator at Thermo Fisher web page for Phusion polymerase. We have successfully used a primer mix the degeneracy of which was 1152.

Citation: Kaia Kukk (04/16/2021). How to get coding sequences of proteins of interest from plants that have no genome data available. https://dx.doi.org/10.17504/protocols.io.btxwnppe

7 Check potential dimers of primers, for example at web page of Thermo Fisher. These should be avoided.

8 Avoid degeneracy at the 3' end of the primer.

It will increase the price of the primer noticeably.

Touchdown PCR

9 Run touchdown PCR with the following reaction mix composition: 2 µl cDNA, 0.2 mM dNTPs, 2 µM degenerate sense primer, 2 µM degenerate antisense primer, 1 × HF buffer, 1 U Phusion High-Fidelity DNA Polymerase (2 U/µL) and water to 50 µl. Start with the highest potential annealing temperature and end at lowest annealing temperature minus a few degrees.

The PCR program we used successfully is presented below in Table 3. In conventional PCR primers are used at 0.5 µM final concentration. Here, the final concentration is 2 µM, i.e. four times higher.

Table 3. The program of touchdown PCR used for amplifying 250-600 bp DNA fragments with degenerate primers and Phusion polymerase

A	B	C	D	E
	Step 1	Step 2	Step 3	Step 4
1. Initial denaturation	98 °C 30 sec			
2. Denaturation	98 °C 10 sec	98 °C 10 sec	98 °C 10 sec	98 °C 10 sec
3. Annealing	-	65 °C 30 sec	58 °C 30 sec	52 °C 30 sec
4. Extension	72 °C 20 sec	72 °C 20 sec	72 °C 20 sec	72 °C 20 sec
5. Final Extension	-	-	-	72 °C 300 sec
End	-	-	-	Hold 4 °C
Cycles of steps 2-4	5	5	5	25

10 Check the PCR products in 1% agarose gel with EtBr. Purify fragments with expected size from agarose gel using GeneJET Gel Extraction and DNA Cleanup Micro Kit.

11 Ligate fragments into pJET1.2 blunt vector using T4 DNA ligase (at room temperature for 1-2 h), transform *E. coli* DH5alpha cells and spread on LB+Amp (50 µg/ml) plate using sterile glass beads.

12 Analyze colonies by running colony PCR with Taq polymerase and pJET1.2 sequencing primers.

Pick a colony with a toothpick or a small (yellow) pipette tip and smear on the inner bottom wall of a 0.2 ml PCR

Citation: Kaia Kukk (04/16/2021). How to get coding sequences of proteins of interest from plants that have no genome data available. https://dx.doi.org/10.17504/protocols.io.btxwnppe

tube. Add 20 µl of the following PCR mix: 2.5 mM $MgCl_2$, 0.25 mM dNTPs, 1 × Taq polymerase buffer, 0.5 µM pJET1.2 forward sequencing primer, 0.5 µM pJET1.2 reverse sequencing primer and Taq polymerase 0.2 µl (1 U). Mix and spin. Run the following PCR program: 95°C 5 min, 30 cycles of 95°C 30 s, 60°C 30 s and 72 °C 1 min/kb; and final extension 72 °C 200 s. Run agarose gel electrophoresis. You should see DNA bands with the size of the insert plus 120 bases.

13 Inoculate 3-5 ml LB+Amp (50 µg/ml) liquid medium with a single colony, grow over night and extract plasmid DNA with GeneJET Plasmid Miniprep Kit.

14 Sequence the fragments with pJET1.2 forward and reverse sequencing primers.

3’ RACE (Rapid amplification of cDNA ends)

15 Use the sequences of fragments obtained in previous step for designing specific sense primer(s). The Tm of the primer should be at least 62 °C.

Although both outer and inner sense primers may be required for amplifying enough product, we managed to get 1 kb sequences with one primer.

16 As the Tm of Oligo$(dT)_{18}$ is only 44 °C, use 3-RACE-CDS primer (5’-AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’) and RevertAid H minus reverse transcriptase to synthesize 3’ RACE ready cDNA. Run conventional PCR using the specific sense primer and the primer 5’-AAGCAGTGGTATCAACGCAGAGT-3’ (Tm 67 °C).

17 Clean the fragments from agarose gel, clone into pJET1.2, transform *E. coli*, analyse the colonies, do mini-preps and sequence as before.

5' RACE

18 Again, use the sequences of fragments obtained in step 14 for designing specific antisense primers. The Tm of the primers should be at least 62 °C. Design two specific primers (i.e. outer and inner antisense primers), if possible, with a 100 bp distance from each other.

19 Perform 5’ RACE using tailing (e.g.SMARTer RACE 5’/3’ Kit) or adapter ligation (e.g. FirstChoice RLM-RACE Kit) method followed by two rounds of PCR with gene specific outer and inner antisense primers and the primers complementary to the tail/adapter sequence.

Although there is protocol

> Pinto FL, Lindblad P (2010). A guide for in-house design of template-switch-based 5' rapid amplification of cDNA ends systems.. Analytical biochemistry.
> https://doi.org/10.1016/j.ab.2009.10.022

that describes obtaining 5’ nucleotide sequences without having to buy an expensive kit, we had no success with

Citation: Kaia Kukk (04/16/2021). How to get coding sequences of proteins of interest from plants that have no genome data available. https://dx.doi.org/10.17504/protocols.io.btxwnppe

this. However, using the kit (we tested FirstChoice RLM-RACE kit) did not guarantee success as well. Obtaining 5' sequences up to 0,5 kb for proteins that do not have isoforms encoded by similar sequences is likely more probable.

20 Clean the fragments from agarose gel, clone into pJET1.2, transform *E. coli*, analyse the colonies, do mini-preps and sequencing as before.

Full-length sequence

21 Analyze 5' and 3' sequences, look for overlapping sequences and combine the fragments into full-length sequence. Look for open reading frames.

22 Design specific primers annealing to 5' and 3' ends of the potential open reading frame and amplify full-length sequence using the cDNA synthesized at step 3 or 16.

23 Purify the potential full-length DNA sequence in gel, clone into pJET1.2, transform *E. coli*, analyse the colonies, do mini-preps and sequencing.

24 Deposit the sequence in a repository (for example GenBank and/or PlutoF).

Citation: Kaia Kukk (04/16/2021). How to get coding sequences of proteins of interest from plants that have no genome data available. https://dx.doi.org/10.17504/protocols.io.btxwnppe