

Design an Expression Construct

The very basic procedure is to find out the gene sequence expressing the protein we want to clone, and to insert the gene sequence into a specific vector.

Protein Sequence to Gene Sequence

Obtain the Protein Sequence in PDB

The screenshot shows the RCSB PDB website interface. At the top, there's a navigation bar with links for Addgene, BLAST, RCSB PDB, Homo sapiens, pET-14b, Homo sapiens, Addgene, regex101, pET-14b, ma restrain, python, 知乎, and others. Below the bar, the RCSB PDB logo is visible along with statistics: 242,874 structures from the PDB archive and 1,068,577 Computed Structure Models (CSM). A search bar allows users to enter search terms, Ligand ID or sequence. A sidebar on the right includes links for Advanced Search, Browse Annotations, Help, Contact us, and MyPDB. Below the header, there are social media icons for Facebook, Twitter, YouTube, and LinkedIn. The main content area displays the protein structure 4JV6, which is a crystal structure of PDE6D in complex to inhibitor-1. It features a ribbon diagram of the protein with various colored regions (orange, green, blue) and a small inhibitor molecule shown as sticks and spheres. To the right of the structure, detailed information is provided: PDB ID (4JV6), PDB DOI (https://doi.org/10.2210/pdb4JV6/pdb), Classification (Protein binding/inhibitor), Organism(s) (Homo sapiens), Expression System (Escherichia coli), and Mutation(s) (No mutations). Below this, the Experimental Data Snapshot section lists Method (X-RAY DIFFRACTION), Resolution (1.87 Å), R-Value Free (0.227 (Depositor), 0.230 (DCC)), R-Value Work (0.175 (Depositor), 0.170 (DCC)), R-Value Observed (0.178 (Depositor)), and Starting Model (experimental). Further down, the wwPDB Validation section provides percentile ranks and values for metrics like Rfree, Clashscore, Ramachandran outliers, Sidechain outliers, and RSRZ outliers. A download button for FASTA Sequence is highlighted with a red box. The bottom of the page shows browser tabs for Explore in 3D, Sequence Annotations, Electron Density, and Validation Report, along with a ligand interaction section. The URL https://www.rcsb.org/fasta/entry/4JV6/display is visible at the bottom left, and the date 2025/10/2 is at the bottom right.

So we get the protein FASTA sequence.

More information about the protein sequence is available in the Sequence Tab or Uniprot.

BLAST Protein to Gene

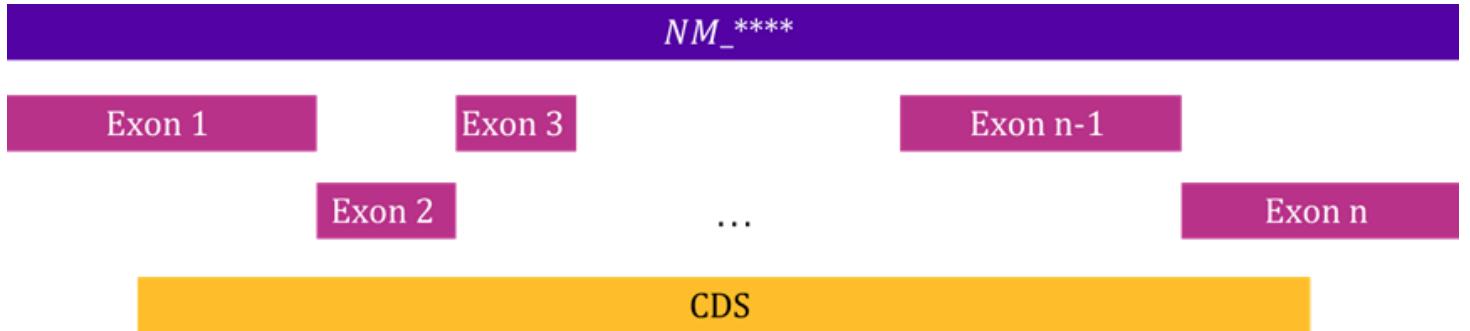
The BLAST tool used here is tblastn, which 'translates' a protein sequence into a nucleic acid sequence.

Remember to change Database to RefSeq Select RNA sequences (refseq_select).

On the new page, information such as biological role, Function/mechanism, Tissue of origin, Disease relevance/known mutation can be found. This time, we just focus on its GeneBank **accession**.

Descriptions	Graphic Summary	Alignments	Taxonomy	Sequences producing significant alignments								Download	Select columns	Show	100	?	
																GenBank	Graphics
<input checked="" type="checkbox"/> select all	6 sequences selected							Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	
<input checked="" type="checkbox"/>	Homo sapiens phosphodiesterase 6D (PDE6D). transcript variant 1.mRNA							Homo sapiens	312	312	99%	3e-107	100.00%	1140	NM_002601.4		
<input checked="" type="checkbox"/>	Mus musculus phosphodiesterase 6D, cGMP-specific, rod, delta (Pde6d). transcript variant 1.mRNA							Mus musculus	306	306	99%	3e-105	98.00%	1129	NM_008801. Show report for NM_002601.4		
<input checked="" type="checkbox"/>	Rattus norvegicus phosphodiesterase 6D (Pde6d). mRNA							Rattus norvegicus	305	305	99%	2e-104	97.33%	1136	NM_001108806.2		
<input checked="" type="checkbox"/>	Rattus norvegicus unc-119 lipid binding chaperone (Unc119). mRNA							Rattus norvegicus	45.1	45.1	63%	2e-05	26.26%	1264	NM_017188.1		
<input checked="" type="checkbox"/>	Mus musculus unc-119 lipid binding chaperone (Unc119). transcript variant 2.mRNA							Mus musculus	44.7	44.7	63%	3e-05	26.26%	1332	NM_011676.3		
<input checked="" type="checkbox"/>	Homo sapiens unc-119 lipid binding chaperone (UNC119). transcript variant 1.mRNA							Homo sapiens	44.3	44.3	63%	5e-05	26.26%	1379	NM_005148.4		

Note that, entries start with accession *NM_* indicate that they are mRNAs. The whole sequence structures in these entries can be represented in the following figure:



Regions beyond the CDS but included in exons are 5' untranslated region (5' UTR) and 3' untranslated region (3'UTR). These sequences are not translated into proteins, but facilitate the initiation or regulation of translations.

Those entries starting with *NG_* or *NC_*, which contain introns, refer to *genome reference sequences* and *chromosome sequences*.

Focus on the following fields in the new page:

Gene: Clicking 'gene' will show you a whole gene sequence at the bottom of the page.

Exon: click 'exon'. A part of the gene sequence will be highlighted.

CDS: click 'CDS' (Coding Sequence). Highlighted part reflects what is translated by a ribosome. The sequence starts with a start codon and terminates with a terminating codon. We can also find a translated protein sequence at the bottom-left area.

What we want in this job is the codon sequence, since some genomic fragments in the whole gene or exons do not express protein, clicking **Display: FASTA** enables us to copy the codon sequence.

Now, we have finished all preparation steps. We will move to [Benchling Lab](#) and start inserting the obtained gene sequence into a vector.

Obtain Vector Sequence

The vector is first obtained via [Addgene](#). Search for the vector we want and copy the sequence.

Insert Sequence into Vector in Benchling Lab

Go back to [Benchling Lab](#), and create a new DNA sequence.

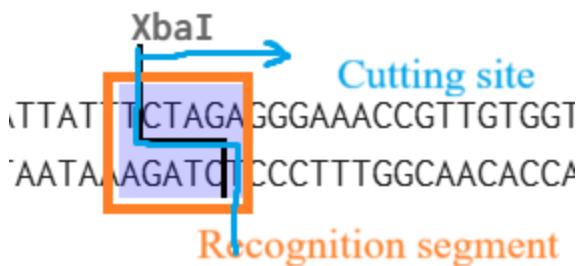
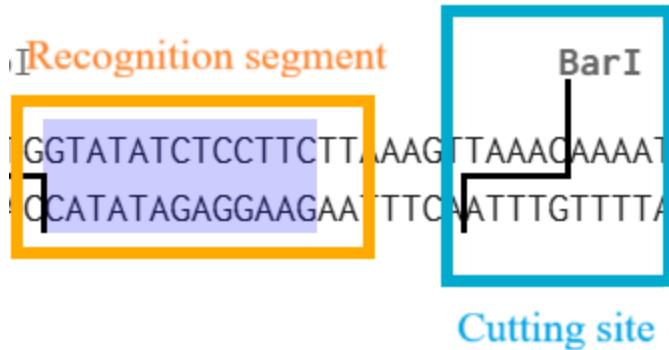
The screenshot shows the Benchling web application interface. At the top, there are several browser tabs: '0e79a011' (PDF), 'RCSB PDB', 'rcsb.org/f...', '1 组', 'Addgene...', and a search bar with the URL 'https://benchling.com/yxtang/f/lib_GaEs5hDbE9-example-project/seq_NU'. On the left, a vertical sidebar has a blue header with icons for Project, Entry, Protocol, a magnifying glass, and a plus sign. Below this is a list of options: 'DNA / RNA sequence' (selected and highlighted in teal), 'AA sequence', 'Oligo', 'Assembly', 'CRISPR', 'Entity from schema', 'Mixture', and 'More'. A tooltip for 'DNA / RNA sequence' says 'pBR322_EGFR SEQUENCE MAP'. To the right of the sidebar, a modal window titled 'New DNA / RNA sequence' is open, showing a sequence map for 'pBR322_EGFR'. The sequence starts with 'gcgcagcg' and ends with 'ctcgccgccaacgccacaaccaccgcgcacg'. A restriction enzyme site 'NotI' is indicated with a vertical line and a bracket. The sequence map includes a scale from 0 to 100. Below the sequence map, there are four options: 'New DNA / RNA alignment', 'Import DNA / RNA sequences', 'Import DNA / RNA sequences from spreadsheet', and 'Assemble DNA / RNA sequences by concatenation'. The 'Import DNA / RNA sequences' option has a small 'YT' icon next to it.

Add Annotations

Some annotations can be added automatically, or add them manually according to instructions.

Insertion

The recognition segment and the cutting site are shown when hovering mouse over a restriction enzyme name.



One critical point to be aware of is how to select the right place to insert the gene.

Taking the second restriction enzyme as an example, the purple highlighted area is the enzyme recognition segment, with a zig-zag solid line referring to the cutting site.

Insertion Location

The insertion of nucleic acid sequence must locate at the right side of A-T pair or the left side of T-A pair on margins of the recognition segment, although the cutting is irrelevant to T-A and A-T, because the insertion within recognition segment may 'damage' the recognition segment.

Avoid self-rejoin

Another point is that both terminals of the cutting site should not be compatible with each other, aiming to avoid the self-rejoin of plasmid terminals

Avoid frame-shift mutation

Moreover, the inserted nucleic acid sequence must line up with the ORF and have a sequence length of $3n$ so there is no 'frame-shift' mutation.

Optimize Inserted Codons

One important consideration is that the insert should not be cut by the restriction enzymes that will be used in cloning.

If we are getting the coding sequence synthesised, we can do two things at once: make sure that

the codons that are being used give the protein the best chance at high expression, and avoid sites recognised by restriction enzymes we are planning on using.

Two objectives:

- Make sure there are no restriction enzyme recognition sites.
- Balance the base usage.

Note that, the optimised codons will not replace the original codons automatically. Copy optimised codons and substitute the original codons to apply the optimisation.

Insertion Direction

Another point to be aware of is the 'direction' of inserted gene segments. Since the DNA is a double-stranded structure, if the protein we desire is expressed by the gene segment on the reverse chain, we should not simply insert the gene in a forward direction, but its reversed complementary nucleic acid sequence.

One example is the question in the homework:

If you have time, try performing the same process with a prokaryotic expression vector: [pET-14b](#) for expression in E. coli with an N-terminal hexahistidine tag. To do this you need to make sure that the codons of the insert line up with the codons of the histidine tag so there is no 'frame-shift' mutation.

The gene sequence expressing 6 His tag is located on the reverse chain. Meanwhile, we have to keep the tag on the N-terminal of the expressed protein.

I first tried to address the problem by reversing the string and mapping its complementary base in a dict in Python:

```
forward_seq = 'ATGTCAGCCAAGGACGAGCGGCCAGGGAGATCCTGAGGGGCTT'
pattern = {
    'A':'T', 'T':'A',
    'C':'G', 'G':'C'
}
implementary_seq = ''

for i in forward_seq[::-1]:
    i_ = pattern[i]
    implementary_seq += i_

print(implementary_seq)
```

But this is obviously not very elegant.

An alternative solution I came up with is to create a new sequence in Benchling Lab, after which, copy its DNA reverse complement.

Copy special



DATA

The following bases were copied from the sequence.

Click the sequence type you would like to copy.

DNA sequence

```
ATGGCGCGTCCTCCCTGGAACAGAAAG
CTGTCCCGCCTGGAAGCAAAGCTGAAG
CAGGAGAACCGGGAGGCCGGCGGAGG
ATCGACCTCAACCTGGATATCAGCCCC
CAGCGGGCCCAGGCCCACCCCTGCAGCTC
```

DNA reverse complement

```
CTACCTGAAGAACGGGCAGGTGGGGCTG
GCTCAGGACGCCGCTAGTCCGCCGTGA
CTCAGTCTTCGCCATGACATCCTTGAA
CCAGGACGCCACGTCCACCTCCAGCGT
CTCGTAGCGCTTGTGAAGCTGTGTT
```

AA translation

```
MAASSLEQKLSRLEAKLKQENREARRR
IDLNLISPQRPRPTLQLPLANDGGSR
SPSSESSPQHPTPPARPRHMLGLPSTL
FTPRTSMESIEIDQKLQEIMKQTGYLTI
GGQRYQAEINDLENLGEMGSGTCGQWV
```

AA reverse translation

```
LPEEGQVGLAQDAASPR*LSLRHDILE
PGRHVHLQRLVALDEAVFK*LIILWSL
PVIFSKAVFDEGLEVPREAHVSGQKRG
LFL*DFGEDLKVRALVLVGKLSCCQLH
QRDAQAPYVGPDVIVRLGGVGVNALG
```

RNA sequence

```
AUGGCGCGGUCCUCCCUGGAACAGAAAG
CUGUCCCGCCUGGAAGCAAAGCUGAAG
CAGGAGAACCGGGAGGCCGGCGGAGG
AUCGACCUAACCUGGAUAUCAGCCCC
CAGCGGGCCCAGGCCCACCCUGCAGCUC
```

RNA reverse complement

```
CUACCUGAAGAACGGGCAGGUUGGGGCUG
GCUCAGGACGCCGCUAGUCCGCCGUGA
CUCAGUCUUCGCCAUGACAUCCUUGAA
CCAGGACGCCACGUCCACCUCCAGCGU
CUCGUAGCGCUUGAUGAAGCUGUGUUC
```

Additional options when copying the sequence or its reverse complement inside Benchling:

- Include annotations/translations
- Include annotations/translations not fully contained by selection

Add Kozak Consensus Sequence

Efficient initiation of translation in eukaryotes requires the presence of the Kozak consensus sequence, most commonly GCCGCCACCAUGG where 'AUG' corresponds with the start

methionine ATG codon. This particular plasmid lacks this sequence, so we can add it by placing the cursor before the start codon and typing GCCGCCACC without anything else selected.

Verify the Construct in InterPro

In this section, I tried to identify whether the InterPro search is sensitive to the inputted sequence. Thus I uploaded:

- The forward protein sequence expressed by the inserted gene segment
- The reverse protein sequence expressed by the inserted gene segment
- The codon-optimized reverse protein sequence expressed by the inserted gene segment
- The reverse protein sequence with a 6-His Tag

Results are shown in the following screenshots:

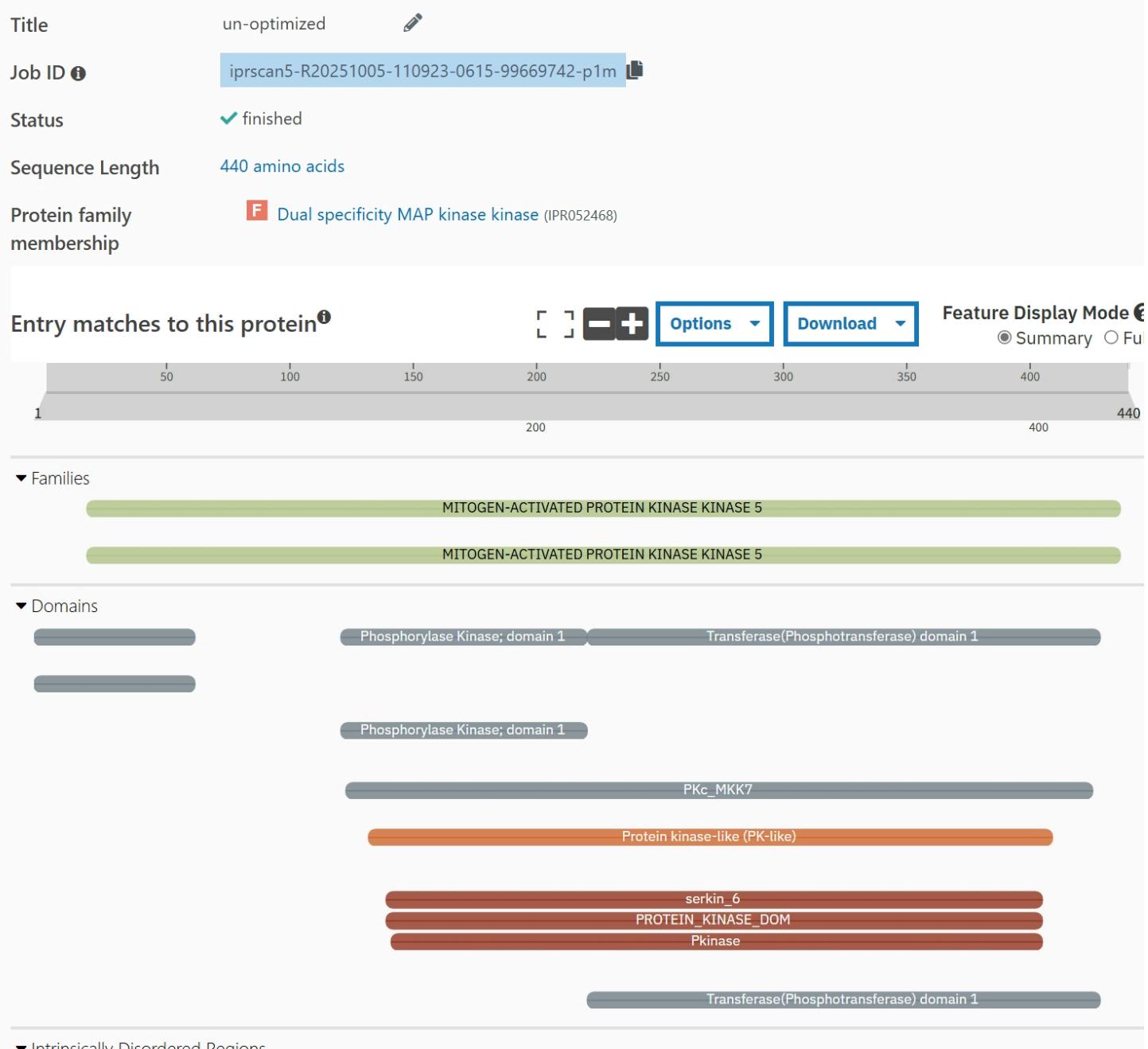
Forward vs reverse

Protein chain expressed by the forward nucleic acid segment matches 0 hit, while the reversed expressed protein matches with 82 hits.

1 - 2 of 2 results		
Sequence	Matches	Sequence Length
forward	0	436
reverse	82	440

Codon optimized vs un-optimized

1 - 2 of 2 results		
Sequence	Matches	Sequence Length
optimized	83	419
un-optimized	82	440



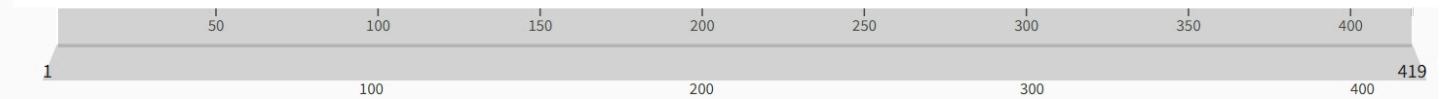
Title	optimized
Job ID	iprscan5-R20251005-110923-0615-99669742-p1m
Status	finished
Sequence Length	419 amino acids
Protein family membership	Dual specificity MAP kinase kinase (IPR052468)

Entry matches to this protein

Options Download

Feature Display Mode

Summary Full



Families

MITOGEN-ACTIVATED PROTEIN KINASE KINASE 5



Domains

Phosphorylase Kinase; domain 1

Transferase(Phosphotransferase) domain 1

Phosphorylase Kinase; domain 1

PKc_MKK7

Protein kinase-like (PK-like)

PROTEIN_KINASE_DOM

serkin_6

Pkinase

Transferase(Phosphotransferase) domain 1

The reverse protein sequence with a 6-His Tag

Title inserted sequence + 6his tag 

Job ID  iprscan5-R20251005-111101-0199-94924529-p1m 

Status  finished

Sequence Length 440 amino acids

Protein family membership  Dual specificity MAP kinase kinase (IPR052468)

Entry matches to this protein     Options Download Feature Display Mode  

Feature Display Mode  Summary  Full sequence

1 50 100 150 200 250 300 350 400 440

▼ Families

MITOGEN-ACTIVATED PROTEIN KINASE KINASE 5

MITOGEN-ACTIVATED PROTEIN KINASE KINASE 5

▼ Domains

Phosphorylase Kinase; domain 1 Transferase(Phosphotransferase) domain 1

Phosphorylase Kinase; domain 1

PKc_MKK7

Protein kinase-like (PK-like)

serkin_6
PROTEIN_KINASE_DOM
Pkinase

Transferase(Phosphotransferase) domain 1

Zooming in, we have a 6-His tag on the N-terminal indeed.

