参赛学生姓名：孔繁淏

中学：北京师范大学附属实验中学

省份：北京市

国家/地区：中国

指导老师姓名：张文彬

指导老师单位：北京大学化学与分子工程学院

指导老师姓名：孔婀静

指导老师单位：北京师范大学附属实验中学

论文题目：通过定向进化提高 GFP 索烃胞内合成效率

Improving intracellular synthesis efficiency of GFP catenane through directed evolution

Researcher: Fanhao Kong

Experimental High School attached to Beijing Normal University

Tutor: Professor Wenbin Zhang

College of Chemistry and Molecular Engineering , Peking University

Tutor: Ejing Kong

Experimental High School attached to Beijing Normal University

Abstract

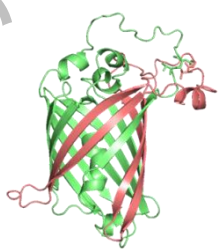
After millions of years of evolution, life has chosen proteins as the ultimate functional molecules, with their rich functionality far beyond our imagination. However, these tools originating from nature are far from perfect and often require topological modifications to construct proteins with better properties to meet certain special needs.

Green fluorescent protein (GFP) is an important protein in molecular biology. The research team led by Zhang Wenbin from Peking University has successfully completed the catenane modification of GFP, significantly improving its thermal recovery. However, after topological modification of proteins, certain properties are often compromised to some extent. Whether it is the traditional two-step synthesis of GFP catenane inside and outside the cell, or the one-step synthesis inside the cell, it will cause loss of its fluorescence intensity and other properties, resulting in low synthesis efficiency.

To recover this loss, this study innovatively introduced the method of directed evolution. By simulating natural evolution mechanisms, DNA recombination can be achieved in the laboratory. By establishing a mutant library and conducting high-throughput screening, targeted gene modification is carried out, and protein molecules with excellent properties suitable for catenane transformation are selected to improve the intracellular synthesis efficiency of GFP catenane.

Firstly, collect and construct wild-type GFP sequences in nature that are similar to existing sequences, and express wild-type proteins. Secondly, the sequence was redesigned using the intracellular one-step synthesis method of GFP catenane, molecular cloning was performed using overlapping extension, and the modified catenane protein was expressed. After exploring the reaction conditions, a mutant library was constructed using a staggered extension process and the efficiency of DNA recombination was evaluated. Finally, flow cytometry was used to select mutants that meet the fluorescence intensity requirements, and further screening was conducted for thermal recovery.

The final screened protein is suitable for catenane transformation, which is beneficial for improving intracellular synthesis efficiency. These directed green fluorescent proteins have great potential in biomedicine and functional materials, and can serve as the starting point for the next round of directed evolution, iteratively screening proteins with better traits to achieve the ultimate goal of directed evolution.



Keywords

Protein topological engineering; Catenane; Green Fluorescent Protein; Directed Evolution

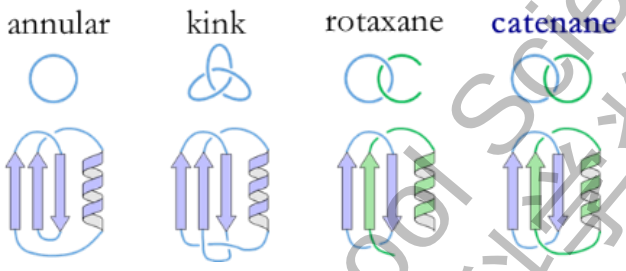
Catalogue

1. Introduction

1.1 Topological proteins

1.1.1 Natural topological proteins

Proteins are molecular machines with precise structures and rich functions, composed of amino acids connected by peptide bonds in a specific order, playing an important role in life processes. Under cellular specific translation mechanisms, most proteins synthesized within the cell have linear main chain structures. However, natural proteins with non-linear topological structures are not uncommon, such as cyclic protein[1], knot proteins[2], lasso peptides[3] and catenane[4].



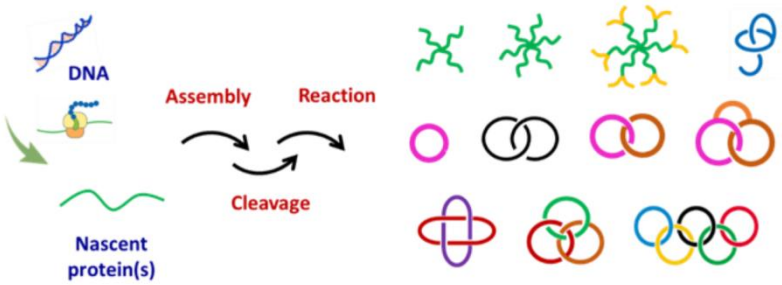
These topological proteins have many functional advantages due to their stable mechanical interlocking spatial structure, such as better thermal stability, mechanical stability, anti-denaturation stability, and anti-enzymatic stability[5]. Therefore, studying the topological structure of biomolecules in nature and synthesizing proteins with topological effects through advanced methods has become an important research topic.

1.1.2 Artificial synthesis of topological proteins

Protein engineering involves modifying proteins in nature to construct variants with better properties to meet specific needs.

There are various strategies for protein engineering, such as site-specific mutagenesis[6], cyclic rearrangement[7], domain exchange[8] and split recombination[9]. With the emergence and development of protein reaction tools such as ligase based coupling[10], peptide protein reaction pairs[11] and isolated peptides[12], protein engineering is gradually moving beyond the traditional limitations of only changing sequences and lengths, from one-dimensional single stranded linear species to multi-dimensional multi stranded topological structures, with richer modifiability and wider application prospects.

The research group led by Zhang Wenbin from Peking University is committed to using assembly reaction synergy, utilizing genetically encoded assembly and reaction elements, to spontaneously fold, assemble, shear, connect and other processes in newly formed linear proteins in vivo, forming complex topological structures of protein, and achieving performance modification and improvement.

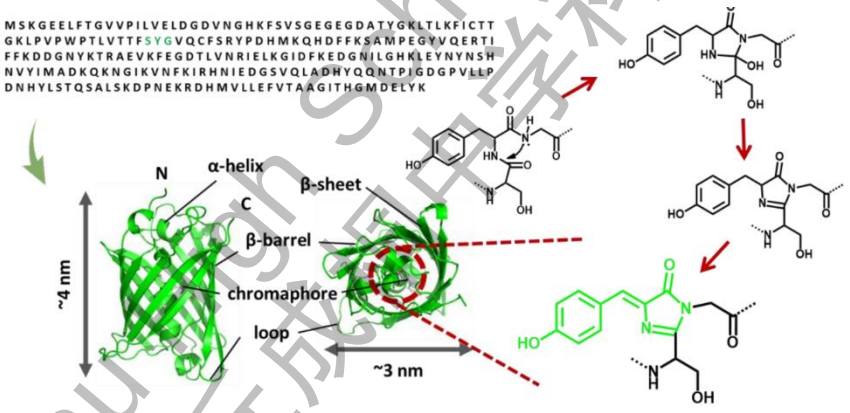


1.2 Green Fluorescent Protein (GFP) catenane

1.2.1 Green Fluorescent Protein

Green fluorescent protein is a protein composed of approximately 238 amino acids, which can be excited by both blue and ultraviolet light, emitting green fluorescence. This protein was first discovered by Xiu Shimamura and others in 1962 in Victoria's multi tube luminescent jellyfish[13].

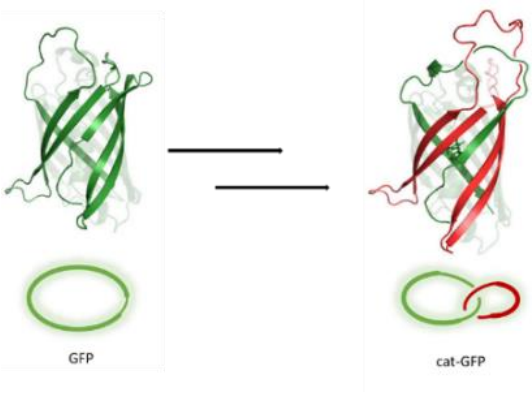
Crystal structure analysis shows that the peptide chain of green fluorescent protein folds to form a β-barrel. The structure of the bucket consists of three adjacent amino acids, serine tyrosine glycine (SYG), which, due to their close proximity in space, undergo cyclization reactions and form a conjugated chromophore through dehydration and oxidation. The chromophore is located in a rigid hydrophobic environment, isolated from environmental water molecules, so its fluorescence will not be quenched, maintaining good fluorescence quantum efficiency.



In molecular biology, green fluorescent protein is commonly used as a reporter gene. The green fluorescent protein gene can also be cloned into vertebrates for expression and used to demonstrate a certain hypothesis in experimental methods. Through genetic engineering technology, its genes can be transferred into the genomes of different species and continuously expressed in offspring.

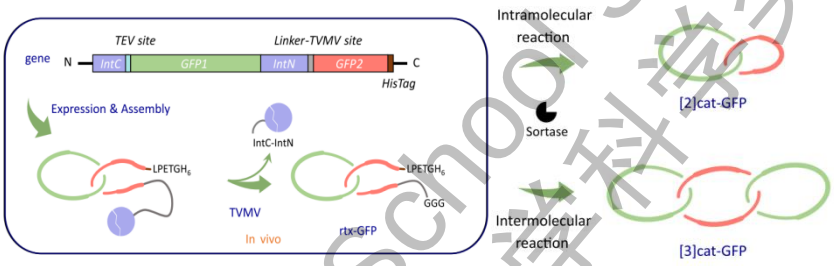
1.2.2 Green Fluorescent Protein catenane

In 2023, the research group led by Zhang Wenbin from Peking University successfully completed the catenane construction of GFP by introducing an intrinsic peptide[14]. Compared with ordinary green fluorescent proteins, GFP catenane exhibit good thermal stability and thermal recovery due to their interlocking configuration limitations.

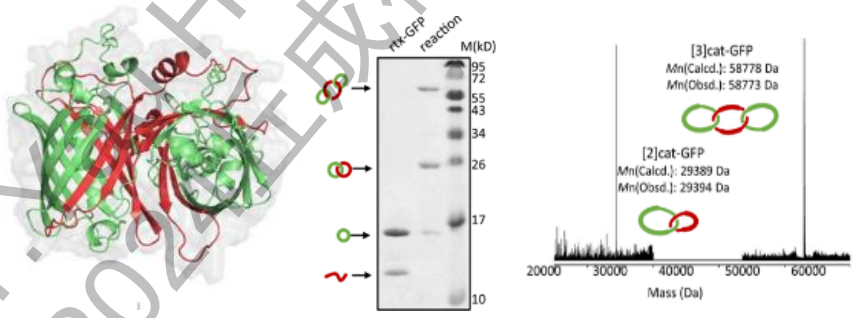


1.2.3 Two step synthesis inside and outside the cell

When synthesizing GFP catenane in two steps, the first step involves the closure of intracellular GFP ring 1. The gene sequence is designed as IntC-GFP1-IntN-Linker-GFP2, and the two segments of the peptide recognize each other in the cell, forming a rotane intermediate. Afterwards, the closure of GFP loop 2 is completed using transpeptidase outside the cell.



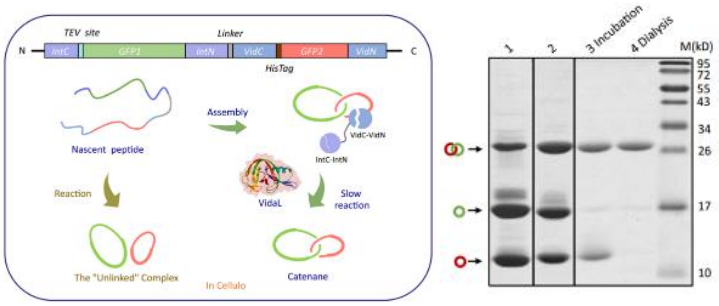
However, when performing transpeptidase cyclization, it is easy to form intermolecular cyclization products and other polymers formed by GFP ring 2 passing through two rings 1; And the two-step synthesis process is longer, which reduces the yield of [2] catenane.



1.2.4 Intracellular one-step synthesis

To shorten the synthesis steps and avoid the generation of the above-mentioned by-products, the intracellular one-step synthesis method is adopted. Design two sets of mutually recognized intropeptides in the gene sequence, IntC-GFP1-IntN-VidC-GFP2-VidN, which can directly synthesize GFP catenane in the cell.

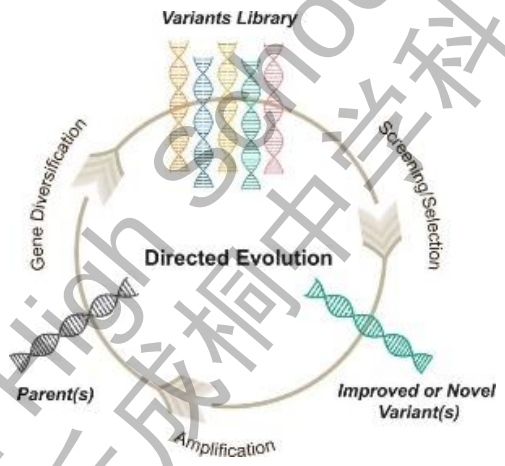
However, the experiment found that the yield of generating GFP catenane was low, and the SDS-PAGE showed that most of the products were single ring product. This indicates that the one-step synthesis of GFP in cells has a lower efficiency in catenane synthesis, and the topological modification of proteins to some extent reduces the probability of their correct assembly.



1.3 Directed evolution of proteins in vitro

1.3.1 Directed evolution

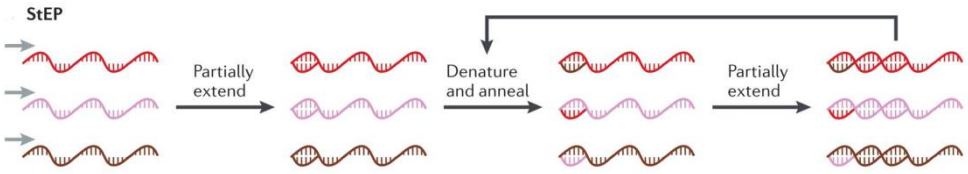
Evolution is constantly taking place in nature. Directed evolution simulates natural evolution mechanisms to achieve random mutations, DNA recombination, and screening in the laboratory, thereby modifying genes and selecting protein molecules with desired properties. The two most crucial steps in directed evolution are creating mutant libraries and developing appropriate screening methods[15]. Under appropriate screening criteria, directed evolution can accumulate beneficial mutations and potentially produce novel functions.



The main methods of directed evolution include error prone PCR[16], site-specific saturation mutation[17], DNA shuffling[18], and so on. Due to the difficulty in controlling mutation frequency through error prone PCR, and the difficulty in selecting suitable mutation sites for fixed-point saturation mutations, this study used a staggered extension process improved by DNA shuffling.

1.3.2 Staggered Extension (StEP) process

In vitro DNA recombination is based on PCR, which exchanges sequences between multiple homologous templates, which is beneficial for accumulating beneficial mutations without affecting the original function of proteins, thus completing directed evolution[19]. On this basis, the interleaved extension process is improved by combining primers and amplification products with different template sequences in a cycle of denaturation, rapid annealing, and extension. Based on complementary pairing of sequences, further extension is achieved until full-length genes are obtained. After obtaining the mutant library, high-throughput screening is carried out, using mutants with excellent properties as the starting point for a new round of cross extension. After multiple rounds of library construction and screening, proteins can be effectively modified.



1.4 Proposal of the project

In summary, the catenane transformation of green fluorescent protein not only improves thermal recovery but also causes loss of properties, resulting in low synthesis efficiency. And directed evolution happens to be an excellent solution to compensate for this loss. This study aims to obtain protein molecules with excellent properties suitable for catenane transformation through extensive library construction and screening, demonstrating the effectiveness of directed evolution methods and improving the intracellular synthesis efficiency of GFP catenane.

The research includes the following parts:

1. Collect and construct wild-type GFP sequences that are similar to existing sequences.
2. Design the sequence of catenane protein and express the wild-type and catenane protein.
3. Use the staggered extension process to construct a mutant library and evaluate the efficiency of DNA recombination.
4. Screen the fluorescence intensity and thermal recovery, and test the synthesis efficiency.

2. Experiment Methods

2.1 Construction of recombinant plasmids

Centrifuge the ordered primers using a small centrifuge, let the DNA reach the bottom of the centrifuge tube, and add ddH_2O to dilute to 10 μ M solution. Add 1 to the PCR tube μ L template, 2 forward and 2 reverse primers each μ L. 2x Pfu MasterMix 20 μ L. And supplement the total volume with ddH_2O to 40 μ L. Set the annealing temperature based on the melting temperature Tm value of the primer template binding part, and set an appropriate extension time according to the length of the target fragment. Usually, the denaturation annealing extension cycle takes about 30 rounds. Use a dual slot gradient gene amplification instrument for amplification.

When splicing fragments from multiple different sources, overlap extension PCR is used. Overlap extension PCR is a technique that uses primers with complementary ends to form overlapping chains in PCR products, and then, in subsequent amplification reactions, overlaps and splices amplification fragments from different sources through the extension of the overlapping chains. The fragments of the first round of conventional PCR have overlapping and complementary parts at the end. The second round uses multiple products from the first round of PCR as templates, with 1 template for each μ L. Add 2 forward and 2 reverse primers to the target gene fragment μ L. 2x Pfu MasterMix 20 μ L. And supplement the total volume with ddH_2O to 40 μ L. Set the annealing temperature based on the melting temperature Tm value of the primer template binding part, and set an appropriate extension time according to the length of the target fragment.

After PCR, use agarose gel electrophoresis experiment to separate and confirm the reaction system,

and cut the band that meets the expected molecular weight. Add 300 to the cut gel containing DNA μ L GDP, 55 ° C metal bath for 15 minutes, during which it is reversed several times to completely dissolve. Transfer to the column, centrifuge at 13000 r for 30 s, discard the waste liquid and add 300 more μ L GDP. Add 600 μ Wash the DW2 of L twice and idle it again. Add 15 to 20 μ Wash DNA with ddH_2O of L, let it stand for 1 minute, centrifuge at 13000 r for 1 minute. Measure concentration and purity using an ultra micro spectrophotometer.

Inoculate Escherichia coli containing the target plasmid vector into 5 mL of 2xYT medium containing corresponding resistance at a ratio of 1:100, and incubate in a full temperature shaking incubator at 37 ° C and 220 rpm for 12-16 hours. Collect the bacterial cells, transfer 5 mL of culture medium into a centrifuge tube, centrifuge at 13000 r for 1 minute, discard any unnecessary culture medium, and use a small plasmid DNA extraction kit to extract plasmids.

Take purified target genes and vectors 10 μ L (approximately 1) μ g) 1 restriction endonuclease 1 and 1 restriction endonuclease 2 each μ L. 10x FastDigest buffer 4 μ L and ddH_2O 24 μ L. 37 ° C metal bath for 1 hour. During purification, add 300 μ Mix L GDP well, transfer to a column, centrifuge at 13000 r for 30 s, and discard the waste liquid. Add another 300 μ L GDP repeat operation. Add 600 μ Wash the DW2 of L twice and idle it again. Add 15 to 20 μ Wash DNA with ddH_2O of L and centrifuge at 13000 r for 30 s. Measure concentration and purity using an ultra micro spectrophotometer.

Mix the vector and fragment after dual enzyme digestion in a molar ratio of 1:4 to 1:10, and add 1 μ L T4 ligase and 2 μ L ligase buffer and add volume to 20 with ddH_2O μ L. 22 ° C metal bath for 2 hours.

Add 20 μ The connecting system obtained from L molecular cloning is converted to 50 μ Incubate the top 10 receptive cells of L on ice for 10 minutes, apply onto an agar plate containing corresponding resistance (shake out scratches with 3 glass beads to indicate uniform application), and incubate overnight at 37 ° C in an electric constant temperature incubator.

Pick out 5 large and round colonies on the plate, and add 500 to each colony μ L medium, shake for 8 hours, and send to the sequencing company for sequencing. Further expand the culture after obtaining the recombinant plasmid with the correct sequence.

2.2 Protein expression and purification

Expand the culture to an OD600 of 0.6~0.8 in the bacterial solution, add IPTG and express at 16 ° C for 12~20 hours, then transfer to a storage bottle. Separate the culture medium and bacterial cells using a ground centrifuge at 5000 r for 20 minutes. Discard the waste liquid of the culture medium, and add 30 mL Lysis Buffer to the bacterial body for suction and mixing. Transfer to a 50 mL centrifuge tube and place it in an ice bath. Use an ultrasonic cell lysis device to perform ultrasonic lysis on the bacterial cells (ultrasonic treatment for 5 seconds with an interval of 8 seconds, total working time of 30 minutes). Centrifuge the cracking mixture at 10000 r for 30 minutes.

Mix the centrifuged supernatant with Ni NTA resin and incubate at 4 ° C for 30 minutes to 1 hour. Pour the mixture of protein and resin into the washed column, and add 10 times the volume of Wash Buffer (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole, pH=8.0) to wash out the impurities. Add 2 times the volume of Elution Buffer (50 mM NaH_2PO_4, 300 mM NaCl, 250 mM imidazole, pH=8.0) to wash out the target protein with HisTag. Recycle Ni NTA using ethanol.

2.3 Electrophoretic characterization of proteins

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) is a common electrophoresis technique that uses polyacrylamide gel as a support medium to separate proteins and oligonucleotides. The small vertical electrophoresis tank and gel were used for vertical electrophoresis, and the buffer solution was Tris glycine system. Electrophoresis at a voltage of 90 V, concentrated and changed to 140 V for approximately 1 hour. After electrophoresis, stain with Coomassie Brilliant Blue dye for about 30 minutes, and if necessary, microwave heat for 1 minute. After dyeing, decolorize with a decolorizing solution (50% ddH_2O, 40% methanol, 10% acetic acid). Once the protein band appears, imaging can be performed and the position of the band can be observed.

2.4 Electrical conversion

Electrotransformation, also known as high voltage electroporation, can introduce nucleotides, DNA, RNA, proteins, sugars, dyes, and viral particles into prokaryotic and eukaryotic cells. The high-intensity electric field instantly increases the permeability of the cell membrane, thereby absorbing foreign molecules in the surrounding medium. Electroconversion is a valuable and effective alternative method compared to other physical and chemical conversion methods.

The electric shock cup is pre cooled on ice, and the frozen receptive cells are taken out and melted on ice. one hundred μ Add 10 to L receptive cells μ Purified connecting solution, incubate on ice for 5 minutes. Add the mixture to a pre cooled electric shock cup and ice bath for 10 minutes.

When electric shock occurs, push the electric shock cup into the electric conversion device, press the pulse button, and hear a buzzing sound. Electric shock conditions are 2.5 kV, 200 ohms, 25 μ F. The time constant is approximately 6 ms. Quickly add 1 mL of culture medium insulated at 37 ° C, aspirate the mixture, and transfer it to a 1.5 mL centrifuge tube. Shake and incubate at 37 ° C and 220 rpm for 1 hour to fully express resistance. Dilute the bacterial solution 101-105 times and take 100 μ Apply L onto the plate and incubate overnight at 37 ° C. Clean the electric shock cup after electrical conversion.

2.5 Flow cytometry sorting

When the single-cell suspension stained or labeled with fluorescence is placed in the sample tube, it is pressurized into the flow chamber under high pressure. The flow chamber is filled with sheath fluid. Under the wrapping and pushing of the sheath fluid, cells are arranged in a single row and sprayed out from the nozzle of the flow chamber at a certain speed. Equipped with an ultra-high frequency piezoelectric crystal at the nozzle of the flow chamber, it vibrates after charging, causing the sprayed liquid flow to break into uniform droplets, and the test cells are dispersed among these droplets. Charge these droplets with different positive and negative charges. When the droplet flows through a deflection plate with several thousand volts, it deflects under the action of a high-voltage electric field and falls into its respective collection containers. The uncharged droplets fall into the middle waste liquid container, thus achieving cell separation.

series of optical systems in the instrument can collect signals such as fluorescence, light scattering, light absorption, or cell impedance. Computer systems collect, store, display, and analyze various measured signals, and perform statistical analysis on various indicators.

2.6 ELISA reader screening

An enzyme-linked immunosorbent assay (ELISA) reader, also known as a microplate detector, is a variable phase photoelectric colorimeter or spectrophotometer. The light waves emitted by the light source lamp are transformed into a single color beam through a filter or monochromator, and enter a 96 well plate (a transparent plastic plate used to place the test sample, with uniformly sized small holes on the plate, each of which can hold hundreds of microliters of solution) for the test sample. Part of the monochromatic light is absorbed by the specimen, while the other part is irradiated onto the photodetector through the specimen. The photodetector converts the light signals of different strengths and weaknesses of the specimen into corresponding electrical signals. The electrical signals are processed by pre amplification, logarithmic amplification, analog-to-digital conversion, etc., and then sent to a microprocessor for data processing and calculation. Finally, the results are displayed on a display and printer.

The enzyme-linked immunosorbent assay (ELISA) reader can be divided into grating ELISA reader and filter ELISA reader in principle. The grating type ELISA reader can capture any wavelength within the wavelength range of the light source, while the filter type ELISA reader can only capture specific wavelengths for detection based on the selected filter.

3. Directed evolution of GFP catenane

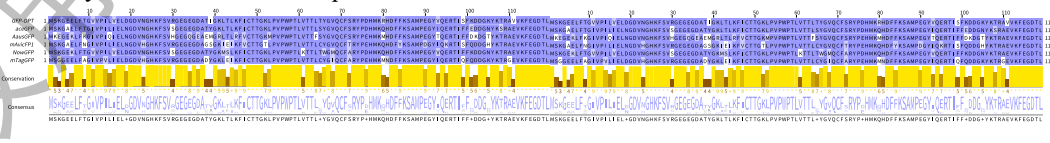
3.1 Collection and Construction of Similar GFP Sequences

3.1.1 Protein sequence collection

In order to use DMA shuffling for efficient library construction, it is necessary to search for multiple protein sequences that have a consistency of about 80% with the template sequence. Using the GFP-OPT sequence as a template, 5 GFP sequences were screened from FPbase, and the excess amino acids at the N-terminus were removed to ensure a length of 238 aa. The protein information and multi sequence alignment results are as follows:

Name	Identity	Reference
AceGFP	86.1%	J. Biol. Chem. 2010, 285, 15978−15984
AausGFP	78.6%	PLoS Biol. 2020, 18, e3000936
mAvicFP1	78.2%	
NowGFP	87.8%	Biophys. J. 2015, 109, 380−389
mTagGFP	79.0%	Chem. Biol. 2008, 15, 1116−1124

The consistency between the above 5 sequences and GFP-OPT ranges from 78% to 88%, and the consistency between each two sequences does not exceed 88%.



3.1.2 DNA sequence construction

Write Python scripts to convert protein sequences into DNA sequences. In order to switch between

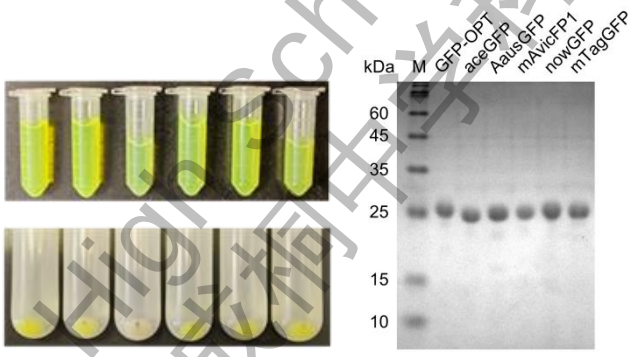
templates at a higher frequency during shuffling, the template DNA sequence difference should be minimized. Conservative sites directly replicate template codons, while mutation sites minimize the number of base mutations. If there are multiple possibilities, choose the codon preferred by Escherichia coli. Python scripts can provide mutation sites, possible mutations, codons before and after mutations, and complete DNA sequences.

The generated mutation sequence may contain common cleavage sites, all of which are marked and manually replaced in ApE, following the rules of least mutation and E. coli preference.

The mutation rate of DNA bases ranges from 5% to 10%, which is half lower than the protein sequence mutation rate of 12% to 21%, because most residues mutate 1-2 bases. Order the aforementioned genes from a gene synthesis company and use the same pET-22b vector for protein expression.

3.2 Expression of wild-type and GFP catenane

Transform the plasmid into BL21 competent cells and culture in 200 mL of medium. Induce with 0.5 mM IPTG and express at 16 ° C for 16 hours. Soluble proteins and precipitates after centrifugation are shown in the following figure. From left to right are GFP-OTP, aceGFP, AausGFP, mAvicFP1, nowGFP, and mTagGFP. All GFPs produce green fluorescence, with only AausGFP showing weak fluorescence, consistent with the literature.

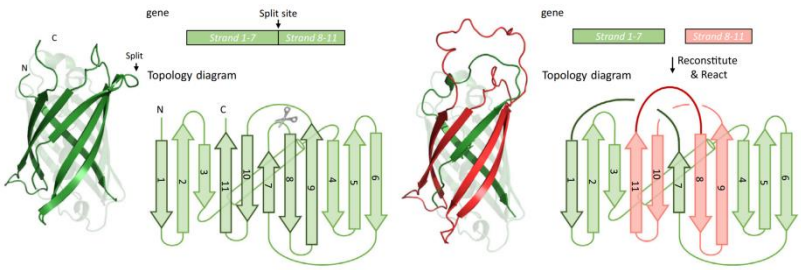


The SDS-PAGE characterization results are shown in the following figure. The molecular weight of six linear GFPs is about 28 kDa, with the main band at 25 kDa, slightly smaller than the actual molecular weight; All proteins can obtain a single main band, indicating that linear GFP is well expressed in Escherichia coli.

3.3 Expression of GFP derived from different sources

3.3.1 Sequence design

Construct rings 1 and 2 of cat-GFP-OPT, cat-aceGFP, cat-AausGFP, cat-mAvicFP1, cat-nowGFP, and cat-mTagGFP using the same splitting method as cat-GFP (between 156 and 157).



3.3.2 PCR and Overlap PCR

Design 5 primers for ring 1 and ring 2 of the catenane protein and primers for Npu-N-VidaL-C as follows. Using the wild-type as the template, PCR to obtain ring 1 and ring 2. Use PCR to obtain Npu-N-VidaL-C using cat GFP as the template.

By overlapping extension PCR and using some previous primers, the three gene segments were assembled into complete fragments using GFP 1, Npu N-VitaL-C, and GFP 2 as templates, and connected to the pQE-80L vector. This method only requires one sequencing, greatly reducing the experimental period.

BamHI-AausGFP-F	ATTAGGATCCATGAAAGAAGGAGAAAAACTTTTC
BamHI-ace-mAvicGFP-F	ATTAGGATCCATGAGCAAAGGAGCAGAA
BamHI-nowGFP-F	ATTAGGATCCATGAGCAAAGGAGAAAAACTT
BamHI-mTagGFP-F	ATTAGGATCCATGAGCGGAGGAGAAGAA
AausGFP-Npu-N-R	ATACGACAGGCAGGTACCTTTGTCTGACATGATGTATACATT
aceGFP-Npu-N-R	ATACGACAGGCAGGTACCTTTGTCTGTCATGATGTATACATT
mAvicGFP-Npu-N-R	ATACGACAGGCAGGTACCTTTGTCTGACAGGACGTATAC
nowGFP-Npu-N-R	ATACGACAGGCAGGTACCTTTGTCTGCCGTGATGTT
mTagGFP-Npu-N-R	ATACGACAGGCAGGTACCTTTGTCTGGCCTGATGTATAC
HisTag-Aaus-aceGFP-F	CATCATCATCATCATGAGCTCGCAAAGAATGGAATCAAAGTT
HisTag-mAvicFP1-F	CATCATCATCATCATGAGCTCGCAAACAATGGAATCAAAGTT
HisTag-nowGFP-F	CATCATCATCATCATGAGCTCCAAAAGAATGGAATCAAAGCG
HisTag-mTagGFP-F	CATCATCATCATCATGAGCTCGCAAACAATGGACTCGAA
Aaus-mAvicGFP-SpeI-R	TAATACTAGTACCTCCGCCGTGAGTGATCCC
aceGFP-SpeI-R	TAATACTAGTACCTCCGCCGTGAGTGATCGC
nowGFP-SpeI-R	TAATACTAGTACCTCCGCCGAGAGGAATCCC
mTagGFP-SpeI-R	TAATACTAGTACCTCCGCCGTGAGTGTGGCA

KpnI-Npu-N-F	GGTACCTGCCTGTCGTAT
HisTag-SacI-R	GAGCTCATGATGATGATGATG

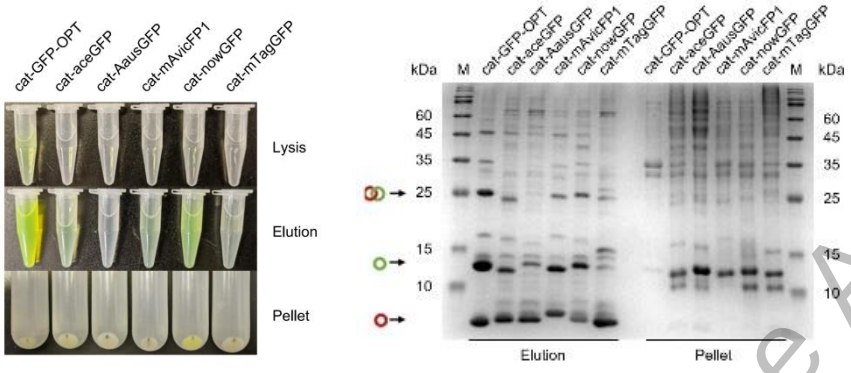
BamHI-AausGFP-F	ATTAGGATCCATGAAAGAAGGAGAAAAACTTTTC
BamHI-ace-mAvicGFP-F	ATTAGGATCCATGAGCAAAGGAGCAGAA
BamHI-nowGFP-F	ATTAGGATCCATGAGCAAAGGAGAAAAACTT
BamHI-mTagGFP-F	ATTAGGATCCATGAGCGGAGGAGAAGAA
Aaus-mAvicGFP-SpeI-R	TAATACTAGTACCTCCGCCGTGAGTGATCCC
aceGFP-SpeI-R	TAATACTAGTACCTCCGCCGTGAGTGATCGC
nowGFP-SpeI-R	TAATACTAGTACCTCCGCCGAGAGGAATCCC
mTagGFP-SpeI-R	TAATACTAGTACCTCCGCCGTGAGTGTGGCA

KpnI-Npu-N-F	GGTACCTGCCTGTCGTAT
HisTag-SacI-R	GAGCTCATGATGATGATGATG

3.3.3 Protein expression and purification

Transform the plasmid into BL21 competent cells, express at 16 ° C for 16 hours, centrifuge for collection, and sonicate for rupture. The flow through, elution solution, and precipitation after centrifugation of GFP from different sources purified by nickel columns are shown in the following figure. Among them, cat-GFP-OPT has the strongest fluorescence and the least precipitation, while cat-nowGFP has a certain fluorescence and the precipitation is also green. The other four cat GFPs have weak or no fluorescence.

The SDS-PAGE characterization is shown in the following figure. The bands near 25 kDa in the eluent are catenane bands, bands between 10-15 kDa are ring 1, and bands below 10 kDa are ring 2. Cat GFP-OPT, cat aceGFP, cat mAvicFP1, and cat nowGFP have a chain and a ring 1 band, and all proteins have a ring 2 band. The upper band at 10-15 kDa in the precipitate is ring 1, while the lower band may be ring 2 that has not been closed, indicating that only correctly folded and closed ring 2 is soluble. The above results indicate that GFP-OPT is still the optimal variant for constructing GFP catenane.



3.4 Establishment of GFP directed evolution mutant library

3.4.1 Exploration of reaction conditions for staggered extension PCR

To directly amplify the full-length fragment of GFP1-Npu-N-VidaL-C-GFP2 (1500 bp), different StEP reaction conditions were attempted using six cat GFPs as mixed templates.

No single full-length fragment or band dispersion can be obtained, while conventional PCR using the same system can yield high concentrations of target bands. Therefore, it was decided to use the segmented StEP followed by Overlap PCR method.

Length (bp)	21	24	27	Time (s)	5	10	15
MasterMix	Taq	Pfu	High-GC	Tem. (℃)	55	60	65

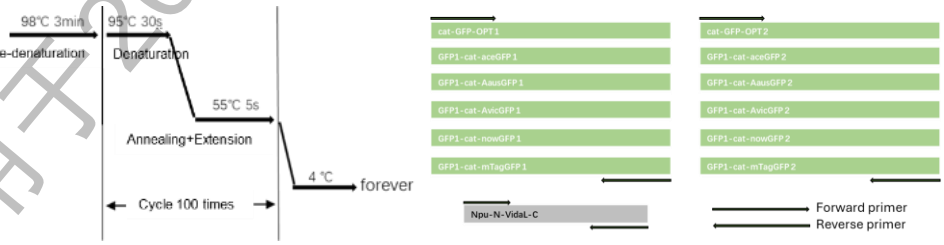
3.4.2 Cross extension PCR construction of mutant library

Replace MasterMix with Hieff Canace Plus, take 100 ng each of the 6 previously obtained cat GFP plasmids, and mix evenly. With 1 μ The L mixed plasmid is used as a template, and the primers and reaction conditions for staggered extension PCR are as follows. Recombinant GFP1 and GFP2 were amplified using staggered extension PCR, while Npu-N-VidaL-C fragments were amplified using conventional PCR.

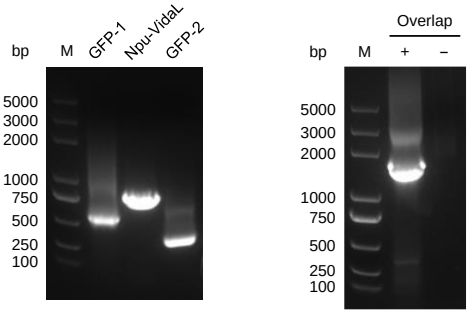
TEV-BamHI-F	AACCTGTATTTTCAGGGCGGATCC
KpnI-Npu-N-R	CTCATACGACAGGCAGGTACC
HisTag-SacI-F	CATCATCATCATCATCATGAGCTC
SpeI-VidaL-R	CGGCAGACAGCCAGATTCACTAGT

KpnI-Npu-N-F	GGTACCTGCCTGTCGTATGAG
HisTag-SacI-R	GAGCTCATGATGATGATGATGATG

Using Overlap PCR, full-length fragments were obtained using recombinant GFP ring 1, Npu-N-VidaL-C, and recombinant GFP ring 2 as templates, using the same method as in 3.2.2.



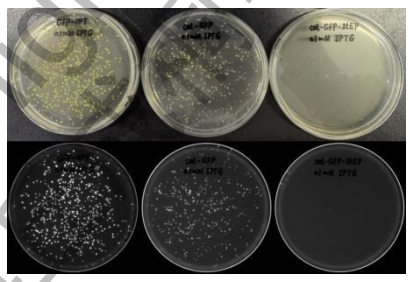
The electrophoresis results of GFP 1, Npu-N-VidaL-C, GFP 2 and overlapping extended full-length fragment agarose gel are as follows, and the expected molecular weight products are obtained.



3.4.3 Evaluation of DNA Recombination Results

The cat GFP recombinant plasmid obtained from StEP PCR was transformed into TOP10 competent cells, and 15 single colonies were sequenced, of which 9 were correctly recombined sequences. The sequence alignment after removing the intein is shown in the following figure. The mutations in the recombinant sequence all come from six template sequences, with a proportion roughly similar to that in the template. The results indicate that a DNA recombinant library has been successfully constructed using staggered extension PCR, which can be used for subsequent screening.

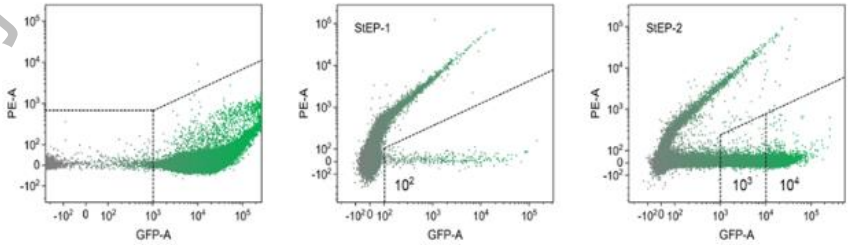
Expression of wild-type GFP-OPT, cat-GFP, and DNA recombinant cat-GFP StEP on the agar plate, as shown in the following figure. The wild-type GFP-OPT was observed to have the strongest fluorescence, followed by cat-GFP. cat-GFP-StEP was unable to directly observe fluorescence due to the small number of clones grown. Therefore, subsequent electroconversion was performed and flow cytometry was used to select proteins that produced strong fluorescence for final screening.



3.5 Screening of GFP directed evolutionary mutant library

3.5.1 Electroconversion and Fluorescence Intensity Sorting of Cat-GFP Library

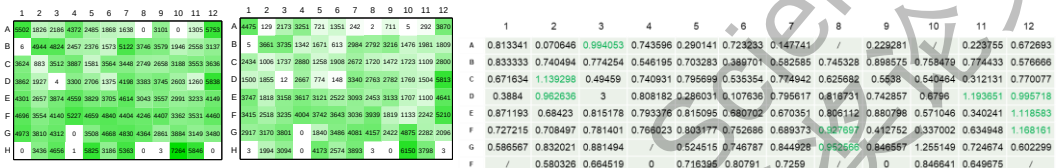
Perform electroconversion of GFP-OPT, cat-GFP, and cat-GFP-STEP. Using a flow cytometer for sorting, collect bacteria with intensities above 10^2, 10^3, and 10^4, as shown blow.



3.5.2 Heat recovery screening of cat GFP libraries

Cultivate bacteria with fluorescence intensity above 10^3 and 10^4 after flow cytometry sorting using a 96 well plate, induce with 0.5 mM IPTG, and express at 30 ° C for 8 hours. Take 50 μ Measure the fluorescence intensity of 480/510 nm on an enzyme-linked immunosorbent assay (ELISA) reader using L-bacterial lysate. Using a 96 well plate in a boiling water bath, let it cool naturally and then measure the fluorescence intensity again. The results of the two tests are shown in the left and right figures, respectively.

Among them, lines A to D correspond to 10^3 groups, lines E to H correspond to 10^4 groups, H10 is the linear GFP control, and H11 is the cat-GFP control. Divide the fluorescence intensity after water bath by the fluorescence intensity before water bath, select 10 wells with a ratio greater than 90%, namely A3, C2, D2, F8, G8, G10, D11, D12, E12, F12, and sequence the corresponding bacteria.

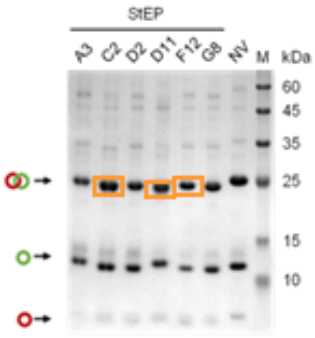


	1	2	3	4	5	6	7	8	9	10	11	12
A	0.813341	0.070646	0.994053	0.743596	0.290141	0.723233	0.147741	/	0.229281	/	0.223755	0.672693
B	0.833333	0.740494	0.774254	0.546195	0.703283	0.389701	0.582585	0.745328	0.898575	0.758479	0.774433	0.576666
C	0.671634	1.139288	0.49459	0.740931	0.795699	0.535354	0.774942	0.625682	0.5638	0.540464	0.312131	0.770077
D	0.3884	0.962636	3	0.808182	0.286031	0.107636	0.795617	0.816731	0.742857	0.6796	1.193651	0.995718
E	0.871193	0.68423	0.815178	0.793376	0.815095	0.680702	0.670351	0.806112	0.880798	0.571046	0.340241	1.118583
F	0.727215	0.708497	0.781401	0.766023	0.803177	0.752686	0.689373	0.937697	0.412752	0.337002	0.634948	1.168161
G	0.586567	0.832021	0.881494	/	0.524515	0.746787	0.844928	0.952568	0.846557	1.255149	0.724674	0.602299
F	/	0.580326	0.664519	0	0.716395	0.80791	0.7259	/	0	0.846641	0.649675	/

3.5.3 Evaluation of screening results

The screened cat-GFP variants with high heat recovery were sequenced, and the protein sequence alignment after removing the intein is as follows. Four out of ten variants (F8, G10, D12, E12) were found to be template cat-GFP-OPT sequences. Among the remaining six variants, A3, C2, and D11 had more than 10 mutations relative to cat-GFP-OPT, while D2, F12, and G8 had less than 10 mutations.

Select 6 variants that do not replicate cat-GFP-OPT, induce with 0.4 mM IPTG in 50 mL medium, and express at 30 ° C for 7 hours. The SDS-PAGE characterization is shown in the following figure. The bands near 25 kDa in the eluent are hydrocarbon bands, bands between 10-15 kDa are ring 1, and bands below 10 kDa are ring 2. cat-GFP-OPT (denoted as NV) is used as a control.



Among them, the C2, D11, and F12 mutants have a higher proportion of catenane than NV, indicating a higher intracellular synthesis efficiency. Other proteins that can improve protein thermal recovery and fluorescence intensity also have a similar proportion of hydrocarbons as NV.

A round of library building and screening resulted in mutants that improved intracellular synthesis

efficiency without affecting protein thermal recovery and fluorescence intensity. These mutants can serve as the starting point for subsequent evolution. After multiple rounds of iteration, mutants with greater improvement in intracellular synthesis efficiency can be selected.

4. Conclusion and Prospects

The complex structure formed by protein folding is like a beautiful gem, and the topological modification of proteins based on structure is like the process of meticulous carving on this gem. The unique topological structure modification can bring specific functional advantages to proteins and to some extent damage their biophysical properties: when synthesizing green fluorescent protein (GFP), it inevitably causes losses in fluorescence intensity and other properties, resulting in low synthesis efficiency. This study found a solution for it through the introduction of directed evolution methods. The specific experimental results are as follows:

1. 5 protein sequences with a consistency of about 80% with the template GFP-OPT sequence were collected from nature for PCR experiments and protein expression. The protein fluorescence intensity was similar to that described in the literature.

2. GFP was subjected to hydrocarbon transformation using the same splitting and modification method as cat-GFP-OPT, followed by segmented PCR followed by overlapping extension PCR to obtain full-length genes and express proteins. Among them, cat-GFP-OPT has the strongest fluorescence intensity, while the rest of cat-GFP have weak or no fluorescence. The universality of using this method for intracellular one-step synthesis of GFP has been demonstrated, and the issue of synthesis efficiency has been once again verified.

3. Using the previous six cat-GFP templates as templates, staggered extension PCR was performed, and then overlapped with the Npu-N-VidaL-C fragment to obtain the recombinant full-length fragment. The results showed that the mutations in the recombinant sequence were all from 6 template sequences, and a DNA recombinant library was successfully constructed. Due to the limited number of cat GFP StEP strains, fluorescence could not be directly observed, and the cells were subjected to electroconversion treatment.

4. Perform flow cytometry sorting on bacteria expressing cat GFP StEP protein. About 100000 and 30000 bacteria with fluorescence intensities above 103 and 104 were collected. Select these two groups of proteins for thermal recovery screening in a 96 well plate. Six mutant proteins with different templates showed better heat recovery, and three of them showed a higher proportion of hydrocarbon formation during expression compared to GFP-OPT, resulting in improved intracellular synthesis efficiency compared to the template.

To further improve the intracellular synthesis efficiency of GFP, further research can be conducted in the following areas in the future.

1. Using the mutants obtained from the first round of library construction and screening as the starting point for the next round of directed evolution, conduct 2-3 iterations to increase screening pressure, search for mutants with improved fluorescence intensity and thermal recovery, explore key sites that affect the intracellular synthesis efficiency of GFP hydrocarbon, and achieve precise regulation of hydrocarbon transformation.

2. Change the strategy of GFP hydrocarbon transformation, or improve experimental methods and conditions, and undergo re evolution to obtain mutants that significantly improve intracellular synthesis

efficiency after a round of library building and screening.

3. Apply the method of directed evolution to the topological modification of other proteins, enhance the specific properties of proteins, and further understand the functional advantages that topological modification brings to proteins.

Reference

[1] Cascales, L.; Craik, D. J. Naturally occurring circular proteins: Distribution, biosynthesis and evolution. *Org. Biomol. Chem.* **2010***, 8*, 5035-5047.

[2] Sułkowska, J. I.; Sułkowski, P.; Szymczak, P. et al. Stabilizing effect of knots on proteins. *Proc. Natl. Acad. Sci. U. S .A.* **2008**, *105*, 19714-19719.

[3] Maksimov, M. O.; Pan, S. J.; Link, J. A. Lasso peptides: Structure, function, biosynthesis, and engineering. *Nat. Prod. Rep.* **2012**, *29*, 996-1006.

[4] Boutz, D. R.; Cascio, D.; Whitelegge, J. et al. Discovery of a thermophilic protein complex stabilized by topologically interlinked chains. *J. Mol. Biol.* **2007**, *368*, 1332-1344.

[5] Wang, X.-W.; Zhang, W.-B. Chemical topology and complexity of protein architectures. *Trends Biochem. Sci.* **2018**, *43*, 806-817.

[6] Chiu, J.; March, P. E.; Lee, R. et al. Site-directed, Ligase-Independent Mutagenesis (SLIM): a single-tube methodology approaching 100% efficiency in 4 h. *Nucleic Acids Res.* **2004**, *32*, e174.

[7] Martinez, J. C.; Viguera, A. R.; Berisio, R. et al. Thermodynamic analysis of α-spectrin SH3 and two of its circular permutants with different loop lengths: Discerning the reasons for rapid folding in proteins. *Biochemistry* **1999**, *38,* 549-559.

[8] Koon, N.; Squire, C. J.; Baker, E. N. Crystal structure of LeuA from *Mycobacterium tuberculosis*, a key enzyme in leucine biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 8295-8300.

[9] Shekhawat, S. S.; Ghosh, I. Split-protein systems: beyond binary protein–protein interactions. *Curr. Opin. Chem. Biol.* **2011**, 15, 789-797.

[10] Chen, I.; Dorr, B. M.; Liu, D. R. A general strategy for the evolution of bond-forming enzymes using yeast display. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 11399-11404.

[11] Zakeri, B. Synthetic biology: A new tool for the trade. *ChemBioChem* **2015**, *16*, 2277-2282.

[12] Perler, F. B. InBase, the New England Biolabs Intein Database. *Nucleic Acid Res.* **1999**, *27*, 346-347.

[13] Cantrill S. Green fluorescent protein. *Nat. Chem.* **2008**, DOI: 10.1038/nchem.75.

[14] Qu, Z.; Fang, J.; Wang, Y.-X. et al. A single-domain green fluorescent protein catenane. *Nat. Commun.* **2023,** *14*, 3480.

[15] Wang, Y.; Xue, P.; Cao, M. et al. Directed evolution: Methodologies and applications. *Chem. Rev.* **2021,** *121*, 12384-12444.

[16] Lee, S. O.; Fried, S. D. An error prone PCR method for small amplicons. *Anal. Biochem.* **2021,** 628, 11426.

[17] Zhang, K.; Yin, X.; Shi, K. et al. A high-efficiency method for site-directed mutagenesis of large plasmids based on large DNA fragment amplification and recombinational ligation. *Sci. Rep.* **2021,** *11*, 10454.

[18] Zhang, K.; Yin, X.; Shi, K. et al. A high-efficiency method for site-directed mutagenesis of large plasmids based on large DNA fragment amplification and recombinational ligation. *Sci. Rep.* **2021,** *11*, 10454.

[19] Wang, Y.; Xue, P.; Cao, M. et al. Directed evolution: Methodologies and applications. *Chem. Rev.* **2021,** *121*, 12384-12444.

致谢

我于 2023 年 12 月入选中科协“英才计划”，通过笔试及面试，进入北京大学化学与分子工程学院张文彬课题组研究学习，获得张文彬教授同课题组成员的指导帮助。“英才计划”为无偿项目，旨在选拔品学兼优、学有余力的中学生走进大学，在自然科学基础学科领域的科学家指导下参加科学研究、研讨和实践，推动高校和中学联合培养基础学科拔尖人才。

该研究课题《通过定向进化提高 GFP 索烃胞内合成效率》由我阅读课题组论文及相关文献后与张文彬教授讨论形成，运用定向进化的手段使 GFP 索烃突变体胞内合成效率相较模板有所提升，为蛋白质拓扑改造给蛋白质性质带来的损失找到了合理的解决方案。该研究项目在高分子拓扑改造及蛋白质工程领域有积极影响。

自 2023 年 1 月进入实验室至 2024 年 2 月，我提出定向进化的研究想法，完成了为其一年多的研究，从实验设计，分子克隆的工艺选择，实验操作，结论分析和 20 页英文论文产出及修改完善。在研究中，我积极创新，如创造性地将定向进化运用在蛋白拓扑改造上，在遇到扩增全长重组蛋白不成功的问题时，将方法改为分段交错延伸 PCR 后重叠延伸 PCR 等。

张文彬教授在项目中为我提供了关于蛋白拓扑工程和定向进化的理论指导、选题指导和实验指导，监督研究全程的顺利开展。学校老师孔婀静提供了分析数据指导和学术论文撰写指导，帮助适应大学科研环境及项目结果到实际产出的转化。实验过程中，部分蛋白表达接受了来自课题组蒋冯逸师兄的指导和帮助。

经历“英才计划”开题报告、中期答辩和终期答辩，该项目获中国科学技术大学包信和等院士认可。我被评为“英才计划年度优秀学生”并入选中科协 2024 国际科技交流项目冬令营，4 月赴日本进行“樱花计划”科研国际交流。

在此感谢张文彬教授、孔婀静老师和课题组师兄师姐在研究过程中提供的一切帮助，让我在高中走进梦寐以求的学府，让我在热爱的化学领域发展学术兴趣，开始我的探索之旅。

张文彬 博士

教授，博士生导师
研究方向：高分子科学，蛋白质工程，生物材料
电话：+86-10-62766876
邮箱：wenbin@pku.edu.cn
课题组网站：http://www.chem.pku.edu.cn/zhangwb/
简历:
2013.08-至今，特聘研究员，博士生导师，北京大学化学与分子工程学院
2013.08-至今，PI，北京大学软物质科学与工程中心
2011.09-2013.08，博士后，美国加州理工学院（Prof. David A. Tirrell）
2010.06-2011.08，博士后，美国阿克伦大学高分子科学系（Prof. Stephen Z. D. Cheng）
2006.01-2010.05，理学博士，美国阿克伦大学高分子科学系（Prof. Stephen Z. D. Cheng and Prof. Roderic P. Quirk）
2000.09-2004.06，理学学士，北京大学化学与分子工程学院

孔婀静

北京师范大学附属实验中学化学教师，国际部化学学科组组长
电话：+86 13661058326
邮箱：ejing.kong@sdszintl.com
简历：
2004.09-至今，北京师范大学附属实验中学化学教师，国际部化学学科组组长
2001.09-2004.07 理学硕士，北京师范大学化学系有机化学专业
1997.09-2001.07 理学学士，北京师范大学化学系

孔繁淏

电话：+86 15011553279
邮箱：rickykfh@sina.com
简历：
2022.09-至今 北京师范大学附属实验中学
2024.06-2024.08 暑期科学项目（SSP 生化，美国加州理工学院）
2019.09-2022.06 北京市第二中学分校
个人荣誉：
2023 英才计划年度优秀学生
中科协 2024 国际科技交流项目冬令营
中国化学奥林匹克竞赛，二等奖
加拿大化学奥林匹克竞赛，国家集训营
英国化学奥林匹克竞赛，金奖