How to efficiently connect by using Golden Gate?

Characterize and debug TAL effectors parts
designed by 2012 Freiburg

Why we want to improve it?

Whether the Freiburg's design is efficient or not

According to the experimental record of Freiburg, the success rate is higher than 95%(32/33). However, this result apparently lacks statistical significance.

In the result section, they emphasize that there is a light band at 1200bp, which they believe could indicate that the Golden Gate connection works well. However, after conducting several experiments by ourselves, we find that the key point to indicate whether Golden Gate connection works is not the band at 1200bp. If the band is not clear and specific in the gel, it indicates the experiment doesn't go well. We can easily find several light bands under the band of 1200bp. Moreover, the second light band is somewhat lighter than the band at 1200bp. Although the Freiburg can explain the results with the repeatability of the TALE sequence, we suppose that the possibility of the mismatch of the sticky ends still can't be excluded. Frankly speaking, we try to believe that they really made it, but if the success cannot be repeated, there must be something wrong with their system.

The protocol we take to connect the parts of TALE

1. Freiburg's protocol
2. Restriction enzyme digestion of plasmid and TAL repeats and gel extraction respectively. By the mole ratio, plasmid to TALE is 1 to 5 and TALE to TALE is 1 to 1. Ligation with T4 ligase in 22 ℃ over night.
3. The same ratio of plasmid and TALE repeat, but add the TALE repeats one by one and ligation in 22 ℃, 30 minutes
4. Every two parts connect at one time, and try to make three intermediates of 400bp, and then mix the plasmid to make the complete TALE.
5. The same ratio and ligation with the program of 22℃ 2min, 40℃ 30,25 repeats.

The motivation to debug 2012 Freiburg's parts

Unfortunately, all of our attempts failed. We didn't manage to make a complete TALE, or even make two of them together. However, what is important for us is that when we try the 5th protocol, we notice an unexpected result. When we analyze the sequence result, we find that our left adaptor, 1st part and right adaptor connect together. Why do we get this result? We notice that their sticky end is TGAC, GCTC, and ACTC. That is to say, GCTC and ACTC connect with each other by mistake. In another word, if the sticky ends are very similar, they probably connect with each other. Although we failed again, the result gives us confidence to debug 2012 Freiburg's parts.

How do we connect certain monomer?

Some advanced tips for TALE protein

1) Given a sample sequence with repeating amino acids:

LTPEQVVAIAS(XX)GGKQALETVQRLLPVLCQAHG

XX= NG→T
HD→C
NI→A
NN→G or A (NN & NK → G)

What XX means is that it determine the certain kind of base. For one unit of repetition, other amino acids can be identical.

2) A fully functional TALE protein contains one sequence, that does not have repetitive units, recognizing base T, and similar sequence but is only half length as its end. That is, one complete TALE protein is able to recognize certain number of repetitive units and two bases.
3) The length that can be recognized is not strictly twelve or fourteen. According to the published results, the length and certain sequence are dependent on number and type of monomer.

We can gather 96 bioparts based on Freiburg, and each part has its counterproductive base on certain location(1,2,3,4,5 or 6). By picking two bases on certain location, we are able to design one TALE protein sequence.

Previous Review: Freiburg's way of connection

The main principles of connection is built upon the idea of Golden Gate Connection.(Sanjana, N. E. et al. A transcription activator-like effector toolbox for genome engineering. Nature Protocols 7, 171–192 (2012).)

The gust of these procedures is more related to one type of restriction enzyme, type II Restriction Enzyme, especially BsmBI enzyme.

5'...C G T C T C (N)1↓...3'
3'...G C A G A G (N)5↑...5'

The main feature of this enzyme is the recognition sequence is on only one side of cleavage site. It provides the way which can be used to get certain incision without damaging the whole sequence. The sticky end has 4bp base, and it could be designed even for combination of multiple sticky end. That feature is fancy at first, but we cannot regardless its latent shortcomings.

Let's analyze the example(AA1) provided by Freiburg.

CGTCTCA|5'-TGACCCCGGAACAGGTGGTGGCCATCGCCTCCAACATTGGTGGTAAGCAAG
CCCTCGAAACTGTGCAGCGGCTGCTTCCAGTCTTGTGCCAGGCTCACGGCCTGACACCG
GAGCAGGTGGTTGCAATCGCGTCTAATATCGGCGGCAAACAGGCATTGGAGACCGTGCA
GCGCTTGCTTCCAGTGCTGTGTCAGGCCCACGG|GCTCTGAGACG

The underlined parts are recognized by BsmBI. Vertical bar(|) is the cutting position. As for this sample, TGAC is one sticky end which can combine with other seven sticky ends.

Evaluate seven sticky ends designed by 2012 Freiburg

2012 Freiburg's parts have seven sticky ends:

TGAC,GCTC,CTTG,GCTT,ACTG,CCTG,ACTC

We all know that certain two parts can combine together, under base-pair rule. However, whether it is possible that unpaired sticky ends can bind together? In fact, the more similar they are, the more possibility that can form new but error base pairs.

Spired by BLAST algorithm, we calculate the similarity of each other sticky ends.

	TGAC	GCTC	CTTG	GCTT	ACTG	CCTG	ACTC
TGAC	\	2	2	2	2	2	2
GCTC	2	\	2	3	2	2	3
CTTG	2	2	\	3	3	3	2
GCTT	2	3	3	\	2	2	2
ACTG	2	2	3	2	\	3	3
CCTG	2	2	3	2	3	\	2
ACTC	2	3	2	2	3	2	\

The higher score, the higher similarity, and the higher possibility of mismatch.

The table shows that more than 30% of pairs' score is equal to 3, which means that the possibility of mismatch cannot be neglected.

Even if we employ the relatively loose rule to calculate the similarity, we can still find that error rates cannot be neglected.

	TGAC	GCTC	CTTG	GCTT	ACTG	CCTG	ACTC
TGAC	\	1	2	1	2	2	2
GCTC	1	\	2	3	2	2	3
CTTG	2	2	\	3	2	3	2
GCTT	1	3	3	\	2	2	2
ACTG	2	2	2	2	\	3	3
CCTG	2	2	3	2	3	\	2
ACTC	2	3	2	2	3	2	\

Why did Freiburg insist to use these sticky ends, even the mis-pairing rates are so high? Why not other sticky ends?

The Reason why Freiburg used these sticky ends

Failed to contact the original designers of these sticky ends, what we can do is just to find feasible advantages of these combinations.

Review the TALE repeated amino acids sequence:

LTPEQVVAIAS(XX)GGKQALETVQRLLPVLCQAHG（34aa）

The first amino acid is Leu, which is essential for all connection process. There are six different types of base arrangement for Leu, one of the most number of base arrangement.

UUA, UUG, CUU, CUC, CUA, CUG

The counterproductive sticky ends:

（C）TGAC, GCTC, CTTG, GCTT, ACTG, CCTG, ACTC

The useless of Degeneracy has helped to design seven sticky ends. However, since the codons for identical amino acid are highly similar.

This feature, for experimental scientists, is a double-edged sword.

How to improve the Golden Gate sticky ends? A big Table!

Three basic key questions need to be answered:

1. Whether it's possible to find perfect match pair?
2. Whether we can find a certain number of sticky ends with least possibility to be mismatched?
3. How to make this *sticky-end score table*?

Key algorithms derived from BLAST algorithm

Loose rule: Match: 1;Mismatch:-1; Gap: 0

Strict rule: Match: 1;Mismatch:0; Gap: -1

The sticky end is composed of four bases, which means that we can design 256 types of sticky ends at most.

The forming pair is represented as a 256*256 table.

Find target groups of sticky ends

To solve the TALE parts problem, we need find seven sticky ends, and the similarity score(hereafter referred to as Score) of each pair of them are less than or equal to 1.

Number of Ends	Algorithm	
	Strict	Loose
2	11322	15114
3	102844	234904
4	169519	886369
5	11640	624960
6	76	61624
7	0	640

When we select Strict Algorithm to find these ends, it is impossible to find seven sticky ends, that each pair of them has score no more than 1. So we have to select Loose Algorithm.

Four basepair sticky ends convert to amino acid pair

What we are caring about is whether two amino acids can be located on my target sequence, rather than the 4bp bases. So we should convert the sticky ends information to 2 amino acid.

		U	C	A	G
1st position		FLSYCW	LPHQR	IMTNKSR	VADEG
First two postions	U	FL	L	IM	V
	C	S	P	T	A
	A	Y	HQ	NK	DE
	G	W	R	SR	G
Last two postions	U	FLIV	SPTA	THND	CRSG
	C	FLIV	SPTA	YHND	CRSG
	A	LIV	SPTA	QKE	RG
	G	LMV	SPTA	QKE	WRG
3rd position		FSYCLPHRITNVADG	FSYCLPHRITNVADG	LSPQRITAVAEG	LSWPQRMTKVAEG

Based on the above table, we are able to calculate the total scores of each combination and find the least one.

Best choice for seven sticky ends on TALE protein

Best combination:

AAAA	AGGG	GTAC	GCTC	TTTT	TCGA	CCCC

Scores Table(Loose rule):

	AAAA	AGGG	GTAC	GCTC	TTTT	TCGA	CCCC
AAAA	\	1	1	0	0	1	0
AGGG	1	\	1	1	0	1	0
GTAC	1	1	\	1	1	1	1
GCTC	0	1	1	\	1	1	1
TTTT	0	0	1	1	\	1	0
TCGA	1	1	1	1	1	\	1
CCCC	0	0	1	1	0	1	\

Position in TALE amino acids sequence:

Sticky ends	Amino Acids	Sequence positions
AAAA	GK*	LTPEQVVAIAS(XX)GGKQALETVQRLLPVLCQAHG
AGGG	GG*	LTPEQVVAIAS(XX)GGKQALETVQRLLPVLCQAHG
GTAC	VL* VQ*	LTPEQVVAIAS(XX)GGKQALETVQRLLPVLCQAHG
GCTG	LL* RL* VL* AL* GL*	LTPEQVVAIAS(XX)GGKQALETVQRLLPVLCQAHG
TTTT	LL* VL*	LTPEQVVAIAS(XX)GGKQALETVQRLLPVLCQAHG
TCGA	LE*	LTPEQVVAIAS(XX)GGKQALETVQRLLPVLCQAHG
CCCC	TP* LP*	LTPEQVVAIAS(XX)GGKQALETVQRLLPVLCQAHG

Reconstruct DNA Sequence

Two main factors to reconstruct DNA sequence:

1.Use the table of best combination and rearrange the sticky ends with your demand.

2.No BsmBI recognition sequence in the reconstruct DNA sequence.

Final DNA Sequence for TALE protein.

1 CTGACCCCGG AACAGGTGGT GGCCATTGCA AGCAACGGTG GTGGCAAGCA GGCCCTGGAG

61 ACAGTCCAAC GGCTGCTTCC GGTTCTGTGT CAGGCCCACG GCCTGACTCC AGAACAAGTG

121 GTTGCTATCG CCAGCCACGA TGGCGGAAAA CAAGCCCTCG AAACCGTGCA GCGCCTGCTT

181 CCGGTGCTGT GTCAGGCCCA CGGGCTCACC CCGGAACAGG TGGTGGCCAT CGCATCTAAC

241 AATGGCGGTA AGCAGGCACT GGAAACAGTG CAGCGCCTGC TTCCGGTCCT GTGTCAGGCT

301 CATGGCCTGA CCCCAGAGCA GGTCGTGGCA ATTGCCTCCA ACATTGGAGG GAAGCAGGCA

361 CTGGAGACCG TGCAGCGGCT GCTGCCGGTG CTGTGTCAGG CCCACGGCTT GACCCCGGAA

421 CAGGTGGTGG CCATCGCCTC CAACGGCGGT GGCAAACAGG CGCTGGAAAC AGTTCAACGC

481 CTCCTTCCGG TCCTGTGCCA GGCCCATGGT CTGACTCCAG AGCAGGTTGT GGCAATTGCA

541 AGCAACATTG GTGGTAAACA AGCTTTGGAA ACCGTCCAGC GCTTGCTGCC AGTACTGTGT

601 CAGGCCCACG GGCTTACCCC GGAACAGGTG GTGGCCATTG CAAGCAACGG TGGTGGCAAG

661 CAGGCCCTGG AGACAGTCCA ACGGCTGCTT CCGGTTCTGT GTCAGGCCCA CGGCCTGACT

721 CCAGAACAAG TGGTTGCTAT CGCCAGCCAC GATGGCGGTA AACAAGCCCT CGAAACCGTG

781 CAGCGCCTGC TTCCGGTGCT CTGTCAGGCC CACGGACTGA CCCCGGAACA GGTGGTGGCC

841 ATCGCCTCCA ACATTGGTGG TAAGCAAGCC CTCGAAACTG TGCAGCGGCT GCTTCCAGTC

901 TTGTGCCAGG CTCACGGCCT GACACCGGAG CAGGTGGTTG CAATCGCGTC TAATATCGGC

961 GGCAAACAGG CACTCGAGAC CGTGCAGCGC TTGCTTCCAG TGCTGTGTCA GGCCCACGGC

1021 CTGACCCCGG AACAGGTGGT GGCCATCGCC TCTAACAATG GCGGCAAACA GGCATTGGAA

1081 ACAGTTCAGC GCCTGCTGCC GGTGTTGTGT CAGGCTCACG GCCTGACTCC GGAGCAGGTT

1141 GTGGCCATCG CAAGCCATGA TGGCGGTAAA CAAGCTCTGG AGACAGTGCA ACGCCTCTTG

1201 CCAGTTTTGT GTCAGGCCCA CGGA

Final Amino acids are remain the same：

1 LTPEQVVAIA SNGGGKQALE TVQRLLPVLC QAHG
35 LTPEQVVAIA SHDGGKQALE TVQRLLPVLC QAHG
69 LTPEQVVAIA SNNGGKQALE TVQRLLPVLC QAHG
103 LTPEQVVAIA SNIGGKQALE TVQRLLPVLC QAHG
137 LTPEQVVAIA SNGGGKQALE TVQRLLPVLC QAHG
171 LTPEQVVAIA SNIGGKQALE TVQRLLPVLC QAHG
205 LTPEQVVAIA SNGGGKQALE TVQRLLPVLC QAHG
239 LTPEQVVAIA SHDGGKQALE TVQRLLPVLC QAHG
273 LTPEQVVAIA SNIGGKQALE TVQRLLPVLC QAHG
307 LTPEQVVAIA SNIGGKQALE TVQRLLPVLC QAHG
341 LTPEQVVAIA SNNGGKQALE TVQRLLPVLC QAHG
375 LTPEQVVAIA SHDGGKQALE TVQRLLPVLC QAHG

Corresponding part：

PART-left:

…CTGACCCCGGAGACG

PART1(150bp):

CGTCTCGCCCCGGAACAGGTGGTGGCCATTGCAAGCAACGGTGGTGGCAAGCAGGCCCTGGAGACA
GTCCAACGGCTGCTTCCGGTTCTGTGTCAGGCCCACGGCCTGACTCCAGAACAAGTGGTTGCTATC
GTGGCGGAAAATGAGACG

PART2(219bp):

CGTCTCTAAAACAAGCCCTCGAAACCGTGCAGCGCCTGCTTCCGGTGCTGTGTCAGGCCCACGGGC
TCACCCCGGAACAGGTGGTGGCCATCGCATCTAACAATGGCGGTAAGCAGGCACTGGAAACAGTGC
AGCGCCTGCTTCCGGTCCTGTGTCAGGCTCATGGCCTGACCCCAGAGCAGGTCGTGGCAATTGCCT
CCAACATTGGAGGGCGAGACG

PART3(262bp):

CGTCTCTAGGGAAGCAGGCACTGGAGACCGTGCAGCGGCTGCTGCCGGTGCTGTGTCAGGCCCACG
GCTTGACCCCGGAACAGGTGGTGGCCATCGCCTCCAACGGCGGTGGCAAACAGGCGCTGGAAACAG
TTCAACGCCTCCTTCCGGTCCTGTGCCAGGCCCATGGTCTGACTCCAGAGCAGGTTGTGGCAATTG
CAAGCAACATTGGTGGTAAACAAGCTTTGGAAACCGTCCAGCGCTTGCTGCCAGTACGGAGACG

PART4(224bp):

CGTCTCCGTACTGTGTCAGGCCCACGGGCTTACCCCGGAACAGGTGGTGGCCATTGCAAGCAACGG
TGGTGGCAAGCAGGCCCTGGAGACAGTCCAACGGCTGCTTCCGGTTCTGTGTCAGGCCCACGGCCT
GACTCCAGAACAAGTGGTTGCTATCGCCAGCCACGATGGCGGTAAACAAGCCCTCGAAACCGTGCA
GCGCCTGCTTCCGGTGCTGGGAGACG

PART5(194bp):

CGTCTCCGCTGTGTCAGGCCCACGGACTGACCCCGGAACAGGTGGTGGCCATCGCCTCCAACATTG

GTGGTAAGCAAGCCCTCGAAACTGTGCAGCGGCTGCTTCCAGTCTTGTGCCAGGCTCACGGCCTGA
CACCGGAGCAGGTGGTTGCAATCGCGTCTAATATCGGCGGCAAACAGGCACTCGATGAGACG

PART6(249bp):
CGTCTCATCGAGACCGTGCAGCGCTTGCTTCCAGTGCTGTGTCAGGCCCACGGCCTGACCCCGGAA
CAGGTGGTGGCCATCGCCTCTAACAATGGCGGCAAACAGGCATTGGAAACAGTTCAGCGCCTGCTG
CCGGTGTTGTGTCAGGCTCACGGCCTGACTCCGGAGCAGGTTGTGGCCATCGCAAGCCATGATGGC
GGTAAACAAGCTCTGGAGACAGTGCAACGCCTCTTGCCAGTTTTAGAGACG

PART-right:
CGTCTCATTTTGTGTCAGGCCCACGGA...

The recognition sequence of these TALE protein:

TCGATATCAAGC

(End)

All parts are under artificial synthesis process, so there is few results, which can prove our changes are useful. However, with the principle of complementary base pairing, our chioce should be better than original vision. And if you want our data or use our method to create your own best sticky ends, just contact us!