NucleoMod

NucleoMod can modify CDS based on synoymous mutation. It has 5 applications. Firstly, NucleoMod is used to design CRISPR sites on NeoChr so that we can silence the wild type genes. Secondly, it can erase specific enzyme sites as users command. Thirdly, users can create an enzyme site in selected region of specific genes. Fourthly, it can optimize the codon efficiency to increase the expression level. Finally, it can smash the tandem repeat bases to reduce the synthesis difficulty.

Plugin Scripts

This module contains 5 plugins: CRISPR design, erase enzyme site, create enzyme site, codon optimization, repeat smash. All plugins are included in the main program.

2.1 CRISPR design

This plugin is used to design CRISPR site of NeoChr genes so that we can silence the wild type genes. We use blast+ to ensure the uniqueness of CRISPR sites. If you are using more than one plugin at the same time, this plugin will start firstly and deliver the data to next plugin. Otherwise it will generate a new fasta file for sequence and gff file for annotation.

2.1.1 Internal operation

First, this plugin extracts sequence and annotation from the NeoChr FASTA file and GFF3 file, respectively. Regular expression will be applied to find the 23bp basic structure of CRISPR site, with a head of ‘G’ then following 20 facultative bases and finally followed by ‘GG’. All the sequences and locus will be record in an array.

Second, the blast+ will be used to check whether the 12bp sequences (from 9th to 20th) are uniq in the wild type genome. Only uniq sites will be reserved.

Third, synonymous substitution method will be applied to change one base between the 9th to 20th bases of the CRISPR structure. The result will be record in GFF as an element of gene. If –verbose is set, the designed number will be report in STDOUT.

Finally, if this plugin is the last module, the sequence and annotation information will be recreated in FASTA and GFF format.

2.1.2 Example

We have two input forms to execute the plugin:

Run CRISPR design plugin only:

perl NucleoMod.pl -inputfa NeoChr.fa -inputgff NeoChr.gff -outputgff new_annotation.gff -outputfa new_chr.fa -crisprnum 2 -database saccharomyces_cerevisiae_chr.fa

2.1.3 Parameters

Parameter	Description	Default	Selectable range
inputfa	The NeoChr sequence file in FASTA format		string
inputgff	The NeoChr annotation file in GFF3 format		string
outputgff	Output of new chromosome annotation in GFF3 format		string
outputfa	Output of new chromosome sequence in FASTA format		string
verbose	Output the detailed information in STDOUT	none	option
crisprnum	Number of CRISPR site to be design per gene		Int (>0)
database	The sequence of reference genome, used as blast+ database		string
help	Show help information		

2.1.4 The format of output file

The output files are standard GFF and FASTA format files.

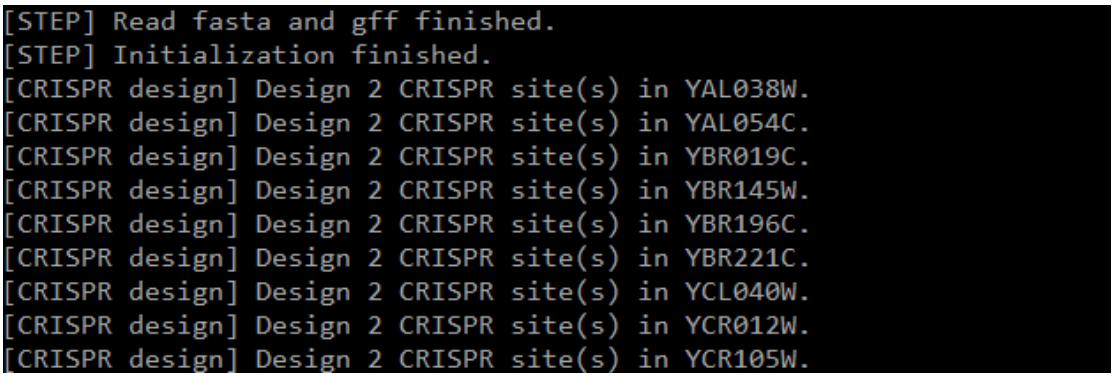
1. GFF file

```
NeoChr  Genovo  left_telomere   1       689     .       +       .       ID=universal_telomere_cap_left;
NeoChr  Genovo  gene    690     3565    .       -       .       ID=YAL054C;display=Acetyl-coA_synthetase_isoform;r
NeoChr  Genovo  3UTR    690     823     .       -       .       Parent=YAL054C;
NeoChr  Genovo  loxp    693     727     .       -       .       ID=site_specific_recombination_target_region;Paren
NeoChr  Genovo  mRNA    824     2965    .       -       .       Parent=YAL054C;
NeoChr  Genovo  CDS     824     2965    .       -       .       Parent=YAL054C;
NeoChr  Genovo  crispr  2156    2178    .       -       .       Parent=YAL054C;CRISPR_seq=GAAAGCAACAGATGGATTGTTGG;
NeoChr  Genovo  crispr  2303    2325    .       -       .       Parent=YAL054C;CRISPR_seq=GGTAGAATGTTGAACACCCTTGG;
```

2. FASTA file

```
>NeoChr
CACACACACCACACCCACACCACACACACACCACACCCACACCACACCCACACACACCCA
CACCCACACACCACACCCACACACACACCCACACCCACACACCACACCCACACCACACAC
CACACCCACACACCACACCCACACACCACACCCACACCACACACCCACACACCACACCCA
CACACACCACACCCACACCCACACACACACACACCACACCCACCACACCCACACACCCAC
ACCCACACACCACACCCACACACACCACACCCACACACCACACCCACACACCCACACCCA
CACATAACTTCGTATAATGTACATTATACGAAGTTATCACATCATTATGCACGGCACTTG
CCTCAGCGGTCTATACCCTGTGCCATTTACCCATAACTCCCACGATTATCCACATTTTAA
TATCTATATCTCATTCGGCGGCCCCAAATATTGTATAACTGCTCTTAATACATACGTTAT
ACCACTTTTACACCATATACTAACCACTCAATTTATATACACTTATGTCAATATTACAAA
```

3. Detailed information in STDOUT



2.2 Erase enzyme site

Given a list of restriction enzyme information, this plugin will erase the restriction sites in every gene. If you are using more than one plugin at the same time, this plugin will start after CRISPR design and deliver the data to next plugin. Otherwise it will generate a new fasta file for sequence and gff file for annotation.

2.2.1 Internal operation

The enzyme information will be extracted. (If the –borbrick tandard parameter is set, it will also remove EcoRI, XbaI, SpeI, PstI and NotI) The recognize site will be reformatted to regular expression and searched in the CDS regions.

Once a restriction site is matched, synonymous substitution method will be applied to try to erase the enzyme site. When the substitution is finished, the plugin will restart the next search from 1 base after the last matched position.

If this plugin is the last module, the sequence and annotation information will be recreated in FASTA and GFF format.

2.2.2 Example

perl NucleoMod.pl -inputfa NeoChr.fa -inputgff NeoChr.gff -outputgff new_annotation.gff -outputfa new_chr.fa –biobrickstandard [-delenzymelist enzyme.list]

Format of enzyme.list:

Company enzyme_name enzyme_site …

Eg. NEB BamHI G/GATCC

2.2.3 Parameters

Parameter	Description	Default	Selectable range
inputfa	The NeoChr sequence file in FASTA format		string
inputgff	The NeoChr annotation file in GFF3 format		string
outputgff	Output of new chromosome annotation in GFF3 format		string
outputfa	Output of new chromosome sequence in FASTA format		string
verbose	Output the detailed information in STDOUT	none	option
biobrickstandar	Erase the biobrick standard	none	option

d	enzyme site		
delenzymelist	The file of enzyme going to delete		string
detail	Show the erased enzyme site in new gff	none	option
help	Show help information		

2.2.4 The format of output

The output files are standard GFF and FASTA format.

GFF file

```
NeoChr  Genovo  enzyme  2952    2957    .       -       .       Parent=YAL054C;name=XbaI;enzyme_seq=TCTAGA;status=removed;
NeoChr  Genovo  codonoptimize  2960    2960    .       -       .       Parent=YAL054C;origin_codon=AAG;optimize_codon=AAA;
NeoChr  Genovo  5UTR    2966    3565    .       -       .       Parent=YAL054C;
NeoChr  Genovo  gene    3566    5802    .       +       .       ID=YAL038W;display=Pyruvate_kinase;repeat_smash=7/16;best_codon_rate=0.93;
NeoChr  Genovo  5UTR    3566    4165    .       +       .       Parent=YAL038W;
NeoChr  Genovo  mRNA    4166    5668    .       +       .       Parent=YAL038W;
NeoChr  Genovo  CDS     4166    5668    .       +       .       Parent=YAL038W;
NeoChr  Genovo  enzyme  4169    4174    .       +       .       Parent=YAL038W;name=XbaI;enzyme_seq=TCTAGA;status=removed;
```

FASTA file

```
>NeoChr
CACACACACCACACCCACACCACACACACACCACACCCACACCACACCCACACACACCCA
CACCCACACACCACACCCACACACACACCCACACCCACACACCACACCCACACCACACAC
CACACCCACACACCACACCCACACACCACACCCACACCACACACCCACACACCACACCCA
CACACACCACACCCACACCCACACACACACACACCACACCCACCACACCCACACACCCAC
ACCCACACACCACACCCACACACACCACACCCACACACCACACCCACACACCCACACCCA
CACATAACTTCGTATAATGTACATTATACGAAGTTATCACATCATTATGCACGGCACTTG
CCTCAGCGGTCTATACCCTGTGCCATTTACCCATAACTCCCACGATTATCCACATTTTAA
TATCTATATCTCATTCGGCGGCCCCAAATATTGTATAACTGCTCTTAATACATACGTTAT
ACCACTTTTACACCATATACTAACCACTCAATTTATATACACTTATGTCAATATTACAAA
```

Detailed information in STDOUT

```
[STEP] Design CRISPR site finished.
[Erase Enzyme] Delete XbaI in YIL177C, position 28500
[Erase Enzyme] Delete PstI in YIL177C, position 28697
[Erase Enzyme] Delete EcoRI in YIL177C, position 29574
[Erase Enzyme] Delete SpeI in YIL177C, position 29829
[Erase Enzyme] Delete PstI in YIL177C, position 26802
[Erase Enzyme] Delete SpeI in YIL177C, position 27495
[Erase Enzyme] Delete SpeI in YIL177C, position 27531
[Erase Enzyme] Delete SpeI in YIL177C, position 27711
[Erase Enzyme] Delete SpeI in YIL177C, position 27747
[Erase Enzyme] Delete SpeI in YIL177C, position 27783
[Erase Enzyme] Delete SpeI in YIL177C, position 27819
[Erase Enzyme] Delete XbaI in YAL038W, position 4169
[Erase Enzyme] Can not remove all enzyme in YAL038W, recorded in new gff.
```

2.3 Create enzyme site

Given a list of restriction enzyme information, this plugin can create a new enzyme site in specific region of selected gene. If you are using more than one plugin at the same time, this plugin will start after erase enzyme site and deliver the data to next plugin. Otherwise it will generate a new fasta file for sequence and gff file for annotation.

2.3.1 Internal operation

First, information of enzyme site will be extracted. According to 3 reading frames, a searching tree will be constructed and converted to regular expression.

The plugin will search the selected regions and then change the sequence to enzyme site by synonymous substitution method.

If this plugin is the last module, the sequence and annotation information will be recreated in FASTA and GFF format.

2.3.2 Example

perl NucleoMod.pl -inputfa NeoChr.fa -inputgff NeoChr.gff -outputgff new_annotation.gff -outputfa new_chr.fa -addenzymelist enzyme.list -addenzymeconfig gene_id,start_pos,end_pos,enzyme_name

2.3.3 Parameters

Parameter	Description	Default	Selectable range
inputfa	The NeoChr sequence file in FASTA format		string
inputgff	The NeoChr annotation file in GFF3 format		string
outputgff	Output of new chromosome annotation in GFF3 format		string
outputfa	Output of new chromosome sequence in FASTA format		string
verbose	Output the detailed information in STDOUT	none	option
addenzymelist	The file of enzyme to get enzyme site information		string
addenzymeconfig	A array of string to specify enzyme and regions		string,int,int,string
help	Show help information		

2.3.4 The format of ouput

The output files are standard GFF and FASTA format.

GFF file

```
NeoChr  Genovo  gene    5803    8636    .       -       .       ID=YBR019C;display=UDP-glucose-4-epimerase;repeat_smash=
NeoChr  Genovo  3UTR    5803    5936    .       -       .       Parent=YBR019C;
NeoChr  Genovo  loxp    5806    5840    .       -       .       ID=site_specific_recombination_target_region;Parent=YBR0
NeoChr  Genovo  mRNA    5937    8036    .       -       .       Parent=YBR019C;
NeoChr  Genovo  CDS     5937    8036    .       -       .       Parent=YBR019C;
NeoChr  Genovo  enzyme  6497    7942    .       -       .       Parent=YBR019C;name=EcoRI;enzyme_seq=CAATAG;status=addin
NeoChr  Genovo  crispr  7164    7186    .       -       .       Parent=YBR019C;CRISPR_seq=GTGGATATAATCCCTGATCGGGG;sub_se
NeoChr  Genovo  crispr  7688    7710    .       -       .       Parent=YBR019C;CRISPR_seq=GCGCCTATATAAGCACTATCAGG;sub_se
NeoChr  Genovo  5UTR    8037    8636    .       -       .       Parent=YBR019C;
NeoChr  Genovo  gene    8637    10426   .       +       .       ID=YBR145W;display=Alcohol_dehydrogenase_isoenzyme_V;rep
NeoChr  Genovo  5UTR    8637    9236    .       +       .       Parent=YBR145W;
```

FASTA file

```
>NeoChr
CACACACACCACACCCACACCACACACACACCACACCCACACCACACCCACACACACCCA
CACCCACACACCACACCCACACACACACCCACACCCACACACCACACCCACACCACACAC
CACACCCACACACCACACCCACACACCACACCCACACCACACACCCACACACCACACCCA
CACACACCACACCCACACCCACACACACACACACCACACCCACCACACCCACACACCCAC
ACCCACACACCACACCCACACACACCACACCCACACACCACACCCACACACCCACACCCA
CACATAACTTCGTATAATGTACATTATACGAAGTTATCACATCATTATGCACGGCACTTG
CCTCAGCGGTCTATACCCTGTGCCATTTACCCATAACTCCCACGATTATCCACATTTTAA
TATCTATATCTCATTCGGCGGCCCCAAATATTGTATAACTGCTCTTAATACATACGTTAT
ACCACTTTTACACCATATACTAACCACTCAATTTATATACACTTATGTCAATATTACAAA
```

Detailed information in STDOUT

```
[Erase Enzyme] Delete PstI in YAL054C, position 2755
[Erase Enzyme] Delete PstI in YIL177W-A, position 33182
[Erase Enzyme] Delete PstI in YIL177W-A, position 33206
[Erase Enzyme] Delete PstI in YIL177W-A, position 33338
[STEP] Remove enzyme site finished.
[Create Enzyme] Successfully add EcoRI enzyme in YBR019C, position 6497.
[STEP] Add enzyme site finished.
```

2.4 Codon optimization

Given a codon priority list, this plugin is used to optimize the codon so that we can increase the expression of selected genes. If you are using more than one plugin at the same time, this plugin will start after create enzyme site and deliver the data to next plugin. Otherwise it will generate a new fasta file for sequence and gff file for annotation.

2.4.1 Internal operation

The codon with same amino acid will be separated into 3 ranks, best normal and worst. Every codon of selected gene will be check whether the codon is in best rank. The codon in normal or worst will be change to best rank by synonymous substitution method.
If this plugin is the last module, the sequence and annotation information will be recreated in FASTA and GFF format.

2.4.2 Example

perl NucleoMod.pl -inputfa NeoChr.fa -inputgff NeoChr.gff -outputgff new_annotation.gff -outputfa new_chr.fa -codonoptimize CodonPriority.txt -optimizeallgene [-optimizegenelist gene1,gene2,gene3]

2.4.3 Parameters

Parameter	Description	Default	Selectable range
inputfa	The NeoChr sequence file in FASTA format		string
inputgff	The NeoChr annotation file in GFF3 format		string
outputgff	Output of new chromosome annotation in GFF3 format		string
outputfa	Output of new chromosome sequence in FASTA format		string
verbose	Output the detailed information in STDOUT	none	option
codonoptimize	Codon priority list to get the ranking information		string
optimizeallgene	Optimize all genes in inputgff		option
optimizegenelist	A list of gene going to optimize, separate by		string,string,string,…

	comma		
detail	Show the optimization sequence in new gff	none	option
help	Show help information		

2.4.4 The format of ouput

The output files are standard GFF and FASTA format.

GFF file

```
NeoChr  Genovo  mRNA    824     2965    .       -       .       Parent=YAL054C;
NeoChr  Genovo  CDS     824     2965    .       -       .       Parent=YAL054C;
NeoChr  Genovo  codonoptimize  827     827     .       -       .       Parent=YAL054C;origin_codon=TGC;
NeoChr  Genovo  codonoptimize  836     836     .       -       .       Parent=YAL054C;origin_codon=GGA;
NeoChr  Genovo  codonoptimize  854     854     .       -       .       Parent=YAL054C;origin_codon=TTA;
NeoChr  Genovo  codonoptimize  860     860     .       -       .       Parent=YAL054C;origin_codon=TGC;
NeoChr  Genovo  codonoptimize  863     863     .       -       .       Parent=YAL054C;origin_codon=ACG;
NeoChr  Genovo  codonoptimize  866     866     .       -       .       Parent=YAL054C;origin_codon=TAC;
NeoChr  Genovo  codonoptimize  878     878     .       -       .       Parent=YAL054C;origin_codon=CAC;
NeoChr  Genovo  codonoptimize  881     881     .       -       .       Parent=YAL054C;origin_codon=TTA;
NeoChr  Genovo  codonoptimize  887     887     .       -       .       Parent=YAL054C;origin_codon=TCA;
NeoChr  Genovo  codonoptimize  890     890     .       -       .       Parent=YAL054C;origin_codon=TTA;
NeoChr  Genovo  codonoptimize  899     899     .       -       .       Parent=YAL054C;origin_codon=CCG;
NeoChr  Genovo  codonoptimize  911     911     .       -       .       Parent=YAL054C;origin_codon=TAC;
```

FASTA file

```
>NeoChr
CACACACACCACACCCACACCACACACACACCACACCCACACCACACCCACACACACCCA
CACCCACACACCACACCCACACACACACCCACACCCACACACCACACCCACACCACACAC
CACACCCACACACCACACCCACACACCACACCCACACCACACACCCACACACCACACCCA
CACACACCACACCCACACCCACACACACACACACCACACCCACCACACCCACACACCCAC
ACCCACACACCACACCCACACACACCACACCCACACACCACACCCACACACCCACACCCA
CACATAACTTCGTATAATGTACATTATACGAAGTTATCACATCATTATGCACGGCACTTG
CCTCAGCGGTCTATACCCTGTGCCATTTACCCATAACTCCCACGATTATCCACATTTTAA
TATCTATATCTCATTCGGCGGCCCCAAATATTGTATAACTGCTCTTAATACATACGTTAT
ACCACTTTTACACCATATACTAACCACTCAATTTATATACACTTATGTCAATATTACAAA
```

Detailed information in STDOUT

```
[Erase Enzyme] Delete PstI in YIL177W-A, position 33206
[Erase Enzyme] Delete PstI in YIL177W-A, position 33338
[STEP] Remove enzyme site finished.
[Create Enzyme] Successfully add EcoRI enzyme in YBR019C, position 6497.
[STEP] Add enzyme site finished.
[STEP] Codon optimization finished.
```

2.5 Repeat smash

This plugin go through the CDS region to find out the tandem repeat bases. Synonymous substitution method will be applied to break long tandem repeat base to reduce the synthesis difficulty. If you are using more than one plugin at the same time, this plugin will start finally and then it will generate a new fasta file for sequence and gff file for annotation.

2.5.1 Internal operation

Regular expression is used to find out the tandem repeat bases longer then specified length (usually longer than 5bp). From the third of the matched sequence, synonymous substitution method will be applied to break the tandem repeat bases.

If the substitution is successful and the rest sequence is still longer than the cutoff, then it will move to next 3 bases and do the same thing.

The sequence and annotation information will be recreated in FASTA and GFF format.

2.3.2 Example

perl NucleoMod.pl -inputfa NeoChr.fa -inputgff NeoChr.gff -outputgff new_annotation.gff -outputfa new_chr.fa -repeatsmash 5

2.3.3 Parameters

Parameter	Description	Default	Selectable range
inputfa	The NeoChr sequence file in FASTA format		string
inputgff	The NeoChr annotation file in GFF3 format		string
outputgff	Output of new chromosome annotation in GFF3 format		string
outputfa	Output of new chromosome sequence in FASTA format		string
verbose	Output the detailed information in STDOUT	none	option
repeatsmash	The tandem repeat bases longer or equal to this cutoff will be smashed		int
detail	Show the repeat smash result in new gff	none	option
help	Show help information		

2.3.4 The format of ouput

The output files are standard GFF and FASTA format.

GFF file

```
NeoChr  Genovo  repeatsmash     1037    1044    .       -       .       Parent=YAL054C;origin_seq=AAAAAAAA;
NeoChr  Genovo  codonoptimize   1037    1037    .       -       .       Parent=YAL054C;origin_codon=AAG;opt
NeoChr  Genovo  codonoptimize   1049    1049    .       -       .       Parent=YAL054C;origin_codon=ATA;opt
NeoChr  Genovo  codonoptimize   1052    1052    .       -       .       Parent=YAL054C;origin_codon=GAG;opt
NeoChr  Genovo  codonoptimize   1055    1055    .       -       .       Parent=YAL054C;origin_codon=TTA;opt
NeoChr  Genovo  codonoptimize   1058    1058    .       -       .       Parent=YAL054C;origin_codon=ATC;opt
NeoChr  Genovo  repeatsmash     1073    1081    .       -       .       Parent=YAL054C;origin_seq=TTTTTTTTT
NeoChr  Genovo  codonoptimize   1082    1082    .       -       .       Parent=YAL054C;origin_codon=CCT;opt
```

FASTA file

```
>NeoChr
CACACACACCACACCCACACCACACACACACCACACCCACACCACACCCACACACACCCA
CACCCACACACCACACCCACACACACACCCACACCCACACACCACACCCACACCACACAC
CACACCCACACACCACACCCACACACCACACCCACACCACACACCCACACACCACACCCA
CACACACCACACCCACACCCACACACACACACACCACACCCACCACACCCACACACCCAC
ACCCACACACCACACCCACACACACCACACCCACACACCACACCCACACACCCACACCCA
CACATAACTTCGTATAATGTACATTATACGAAGTTATCACATCATTATGCACGGCACTTG
CCTCAGCGGTCTATACCCTGTGCCATTTACCCATAACTCCCACGATTATCCACATTTTAA
TATCTATATCTCATTCGGCGGCCCCAAATATTGTATAACTGCTCTTAATACATACGTTAT
ACCACTTTTACACCATATACTAACCACTCAATTTATATACACTTATGTCAATATTACAAA
```

Detailed information in STDOUT

```
[Create Enzyme] Successfully add EcoRI enzyme in YBR019C, position 6497.
[STEP] Add enzyme site finished.
[STEP] Codon optimization finished.
[STEP] Repeat-smash finished.
[STEP] Ranking optimization finished.
```

3 Description of GFF file

Description	Element	explanation
repeat_smash	gene	Smashed number/total number
best_codon_rate	gene	The rate is equal by best rank codon number/all codon number
CRISPR_seq	crispr	The sequence of wild type, used to construct cas system
sub_seq	crispr	The 12bp uniq sequence
change_pos	crispr	The modified position in cirspr
change_base	crispr	The modified base
name	enzyme	Name of restriction enzyme
enzyme_seq	enzyme	Sequence of restriction enzyme
status	enzyme	Status of specific enzyme. removed means erase successfully; immutable means fail to erase; add means new created enzyme.
origin_codon	codonoptimize	The codon before optimization
optimize_codon	codonoptimize	The codon after optimization
origin_seq	repeatsmash	The tandem repeat sequence before optization
optimize_seq	repeatsmash	The tandem repeat sequence after optization