Scroing siRNA project [python : scoring_siRNA] by Yeom

Purpose: 효능적인 dual targeting siRNA sequence을 scoring scheme을 통해서 평가하고, siRNA candidate를 형성하는 프로그램. Random sequence의 efficacy 대비 scoring을 비교해보고 meta-analysis 하는 프로그램.

1.Introduction

-Dual targeting siRNA

-Reynold scoring

-Thermodynamics

2.Method

-Input data(by. Youn Beong)

Group	Value	GENE	primer	pmol	date	cell	출처	A_siRNA	A_sequenc	A_LOC	length	B_siRNA	B_sequenc	B_LOC	Mis	BOTH	A_GC	B_GC	A_score	Full_score	Top	Bot	RISC_BOTH
#1_MET_N	0.597487	MET	-	50	2021.07.22	C42B	RESULT	AAGUAGA	TCAGCgat	CDS	19	UCAGCUU	AAGTAGA	3'UTR	4	L5R5	42.11	47.37	10	18	4.6	-4.6	0
#2_MET_N	0.630628	MET	-	50	2021.07.22	C42B	RESULT	CUUCUAU	TCTCCCCc	CDS	18	UCUCCCC	CTTCTccC	CDS	4	L5R5	50	61.11	6	18	3.1	-3.1	0
#3_MET_N	0.721726	MET	-	50	2021.07.22	C42B	RESULT	AUGCAAU	CCCAGAT	CDS	19	CCCAGAU	ATGCctTG	3'UTR	4	L4R5	47.37	52.63	11	16	8.1	-8.1	0
#4_MET_N	0.740451	MET	-	50	2021.07.22	C42B	RESULT	GAUUCCC	ATTTCTGa	CDS	19	AUUUCUC	GATTCCtT	3'UTR	4	L5R5	47.37	31.58	8	15	-2	2	0
#5_MET_N	0.831402	MET	-	50	2021.07.22	C42B	RESULT	GAUUCCC	TTTCTGac	CDS	18	UUUCUGU	GATTCCtT	3'UTR	4	L5R5	50	33.33	6	15	-1.8	1.8	0
#6_MET_N	0.936045	MET	-	50	2021.07.22	C42B	GC(O)_NC	UAAAGUC	TATGGcTC	CDS	19	UAUGGGU	TAAtcTaC	3'UTR	4	L3R5	47.37	42.11	12	22	8.9	-8.9	0
#7_MET_N	1.03212	MET	-	50	2021.07.22	C42B	GC(O)_NC	GGCACCA	TcACATTt	CDS	19	UGACAUU	GaCACCg	CDS	4	L1R1	47.37	52.63	9	21	-4.4	4.4	0
#8_MET_N	0.948984	MET	-	50	2021.07.22	C42B	GC(X)_NO	CAAGAAU	TGtCAAAC	CDS	18	UGACAAA	CAAGgAT	5'UTR	4	L4R2	33.33	50	7	20	0.8	-0.8	1
#9_MET_N	0.917495	MET	-	50	2021.07.22	C42B	RESULT	UCAAAAC	TGTGTtTa	CDS	18	UGUGUGU	TCAAgAG	CDS	4	L4R5	33.33	50	11	23	6.8	-6.8	0

각 gene에 대한 mRNA에서 random으로 siRNA candidate를 만들어 gene combination duplex을 위해 mismatch 4 이하로 설정하여, dual targeting siRNA duplex을 형성한다. 이후, 실험을 통해서 regulation efficacy를 도출하고, 이를 database한 excel 파일.

→ 각 antisense(duplex 중, 주 siRNA가닥)을 A_siRNA로, sense를 A_siRNA의 상보적인 가닥으로 설정한다.

→ dual targeting duplex : A_siRNA + B_siRNA → 각 RNA sequence

→ "Value" : 1(no change), v<1 (down regulation), v>1(up regulation)

→ 1-"Value" : siRNA의 regulation efficacy

-Scoring(Thermodynamics, Reynold)

1. Thermodynamics

-first_A/U : antisense first position is A or U base

-last G/C : antisense last position is G or C base

-U_10 : antisense mid position is A base

-GC_stretch : antisense continuous GC stretch less than 9 base

-GC_content :antisense GC content

-Tm_hloop : siRNA duplex Tm (Tm must be less than criteria at which hairpin loop structure is dissociated.)

-seed_3_A/U : number of A or U base at seed sequence

-seed_Tm : Tm of seed seq of antisense with complementary sequence

-GC_seed : GC content of seed sequence

-GC_non_seed ; GC content of non seed part

2. Reynold

-reynold_fixed : Reynold가 제시한 position 위치에 맞게 first, last midpoint는 유지.

-reynold_ratio : Reynold가 제시한 position(19mer 기준)을 sequence길이에 맞춰 재조정

-Correlation analysis : Value(experiment) – Score

[Weight method] : 각 criteria 요소의 중요도를 고려하기 위해 가중치 설정

1. sum : 단순 합산

2. f score : value를 기준에 맞춰 binary로 변환 후, criteria 요소별 f score 계산(정확도, 재현율)

3. Least square(linear regression) :

전체 criteria binary data를 matrix연산으로 binary value에 선형 회귀 연산

Linear regression coefficient를 가중치로 설정

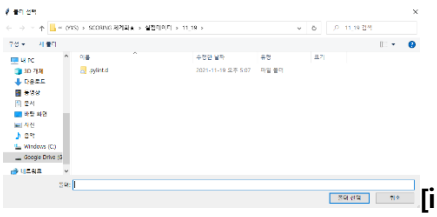
4. Accuracy : binary value와 각 요소별 binary data를 비교하며 각 요소의 Accuracy 계산

5 likelihood ratio(우도비)

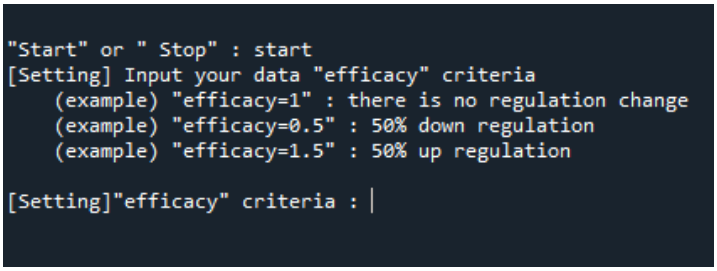
[Polynomial regression]

계산된 가중치로 sequence 별, weighted score 계산.

-Program Process



[input data 파일 위치 지정]



[실험 Value를 binary(0,1)로 변한하기 위한 기준 설정]

```
[Setting] Aricle default values are : "36-54-19-54-20"

[Setting] whole seq GC_content's under limit : 36

[Setting] whole seq GC_content's upper limit : 54

[Setting] seed seq GC_content limit : 19

[Setting] non_seed seq GC_content limit : 54

[Setting] seed seq Tm limit : 20

[Setting] Input your data file name!(format : xlsx):
```

[논문 기반 criteria 요소 별 기준 설정]

[Input data의 파일 이름 알려주기]

[ex] Total_result]

```
[Setting] Select your mode

"yes" : yes! consider sequence mismatches
"no" : no! consider sequence mismatches

[Setting] Select your mode_ (yes_mismatch | no_mismatch ) :   yes
```

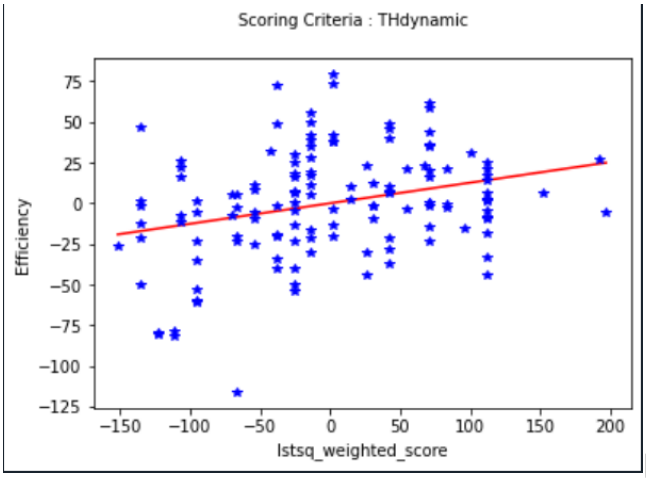
[antisense-sense mismatch 고려여부]

```
[Current] Your current "weight method" is "lstsq_weighted_score"

[Setting] If you want to continue the fitting process, then input Linear fitting degree
[Setting] If you want to stop fitting, then input "stop" command

[Setting] Input Linear Regression polynomial fitting degree (ex_ 1, 2, 3, 4... or stop) :
```

[각 weight calculation에 대해, polynomial fitting 차수 설정]



[차수 설정시, 다음과 같이 plotting 된다.]

```
[Current] Your current "weight method" is "lstsq_weighted_score"

[Setting] If you want to continue the fitting process, then input Linear fitting degree
[Setting] If you want to stop fitting, then input "stop" command

[Setting] Input Linear Regression polynomial fitting degree (ex_ 1, 2, 3, 4... or stop) : 1
[Result] Your fitting coefficient : [0.11605369 0.        ]
[Result] Determinant of Coefficient(Manual) : 0.11386354790325981

[Result] Pearson Coeffiecient : (0.3374367317042706, 7.600145299093366e-05)
```

[plotting시, 다음과 같이 R square value, Pearson coefficient, p-value 제공한다.]

```
[Setting] If you want to continue the fitting process, then input Linear fitting degree
[Setting] If you want to stop fitting, then input "stop" command

[Setting] Input Linear Regression polynomial fitting degree (ex_ 1, 2, 3, 4... or stop) :
```

[결과 값이 괜찮다면, "stop"명령을 아니면, 다른 차수로 fitting을 진행할 수 있다.]

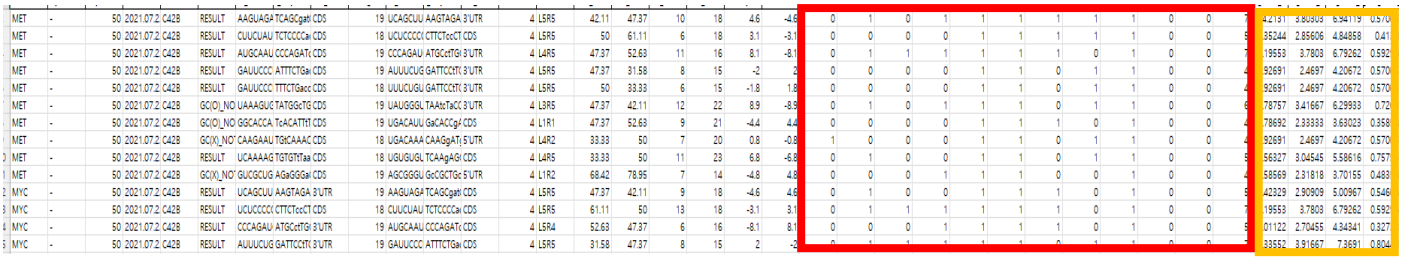
3.Result

-ouput data

이후, program process 끝나면 result 파일 자동 저장

[ouput example : R_yes_mis(2021_11_19)_e1.0]

요소별 scoring **Weighted scoring**



[scoring criteria 별, 가중치 계산 방법 별, linear regression 결과]

	first_A/U	last_G/C	U_10	GC_stretch	GC_conten	Tm_hloop	seed_3_A/U	seed_Tm	GC_seed	C_non_see	coef[High~Low]	degree	R^2_manual	pearson coeff
sum_score	1	1	1	1	1	1	1	1	1	1	[-0.11121563 0.]	1	0.001492982	(-0.03863912349395085, 0.6600299964189329)
f_score	0.636364	0.547945	0.224299	0.788991	0.425532	0.788991	0.565517	0.78341	0	0.129032	[0.06290959 0.]	1	0.001738456	(0.04169479484619909, 0.6350100737208029)
Accuracy	0.575758	0.5	0.371212	0.651515	0.386364	0.651515	0.522727	0.643939	0.348485	0.386364	[0.02628857 0.]	1	0.000350573	(0.018723581711724873, 0.8312565896198252)
LR	1.379437	1.069767	0.713178	1	0.64186	1	1.218346	0.988372	0	3.209302	[0.19231521 0.]	1	0.003056099	(0.055281994808796954, 0.5289677670681412)
lstsq	0.188992	0.075327	-0.0311	-4.8E+13	-0.15556	4.81E+13	0.055219	-0.21173	0	0.321455	[0.12662729 0.]	1	0.085044999	(0.2916247576236317, 0.0006920691632674202

4.Discusstion

-input data vs scoring scheme

다른 scoring scheme을 통해, randomly selected된 sequence을 거른 후(여기서부터 1차 filtering에 대한 유의성 검사 시도가 없음.), 실험값과, scoring scheme에 의한 점수와 비교는 의미가 없다. Scoring scheme의 significance는 scheme에 의한 패턴으로 형성된 sequence에 대한 실험값과, randomly shuffled된 sequence의 실험값 비교를 통해 p-value 도출로 판단할 수 있다.

-gene 별 profile

Input data를 보면 gene별 파악이 아닌, 기존 scoring scheme으로 선별된 sequence에 대한 실험 값이다.

Gene별 별도의 고찰 없이, 단순 sequence에 대한 siRNA 패턴은 의미가 없다고 생각한다.

Gene별 sequence 혹은 dual targeting group별 sequence에 대한 데이터 축적으로 gene별 dual targeting group 별 각기 다른 siRNA pattern 혹은 profile이 형성이 가능하다고 기대한다.

-Further research

[Criteria 독립성 검정]

선형회귀를 위한 가중치 계산할 시, 각 기준에 대한 독립성 혹은 상관성을 검정하지 않은 채, 독립을 가정하고 가중치 계산을 했다. 기준에 대한 독립성 역시 선형회귀에서 다른 결과를 도출할 요인이 될 수 있다.

~~**[Gene 별 siRNA pattern 분석]**~~ – **clustering, 가중치 계산, threshold – Decision Dependent**

Random shuffling된 siRNA sequence pool(길이 일치)을 scoring scheme으로 scoring(단순 sum)한다. ~~이후, clustering을 통해 group화 하여~~ 각각 clustering된 score range을 보고, group을 labeling을 한다. clustering하는 방법 선택은 다음과 같다.

-binary PCA(logistic PCA) clustering & labeling

1. logistic PCA

2.Multiple Correspondence Analysis (MCA)

이후, labeling된 data을 가지고, 다시 가중치를 계산하고(계산 방법은 선택한다.), 점수를 계산한 후, 다시 반복적으로 clustering, labeling, 가중치 계산 점수 계산을 한다. 수렴된 가중치, 점수 값이 생길 때까지 iteration.

이후, 일정한 threshold을 넘는 sequence 한에서 실험을 진행 후, 실험 값(regulation value)과 score에 대한 correlation을 파악한다.(linear regression). 이후, 관련도, regression에 대한 meta analysis을 한 후, 결과가 좋다면 가중치 계산 방법과 clustering 방법 채택, 아니라면 가중치 계산 방법을 바꾼다.

~~**[Gene 별 siRNA profile(PSSM)]**~~ **- Database, initial candidate – Decision Independent(Complete Roughly)**

목적은 gene 별 siRNA의 position별 가중치 profile(Position specific scoring matrix : PSSM)을 만드는 것이다.

Idea ::mRNA binding site(motif finding) + siRNA scoring scheme

일단 더 생각해봐야 한다.