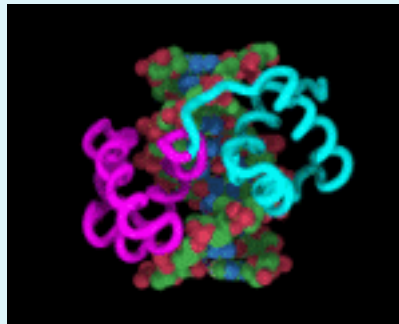


A Levenshtein distance metric for predicting functional siRNA



Liyang Diao, Wake Forest University
**Mentor: Dr. Stanley Dunn, Dept. of Biomedical
Engineering**



❖ Short-interfering RNA

❖ Interferes with mRNA

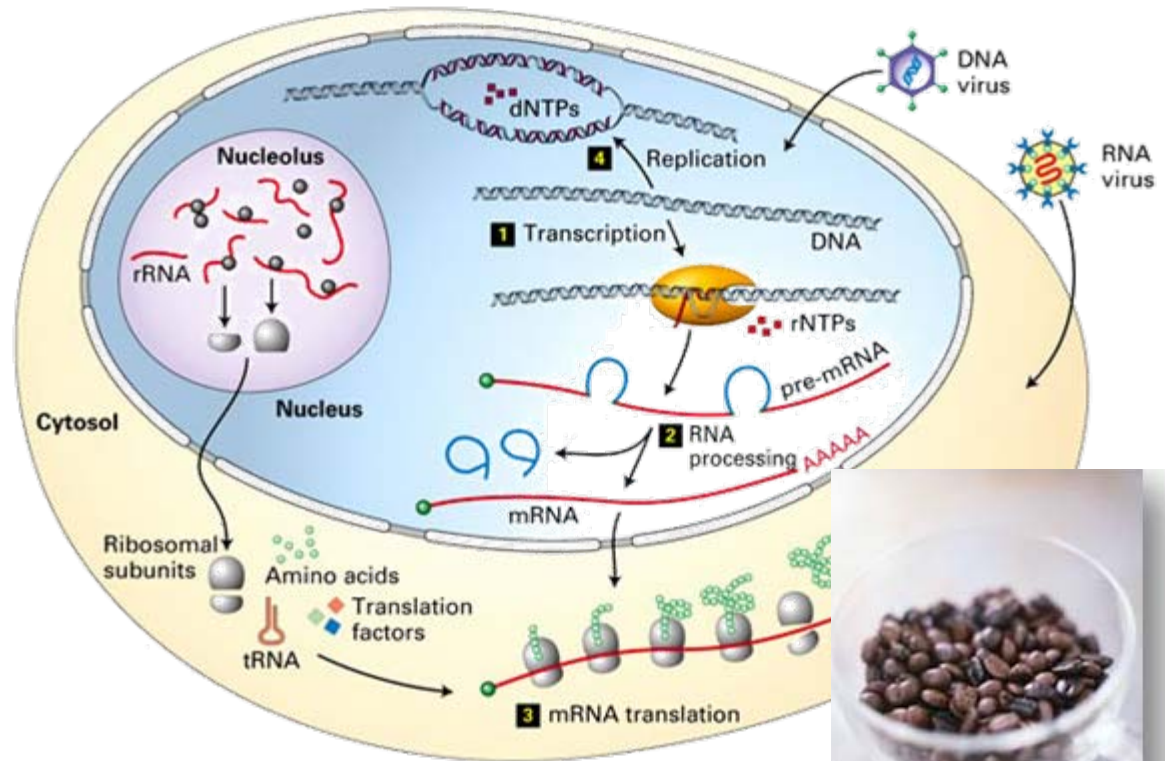
- ❖ Inhibits specific proteins from being produced

❖ How proteins are made

- ❖ **Transcription**
 - ❖ DNA → RNA
- ❖ **Translation**
 - ❖ mRNA → protein
- ❖ **Protein!**

❖ Some proteins we would like to suppress

- ❖ Ex: Knocked out caffeine genes in coffee plants.



What is siRNA?

<http://fig.cox.miami.edu/~cmallery/255/255hist/mcb4.1.dogma.jpg>
<http://www.nature.com/news/2003/030616/full/030616-12.html>

❖ Which strings of siRNA effectively silence genes?

❖ Too many to test every single one

❖ Tried combinatorics

- ❖ Results: About 25% of all strings (of 20 nt strands) fit ideal properties of functional siRNA
- ❖ BUT this amounts to about

274,877,907,000

strings...

hTF173	CAGGCACUACAAAUACUGU
hTF372	GAAGCAGACGUACUUGGCA
PSK208	CUUCAGUGGCACUAAGAGU
PSK229	CGUGUACAAGCACUUAUC
PSK499	GUAUGCCAUCAAGAUGAUU
PSK566	GUGCUGCGUCGGGUGCGUC
PSK793	UGAGAAUCUGCUCUACUAC
CSK909	GCCUCAUUAACCAAAGGU
CSK1079	UGCAUUAAGAACGACGCCA
CSK1121	GCCUCAGUCAUGACGCAAC
CSK1281	GUCUCCUCAAGUUCUCGCU
CSK1481	GCCCUGAGAGAGAAGAAAU
CSK1813	AGAUCAUUGGACCUGGUGCC
CSK2033	ACUCGCCUUCUAGAGUUU
mTF220	CAGGCAUUCAGAGAAAGC
mTF223	GCAUUCAGAGAAAAGCGUU
mTF229	CAGAGAAAGCGUUUAAUUU
mTF321	CAGAUAAAGUGAUCGAUCUA
mTF355	AGUGCUUCUCGACCAACA
hTF77	UGGAGACCCCUGCCUGGCC
hTF158	AGGUGGCCGGCGCUUCAGG
hTF161	UGGCCGGCGCUUCAGGCAC
hTT170	CUUCAGGCACUACAAUAC
hTF176	GCACUACAAAUACUGUGGC
hTF256	CCCGUCAAUCAAGUCUACA
hTF459	CUCCCCAGAGUUCACACCU
hTF478	UACCUGGAGACAAACCUCG
hTF562	CGGACUUUAGUCAGAAGGA
hTF929	GCUGGAAGGAGAACUCCCC
PSK132	UGGGCUGUGGGACAAGCAA
PSK314	UGCCCCGGUCCCCGACUG
PSK382	UAAGUACAGGGCCAAGUUU
PSK411	UUACAGCUAAGUAUGACAU
PSK546	UGUGUGAGUCGGAGCUGCG
PSK739	GGAUGGCGUCCGGUAUCUG
CSK455	AUUGCCAAGUACAACUCC
CSK558	AAGCCAAAAACAAGGUGGG
CSK839	GAGGUGUACUUUGAGAACC
CSK1573	UCCAAGAAUUCUCCUGAAG
mTF217	GUGCAGGCAUUCAGAGAA
mTF226	UUCAGAGAAAGCGUUUAA
mTF245	UUUAACUUGGAUAUCAACU



The Problem...



❖ Levenshtein Distance

- ❖ Calculate distance between strings based on whether character n in string1 is the same as character n in string2.
- ❖ Minimum number of substitutions/insertions required to transform one string to another.

❖ Modifications

- ❖ Used weights from Vert's paper¹
 - ❖ Each substitution no longer increments distance by uniform amount
- ❖ Depends on
 1. Position of nucleotide substitution
 2. Type of substitution

		m	e	i	l	e	n	s	t	e	i	n
	0	1	2	3	4	5	6	7	8	9	10	11
l	1	1	2	3	3	4	5	6	7	8	9	10
e	2	2	1	2	3	3	4	5	6	7	8	9
v	3	3	2	2	3	4	4	5	6	7	8	9
e	4	4	3	3	3	3	4	5	6	6	7	8
n	5	5	4	4	4	4	3	4	5	6	7	7
s	6	6	5	5	5	5	4	3	4	5	6	7
h	7	7	6	6	6	6	5	4	4	5	6	7
t	8	8	7	7	7	7	6	5	4	5	6	7
e	9	9	8	8	8	7	7	6	5	4	5	6
i	10	10	9	8	9	8	8	7	6	5	4	5
n	11	11	10	9	9	9	8	8	7	6	5	4

Levenshtein Distance

1. Vert JP, Foveau N, Lajaunie C, Vandenbrouck Y: "An accurate and interpretable model for siRNA efficacy prediction". *BMC Bioinformatics*. 2006, 7:520.

Table 3: Weights of the linear model trained on the sparse-2l representation.

Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
A1	10.94	C1	0	G1	-2.30	U1	18.39
A2	0	C2	-6.67	G2	-7.28	U2	0.97
A3	1.10	C3	-2.74	G3	-1.23	U3	0
A4	1.19	C4	-1.67	G4	-1.25	U4	0
A5	0.60	C5	-0.78	G5	0	U5	1.32
A6	-0.25	C6	0	G6	0	U6	3.10
A7	0.11	C7	-5.86	G7	-2.02	U7	0
A8	0	C8	0	G8	-1.42	U8	0.95
A9	0	C9	0	G9	-0.55	U9	1.17
A10	3.85	C10	-0.50	G10	0	U10	0
A11	1.69	C11	-0.08	G11	0	U11	1.99
A12	0	C12	0	G12	-0.59	U12	0.01
A13	0.88	C13	-1.06	G13	-1.81	U13	0
A14	1.09	C14	0	G14	-4.37	U14	1.04
A15	0	C15	0.48	G15	0	U15	0
A16	1.35	C16	0	G16	-0.24	U16	0
A17	0	C17	0	G17	-1.70	U17	0.62
A18	-3.61	C18	2.99	G18	0	U18	-0.04
A19	-5.04	C19	6.31	G19	5.32	U19	0
A20	-0.91	C20	-2.64	G20	0	U20	0.65
A21	-0.05	C21	-5.83	G21	1.56	U21	0

This table gives the weights of the linear model trained on the sparse-2l representation to predict the siRNA efficacy (offset = 66.9).



❖ Algorithm

- ❖ C++ implementation
- ❖ Variation on Levenshtein distance algorithm
 - ❖ Added weights
 - ❖ Only considers strings of equal length

❖ Data

- ❖ Data downloaded from siRecords²
- ❖ Used only data for siRNA targeting HEK (human embryonic kidney) mRNAs.
- ❖ Four levels of efficacy
 - ❖ 4=Very High
 - ❖ 3=High
 - ❖ 2=Medium
 - ❖ 1=Low

Position: ...09|10|11|12|13|14|15|16|17|18...
Training string: ...UCCAUA**GUAG**...

Test string: ...AACGU**U**CGGU...

1. Position of
nucleotide

2. Type of
nucleotide
substitution

Modified algorithm

2. http://sirecords.umn.edu/siRecords/download_data.php

Algorithm Details



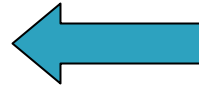
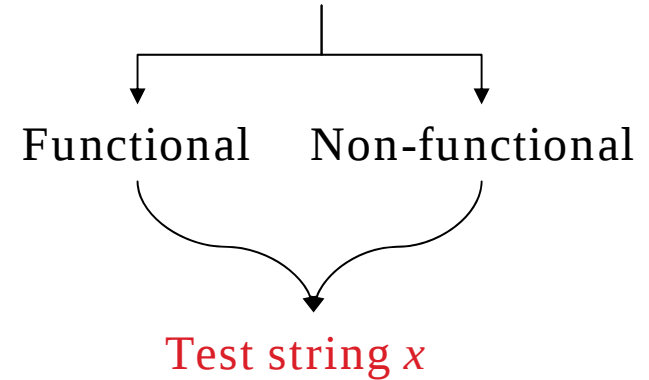
Training Set

Functional	Non-functional
...	...
GAUUACGUCGCCAGUCAAG	AUGAUCCAGGGCGGAGACU
GAGGAUAUGUUUGCAUGGC	UAUUUGGAUCCAUGGUGUA
ACCGAAAGGUCUUACCGGA	UGACAACUUUGGUAUCGUG
CUUACCGGAAAACUCGACG	CAGGGGGGAGCCAAAAGGG
UACCGGAAAACUCGACGCA	AAAACAGUGGAUAAUUUUG
AGGUCUUACCGGAAAACUC	CUUUGGCCUCUGCAAAGAG
CCGGAAAACUCGACGCAAG	UAAUCGGUCCGAAAGAGA
UUACGUCGCCAGUCAAGUA	AAACUAUCGUACCAUGGAA
...	...

Average distance
functional

Average distance
non-functional

Testing Set

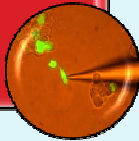


Absolute value:
Smaller value wins

Results

- 61 total functional strings (efficacy = 1)
- 120 total nonfunctional strings (efficacy = 4)

25 splits of
the HEK data



- Matlab algorithm to randomly split data into training and test sets
 - 30 functional training
 - 60 nonfunctional training

Data splitting

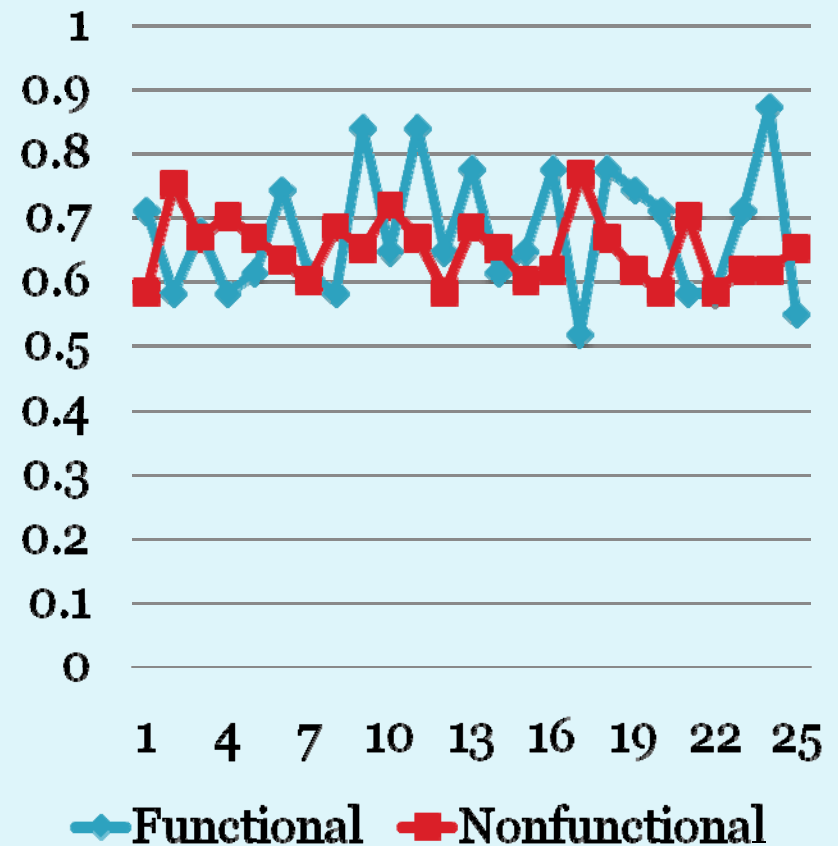


- Functional: 67.6%
- Nonfunctional: 65.0%
- Overall: 65.9%

Average
accuracy



Algorithm Accuracy



Results Details



Dataset	F	NF	%F	%NF	Overall
1	22	25	0.709677419	0.583333333	0.626373626
2	18	15	0.580645161	0.75	0.692307692
3	21	20	0.677419355	0.666666667	0.67032967
4	18	18	0.580645161	0.7	0.659340659
5	19	20	0.612903226	0.666666667	0.648351648
6	23	22	0.741935484	0.633333333	0.67032967
7	19	24	0.612903226	0.6	0.604395604
8	18	19	0.580645161	0.683333333	0.648351648
9	26	21	0.838709677	0.65	0.714285714
10	20	17	0.64516129	0.716666667	0.692307692
11	26	20	0.838709677	0.666666667	0.725274725
12	20	25	0.64516129	0.583333333	0.604395604
13	24	19	0.774193548	0.683333333	0.714285714
14	19	21	0.612903226	0.65	0.637362637
15	20	24	0.64516129	0.6	0.615384615
16	24	23	0.774193548	0.616666667	0.67032967
17	16	14	0.516129032	0.766666667	0.681318681
18	24	20	0.774193548	0.666666667	0.703296703
19	23	23	0.741935484	0.616666667	0.659340659
20	22	25	0.709677419	0.583333333	0.626373626
21	18	18	0.580645161	0.7	0.659340659
22	18	25	0.580645161	0.583333333	0.582417582
23	22	23	0.709677419	0.616666667	0.648351648
24	27	23	0.870967742	0.616666667	0.703296703
25	17	21	0.548387097	0.65	0.615384615
Average:			0.676129032	0.65	0.658901099
Std. Dev:			0.098760968	0.051594932	0.038784619



❖ Accuracy

- ❖ Approximately 66% accurate

❖ Matthews Correlation

- ❖ Between -1 and +1
 - ❖ 0 => random

- ❖ Our Matthews Correlation: 0.31

❖ Other results?

- ❖ Results comparable to other papers (see table)
- ❖ Performed better than 7 of 9 algorithms tested.

Matthews Correlation:

$$M = \frac{Tp \cdot Tn - Fp \cdot Fn}{\sqrt{(Tn + Fn)(Tn + Fp)(Tp + Fn)(Tp + Fp)}}$$

Table 4. siRNA results

Algorithm	ROC	<i>R</i>	<i>P</i>	<i>M</i>	Se	Sp
GP	0.62	−0.27	9.4×10^{-5}	0.19 ± 0.09	0.45	0.73
GPboost	0.70	−0.42	4.0×10^{-10}	0.27 ± 0.18	0.56	0.70
GPboost _{Reg}	0.72	−0.46	4.5×10^{-12}	0.24 ± 0.24	0.50	0.73
<i>v</i> -SVM(seq)	0.57	−0.05	0.48	0.09 ± 0.23	0.52	0.56
<i>C</i> -SVM(seq)	0.53	−0.02	0.78	0.05 ± 0.20	0.35	0.70
<i>v</i> -SVM(1–2)	0.70	−0.21	2.6×10^{-3}	0.31 ± 0.19	0.61	0.68
<i>C</i> -SVM(1–2)	0.64	−0.30	1.3×10^{-5}	0.26 ± 0.14	0.45	0.80
<i>v</i> -SVM(4)	0.62	−0.14	0.046	0.25 ± 0.16	0.49	0.73
<i>C</i> -SVM(4)	0.65	−0.30	1.3×10^{-5}	0.19 ± 0.20	0.35	0.82

Interpretation

Saetrom P: “Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming”. *Bioinformatics*. 2004, 20:17.



❖ Vert's weight data is generated from both mouse and human sources

- ❖ What is functional for mice may not be functional for humans



An issue with data

Future Work



- Incorporate thermodynamic data from Vert into algorithm for additional accuracy
- Test more data sets
- Find more relevant weights

Acknowledgements



- Dr. Stanley Dunn
- DIMACS/DIMATIA REU, NSF
 - REU 2007 participants