An Algorithm for Functional siRNA Detection

Liyang Diao, Wake Forest University
Dr. Stanley Dunn, Biomedical Engineering, Rutgers University

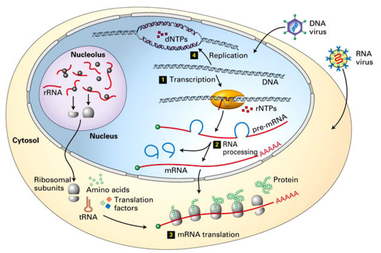
Abstract

Short-interfering RNA, or siRNA, has much potential in the area of gene silencing. The current problem lies in determining which strings of siRNA are effective in suppressing certain genes and which are not: Although siRNA are relatively short, ranging from 20 to 25 nucleotides in length, even for strings of length 20 nt there are 4^{20} number of potential strings of siRNA, only a few of which may actually be useful. We present an algorithm based on a weighted Levenshtein distance that calculates distances between siRNA such that the distance between functional strings is less than distance between nonfunction and functional strings. The algorithm was implemented via C++ and produced results yielding a Matthews Correlation for significance of 0.31.

Background

Information encoded in DNA inside a cell's nucleus is the blueprint of protein production in the body, but the actual assembly of proteins occurs in the endloplasmic reticulum within the cytoplasm, so the transferring of information from DNA requires two major processes: transcription and translation. The former is the process whereby DNA is essentially copied to single-stranded mRNA, which can travel through the nuclear membrane; the latter the process occurs via tRNA and is essentially the process whereby proteins are constructed following the instructions on the mRNA. RNA interference (RNAi) is a disturbance of this process by dsRNA, which is cleaved by the dicer enzyme to produce short double-stranded fragments of 20-25 nt. The most well-known process of RNAi involves double-stranded fragments called siRNA and falls within the class of post-transcriptional gene silencing. The siRNA cleaves key portions of the mRNA before it reaches the RNA translation site, thus effectively silencing the genes which would otherwise be encoded by the cleaved areas of mRNA.

Figure 1



http://fig.cox.miami.edu/~cmallery/255/255hist/mcb4.1.dogma.jpg

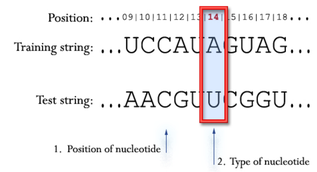
Protein production in the cell: Transcription (1) leads to mRNA synthesis; mRNA travels out of nucleus to protein assembly site (3), where translation occurs: tRNA binds to mRNA, assembling strings of amino acids, which stops at stop codon; protein is released. siRNA-mediated RNAi occurs before translation.

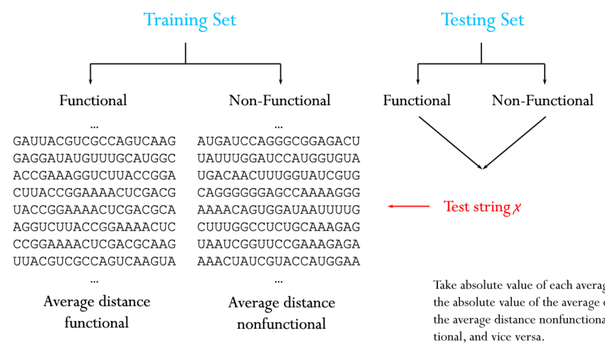
A particular benefit of RNAi is the relative scarcity of unseen side effects, as the target area of siRNAs tend to be quite limited, but the actual mechanism by which this interference occurs is poorly understood. Thus, the ability to predict efficacy of siRNAs is highly desirable, and a great deal of research has been pursued to this end. In previous summers Karen Lostritto, currently at Yale, considered Pancoska's Eulerian graph model for predicting functional siRNAs in depth[1]; in this model the structure of siRNAs was represented as Eulerian graphs of four vertices, first as a directed digraph and then as a weighted, undirected Eulerian graph, but there were several issues concerning the representation including the uniqueness of proposed representations. Thus, we turned to other methods of determining efficacy. Among these was a combinatorial approach which utilized certain characteristics positively associated with gene suppression, including the differential between thermodynamic stability of the 5' and 3' ends, anc a GC content between 30% and 55%. After filtering possible siRNA through these characteristics we were able to narrow down the pool of candidates to 25% of original size, but this was still a far too large a pool to contend with.

Algorithm and Data

The proposed algorithm is a modified Levenshtein distance; Levenshtein distance is a string metric which measures the shortest distance between two strings, where the operations are insertions, deletions, or substitutions of characters. For a simple example, the Levenshtein distance between Sunday and Tuesday is three; particularly, "S" is substituted by "T", "e" is inserted after "u", and "n" is substituted by "s". In theory, the purpose of our algorithm is to differentiate functional from nonfunctional siRNA by defining a distance between strings of siRNA in such a way that the distance between strings of functional siRNA is shorter than the distance between a string of functional siRNA and a string of nonfunctional siRNA. Likewise, the distance between strings of nonfunctional siRNA should be shorter than between strings of functional versus nonfunctional siRNA. However, applying the Levenshtein distance to our data showed that it did not significantly distinguish functional from nonfunctional strings. Thus, we modified the Levenshtein distance first by eliminating the insertion/deletion operations, as our data consisted of strings of constant (19 nt) length, and additionally incorporated weights so that each nucleotide substitution was weighted differently based on the position of the substitution and which nucleotides were involved. The weights were taken from Vert's 2006 paper[2]. Figure 2 (right) displays an example of the general Levenshtein algorithm, along with pseudocode.

Figure 3





Take absolute value of each average distance and compare size: If the absolute value of the average distance functional is less than the average distance nonfunctional, then x is classified as functional, and vice versa.

Our data was downloaded from siRecords[3], which contained various types of siRNA, efficacy levels, and targets. We chose to use only data concerning siRNA which targeted HEK mRNA. These strings were all 19 nt in length and were already discretized into four efficacy classes ranging from low to very high. To test our algorithm, we considered only data of low, high, and very high efficacy levels. We cross-validated our results by randomly generating training and testing sets. In total there were fifty training sets of 30 functional 60 nonfunctional strings, and fifty testing sets of the same number of functional and nonfunctional strings.

Figure 3 is a visual representation of the algorithm. The data is divided into four classes: functional training, nonfunctional training, functional testing, and nonfunctional testing. The training sets are strings of siRNA whose functionality is known; test strings are siRNA used to test the accuracy of the algorithm. The algorithm is implemented in C++ and at each position compares the nucleotide of the test string against the nucleotide of the training string and computes the difference of their weights from Vert's paper. We test each test string x against the set of functional training strings and the set of nonfunctional training strings separately and calculate the averages of these distances. The smallest average (after taking absolute value) determines whether x is classified as functional or nonfunctional.

Figure 4

Often in computing Levenshtein distance an $(n+1)$ x $(m+1)$ matrix is used, where n and m are the lengths of the strings being compared. A variable *cost* is defined such that if the ith letter of string1, s[i], is equal to the jth letter of string2, t[j], then cost = 0; otherwise cost = 1. As we have $i = 0..n$ and $j = 0..m$, at each step we form the matrix M_{ij} so that the element d[i,j] is the minimum of

(1) d[i-1, j] + 1
(2) d[i, j-1] + 1
(3) d[i-1, j-1] + cost

where (1) represents the cost of a deletion, (2) represents the cost of an insertion, and (3) represents the cost of a substitution (which is zero if s[i] = t[j]). The final distance is the value of d[n,m]. The minimum values (left) are set in red.

		k	i	t	t	e	n
	0	1	2	3	4	5	6
s	1	**1**	2	3	4	5	6
i	2	2	**1**	2	3	4	5
t	3	3	2	**1**	2	3	4
t	4	4	3	2	**1**	2	3
i	5	5	4	3	2	**2**	3
n	6	6	5	4	3	3	**2**
g	7	7	6	5	4	4	**3**

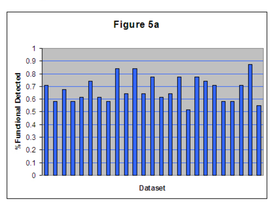
Conclusions

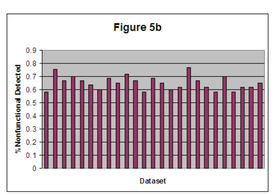
There were a total of 61 functional strings and 120 nonfunctional strings of siRNA pulled from the database and in order to utilize cross-validation, we created twenty-five splits of the data, where each split of data produced 30 functional training strings and 60 nonfunctional training strings, and 31 functional testing strings and 60 nonfunctional testing strings. We treat the testing strings as if we do not know their efficacy and use the algorithm to predict it. Some results from the algorithm are given in Figure 5; the graph shows the results computed by measuring the average percentage of correctly predicted functional siRNA per data split and the average percentage of correctly predicted nonfunctional siRNA per data split.

$$(1) \qquad M = \frac{\mathrm{Tp} \cdot \mathrm{Tn} - \mathrm{Fp} \cdot \mathrm{Fn}}{\sqrt{(\mathrm{Tn}+\mathrm{Fn})(\mathrm{Tn}+\mathrm{Fp})(\mathrm{Tp}+\mathrm{Fn})(\mathrm{Tp}+\mathrm{Fp})}}$$

Additionally we use the Matthews Correlation (Equation 1) to compute the accuracy of the algorithm's predictions taking into consideration the number of true positives, true negatives, false positives, and false negatives. This correlation coefficient is often used in measuring the effectiveness of an algorithm given a binary classification. The value of the coefficient varies from -1 to 1, where 0 implies a random classification, 1 implies a perfect classification, and -1 implies a worse than random classification. Our Matthews Correlation is 0.31, a coefficient which is quite typical in modeling siRNA efficacies. In fact, it compares favorably against seven of nine algorithms presented in Saetrom[4].

Figure 5





The average accuracy of correctly predicted functional siRNA is 67.6% while the average accuracy of correctly predicted nonfunctional siRNA is 65.0%. There are thirty-five more datasets to run through the algorithm, and in addition to these datasets we would like to be able to incorporate thermodynamic data (also in Vert) into the algorithm.

Acknowledgements

I would like to thank Dr. Stanley Dunn at the Rutgers University who worked with the Rutgers DIMACS/DIMATIA REU, without whose support and guidance this project would have been impossible. I would also like to thank the NSF, which funds the DIMACS, DyDAn, and Rutgers University Math REU programs. The REUs are a joint project of Rutgers, Princeton, AT&T Laboratories, Bell Laboratories, Telcordia Technologies, and the NEC Research Institute.

[1] http://dimax.rutgers.edu/~karenlos/
[2] Vert JP, Foveau N, Lajaunie C, Vandenbrouck Y: "An accurate and interpretable model for siRNA efficacy prediction". BMC Bioinformatics. 2006, 7:520.
[3] http://sirecords.umn.edu/siRecords/download_data.php
[4] Saetrom P: "Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming". Bioinformatics. 2004, 20:17.