

An Algorithm for Functional siRNA Detection

Liyang Diao, Wake Forest University

Dr. Stanley Dunn, Rutgers Department of Biomedical Engineering

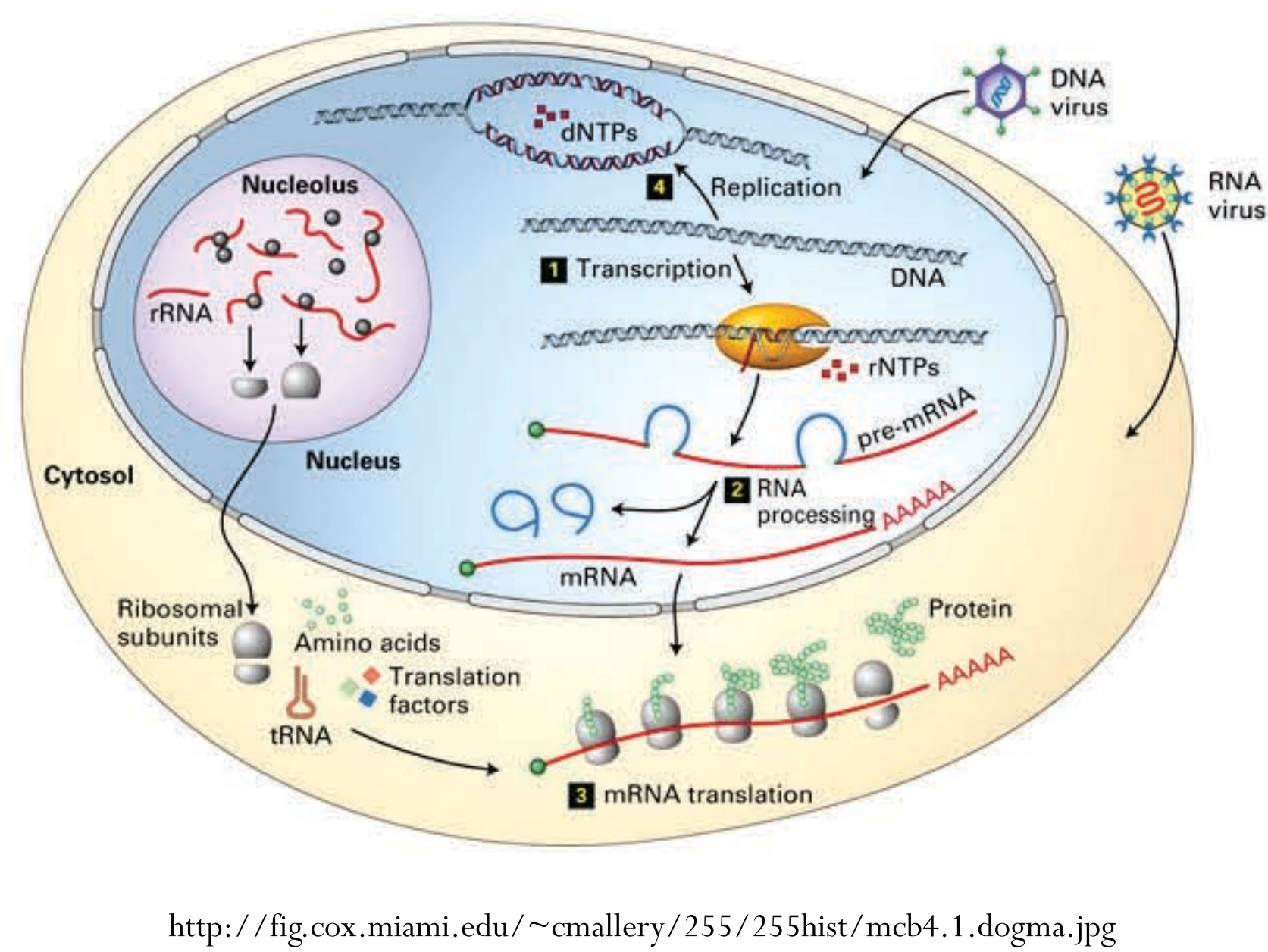
Abstract

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 a C++ program and preliminary results indicate a Matthews correlation 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 endoplasmic 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 nt¹. 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

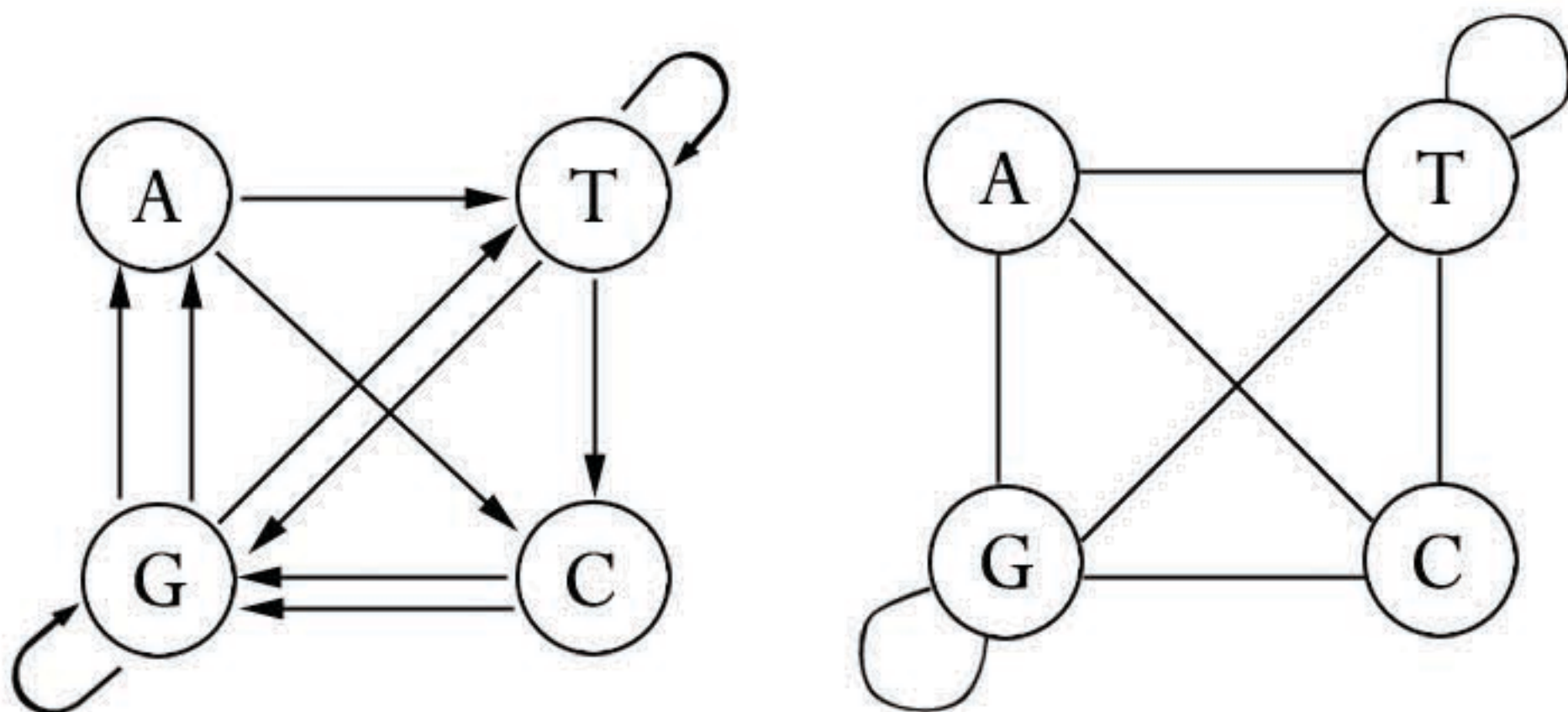


<http://fig.cox.miami.edu/~cmallery/255/255hist/mcb4.1.dogma.jpg>

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 of Brown University considered Pancoska's Eulerian graph model for predicting functional siRNAs in depth; 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:

Figure 2



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 non-functional 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 paper³.

Figure 2

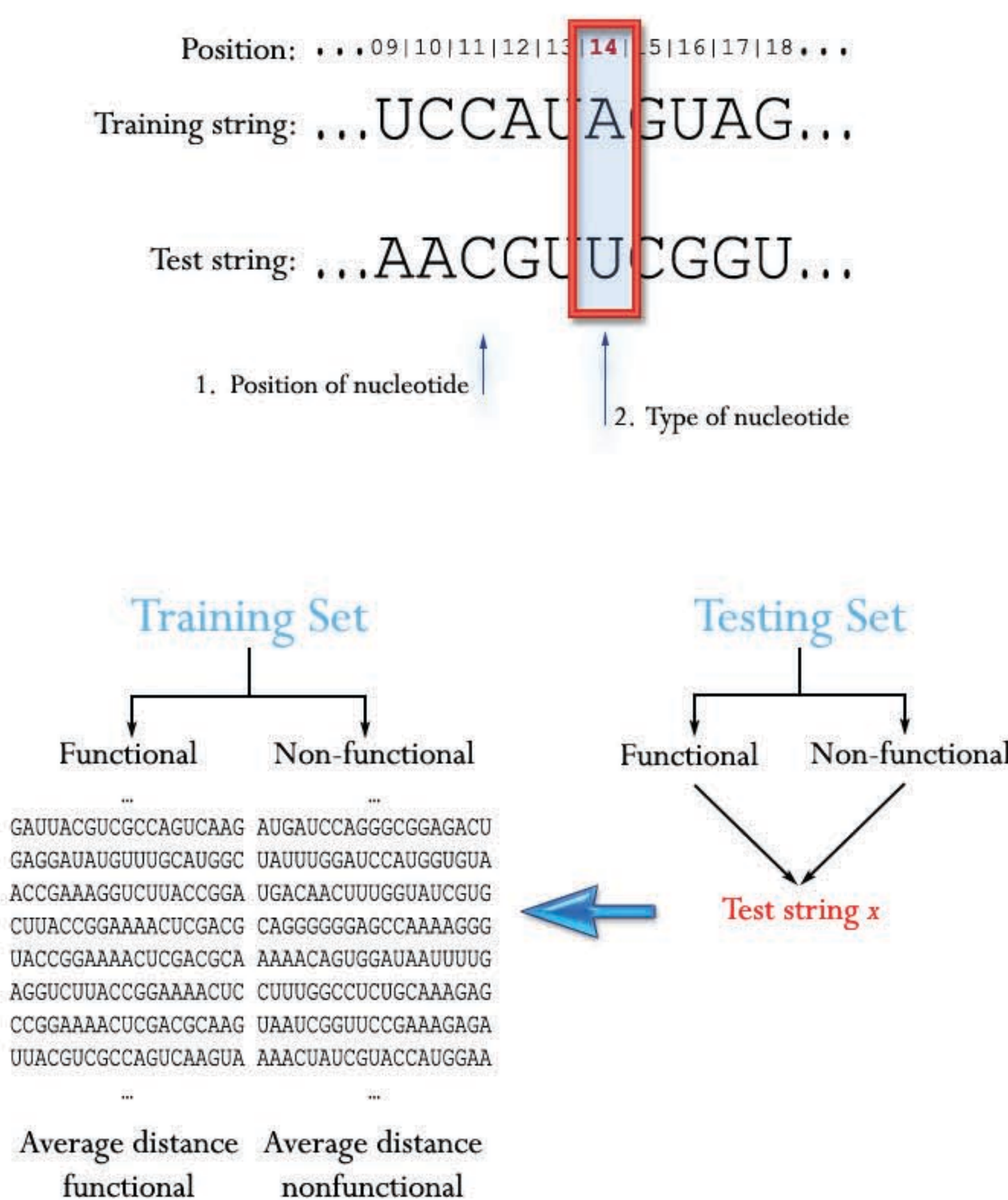


Figure 2 is a visual representation of the algorithm. The data is divided into four classes: functional training, nonfunctional training, functional testing, and nonfunctional testing. The training sets are strings of siRNA whose functionality is known; test strings are siRNA used to test the accuracy 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.

Our data was downloaded from siRecords², 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.

¹<http://en.wikipedia.org/wiki/Rnai>

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

³Vert JP, Foveau N, Lajamie C, Vandenbrouck Y: "An accurate and interpretable model for siRNA efficacy prediction". BMC Bioinformatics. 2006, 7:520.