Technical Report
Predictive models to predict knockdown efficacy of siRNAs

BUI NGOC THANG

January 11, 2018

1 Research Objectives

- Finding characteristics of sequential data
- Building a good representation of sequential data
- Transforming siRNA sequences to a new feature space
- Applying a regression method to learn a predictive model on the new feature space

2 Finding characteristics of sequential data

Many research groups have proposed models to predict knockdown efficacy of siRNAs. However, the performance of existing models is very low. To my knowledge, researchers did not consider the following issues. Firstly, some important features were not found out to represent data. Secondly, siRNA sequences were generated from different protocols in wet-labs, so they may be generated from different distributions in term of statistical view point. Therefore, it is necessary to find out a better representation of siRNA sequences that can overcome above mentioned issues.

We consider the following characteristics of sequential data

- Encoding of each nucleotide on the sequence: represented by one-hot vector representation
- Borrowing ideas of word to vector representation from Nature Language Processing field: using n–gram, l–skip–k–gram, and term frequency–inverse document frequency (TF–IDF) of n–gram, l–skip–k–gram to represent sequences.
- Considering that a sequence is generated by a First Order Markov chain model. Therefore, transition probabilities of the model are computed and represented as features of the sequential data.

- Analyzing the information correlation of nucleotides or k–mers/k-grams in order to use for siRNA representation.

3 Partial representations of siRNA sequence

3.1 One-hot vector representation

Nucleotides A, C, G, and U on a sequence are encoded by binary vectors: $< 1, 0, 0, 0 >, < 0, 1, 0, 0 >, < 0, 0, 1, 0 >$, and $< 0, 0, 0, 1 >$, respectively.

3.2 TF–IDF of l–skip–k gram/k mers

Firstly, l–skip–k mers are generated. We know that the number of k–mers is 4^k. However, k–mer represented sequential data is sparse. In our problem, the length of each sequence is very short so the data represented by k–mer become very sparse. It can lead to lack of some important k–mers features. Therefore, l-skip-k mers are suitable to represent short sequences. Note that the number l–skip–k–mers are the number of k–mers are the same. After obtaining l–skip–k–mers, term frequencies–inverse document frequencies (TF–IDF) are calculated to estimate how important each l–skip–k–mer contributes to a sequence. They are calculated by the following formulas:

$$TF(t,d) = \frac{freq(t|d)}{max\{freq(t'|d) + 1 : t' \in d\}}$$

$$IDF(t,D) = \log(\frac{|D|}{|\{d \in D : t \in d\}|})$$

$$TF - IDF(t,d) = TF(t,d) * IDF(t,D)$$

where TF(t,d) denotes the relative frequency of mer t in the sequence d and IDF(t,D) denotes the inverse relative frequency of sequences in the dataset D having the mer t.

3.3 First Order Markov chain model

Based on the Markov assumption, we computed transition probabilities (conditional probability) of each nucleotide given previous one. This formulas is as follows.

$$P(Y|X) = \frac{P(XY)}{P(X)}$$

where X and Y denote nucleotides satisfying that Y follows X in the sequence. $P(XY)$ and $P(X)$ denote the probabilities of the mers $''XY''$ and $''X''$ in the sequence.

3.4 The information correlation

The information correlation of nucleotides in sub–sequences is estimated to employs the base correlation property of RNA sequence. The information correlation of sub–sequences with size l is the following entropy formula:
$IC_\ell = -2\sum_i P_i \log_2 P_i + \sum_{ij} P_{ij}(\ell) \log_2 P_{ij}(\ell)$

The P_i and P_j denote the probabilities of nucleotides $''i''$ and $''j''$ in the sequence. The $P_{ij}(\ell)$ indicates the probability of bases $''i''$ and $''j''$ at distance l in the sequence.

We combine all of above mentioned representations to represent siRNA sequences. On this new space, we applied Rigde regression method to learn a model on the Huesken training set of 2182 siRNAs. This model was tesed on Husken testing set of 249 siRNAs and three independent data sets: Reynold, Vicker, and Harborth consist of 244, 76, 44 siRNAs, respectively.

The results shown that the performance of our model is better than that of 19 previous models when tested on Harborth and Vicker data sets and are similar to that of these models when tested on the other data sets.