Modified PSWM for RNA editing sites search

SHTOKALO D., NECHKIN S., EREMINA T., CHEREMUSHKIN E., ST. LAURENT G.
Ershov Institute of Informatics Systems
Lavrentjiev ave. 6, Novosibirsk, 630090
RUSSIA
evgeny@iis.nsk.su http://www.iis.nsk.su

Abstract: - The computational method for RNA editing sites search is described. In RNA sequence ADAR ferments convert adenosine (A) into inosine (I), which in turn is read as guanosine (G), and increase transcriptomic diversity. We developed modification of PSWM method that includes information about pairwise correlations between nucleotides in ADAR editing site context. This method was used for *in-silico* discovery of potential RNA editing sites in deep sequencing RNA data.

Key-Words: - ADAR editing, context score, matrix, position weight matrices, PSWM, RNA editing

1 Introduction

Protein composition in cell is regulated by different mechanisms as transcription[1], splicing[2], mRNA stability[3], translation[4,5]. In spite of mechanisms diversity, the transcription regulation is the most widely used and studied [6,7]. It includes specific proteins called transcription factors that bind to cognate DNA fragments, which create protein complex for RNA-polymerase binding. Recently, with development of mass RNA sequencing technology[8,9] it became possible to research post transcriptional regulation mechanism, where in particular ADAR ferments change adenosine (A) into inosine (I) and increase protein diversity, regulate alternative splicing and RNA stability[10,11].

A Position in RNA where ADAR replace A with I is called ADAR editing site. Context area of ADAR editing site was shown to have specific nucleotides in specific positions [12].

Position-specific weight matrices (PSWMs) are widely used for transcription factor binding sites (TFBS) recognition[13]. Our group has developed a modification of PSWM model that counts correlation between nucleotides in different positions, called "correlation matrix" (CM). First, we introduce simple PSWM model and then continue with CM model. PSWM is a 4xN matrix. Each element of PSWM corresponds to probability of meeting of given nucleotide in corresponding position in the set of known binding sites (Fig 1).

1	A	A	G	G	T
2	A	C	G	G	A
3	A	A	T	G	C
4	C	A	T	G	G
5	A	A	T	G	G

M

A	0.8	0.8	0	0	0.2
C	0.2	0.2	0	0	0.2
G	0	0	0.4	1	0.4
T	0	0	0.6	0	0.2

Figure 1. Artificially generated example of PSWM construction. There are five known sites on the top and generated matrix on the bottom. Elements of the matrix correspond to occurrence frequencies of given nucleotide in corresponding position of site. For example A appear at the first position in 4 cases out of 5, therefore M["A",1]=0.8.

Thus, for any sequence of a length N a weight w can be calculated, that characterizes similarity of this sequence to a set of known sites. This weight is calculated as a sum of corresponding weights in each position and then normalized from 0 to 1. Considering this a fragment of sequence is counted as site when its weight is higher then given threshold.

Weight matrices became effective for TFBS search. But one of disadvantages of this method is that it does not consider correlations between nucleotide frequencies in different positions. It gives robustness but decreases sensitivity of a method. In the case of ADAR the context of editing sites doesn't have enough positions with prominent nucleotide frequencies therefore more sensitive model should be used to recognize site by given context. We have developed a context score assessment model that uses correlations between nucleotides. This model is described below.

2 Algorithm Description

Correlation matrices represent more sensitive model for recognition of sites with high degree of diversity. This model is represented in Fig. 2.

1	A	A	G	G
2	A	C	G	G
3	A	A	T	G

4	C	A	T	G
5	A	A	T	G

1	.	A	T	.	0.6
2	.	.	T	G	0.6
3	.	A	.	G	0.6
4	A	.	.	G	0.6

Figure 2. Example of correlation matrix. On the top there is a set of known sites. On the bottom in each line there is a pair of nucleotides and estimated value of probability of this pair to occur. For example 0.6 in the first line corresponds to frequency of meeting A and T in positions 2,3 respectively.

Weight of sequence by correlation matrix is calculated as a sum of weights in lines, where pair of letters coincide with corresponding letters on sequence. For example, for CM represented in Fig 2. weight of sequence ACGG will be calculated as $w=(s-w_{min})/(w_{max}-w_{min})$. $w_{max}=(1+.6*4)=3.4$; $w_{min}=0$: $s=0.6$; So $w=0.17$. Therefore CM similarly to PSWM can be used for recognition of functionally significant sequence fragments.

Before building CM, on the first step, conserved positions of sites context where chosen. Those positions have been used for constructing CM, and other positions were cut out consideration.

Log p-value estimation was used for estimation of weight of each pair. Let T is a test set containing experimentally confirmed sites with flanks, and B is a background set containing control non-sites with flanks.

Let P_i be a pair of nucleotides for which p-value will be calculated. Let s_T is a variate, denoting number of sites that contain P_i in T. Let s_T is from binomial distribution. Let N_T and N_B are numbers of sites that contain P_i in sets T and B respectively. Then p-value$^+$ corresponds to probability of P_i hits not less than N_T strings in the set T.

$$pvalue^+ = \mathrm{P}(s_T \geq N_T) = \sum_{i=N_T}^{N} C_N^i p_T^i p_B^{N-i}$$

Where $N=N_T+N_B$, $p_T=N_T/N$, $p_B=N_B/N$. And p-value$^-$ corresponds to probability of P_i hits not more than N_T strings in the set T is

$$pvalue^- = \mathrm{P}(s_T \leq N_T) = \sum_{i=0}^{N_T} C_N^i p_T^i p_B^{N-i}$$

Weight w_i of pair P_i is set as

$$w_i = \begin{cases} -\log(pvalue^+), if\ \ pvalue^+ < Threshold^+, \\ \log(pvalue^-), if\ \ pvalue^- < Threshold^-. \end{cases}$$

Threshold$^+$ and Threshold$^-$ to be fitted. Weight w_i is whether positive if P_i occurs frequently or negative if P_i occurs rarely. Considering this weight choose most frequent and most rare correlation pairs $P_1,\ldots,P_K$. These pairs and their weights will constitute correlation matrix.

3 Potential ADAR sites discovery

To discover novel ADAR editing sites (a) CM was built, (b) RNA sequences with A/G changes were selected, and (c) sequences from step b are filtered by CM

3.1 CM build

Mathematical description of CM building algorithm was given above. For building CM a set of experimentally approved sites from [14] was taken with 40nt flanks.

As a control set we used 1000 random gene sequences. Thus, using those two sets CM was build. This CM is represented in supplementary data available from http://nprog.iis.nsk.su/supplementary1.doc.

We tested this CM on a set of control sequences that contains randomly chosen gene regions. With fixed recognition cutoff on a target set and on a control set FP and FN errors can be evaluated. We choose cutoff to minimize sum of FP and FN. Obtained FP=0.195 and FN=0.259. To make sure CM model is not overfitted we check it loose recognition power if replace set T with any set of random sequences.

3.2 Pipeline for novel sites discovery

Initial set of RNA fragments was obtained from deep sequencing of human brain cells by Helicos. Data was provided by St. Laurent Institute and contains about 50 millions of short RNA fragments of length 30-120nt. Those fragments called “reads”. This information block was strictly filtered in 5 steps as described below to cut false positive predictions.

1) Keep only reads aligned to gene regions.

2) Find reads with A/G replacements i.e. in genomic sequence fragment in a position there is ‘A’, but at least in 5% of read aligned to this fragment in the position there is ‘G’.

3) Filter out known SNP.

4) Apply CM filter.

5) Apply EST filter that allows selection of sequences that have DNA to RNA A->G replacement in EST database.

As a result of the pipeline 950 potential sites were obtained represented in http://nprog.iis.nsk.su/supplementary2.xls. 213 of them are highly conservative among species. These sites are potential candidates to be real ADAR editing sites which were a goal of this work.

4 Conclusion

In result of this research the algorithm of *in-silico* identification of ADAR editing sites was developed. The algorithm was used for prediction of novel potential ADAR editing sites in Helicos deep sequencing data.

Identification algorithm is based on using so called correlation matrices (CMs) that in contrast to PSWMs consider correlations between pairs of nucleotides. CM model uses mechanism similar to PSWM model: sequence weight is calculated and then compared with given cutoff. Cutoff was selected and stability of algorithm was evaluated.

Using Helicos device, about 50 million of short RNA fragments were collected. Using those fragments 950 potential ADAR editing sites were predicted. These sites can be used further in experimental work.

References:

[1] Beckett, D. Regulated assembly of transcription factors and control of transcription initiation. *J. Mol. Biol.,* Vol. 314, 2001, pp. 335–352.

[2] Singh, R. RNA–protein interactions that regulate pre-mRNA splicing. *Gene Expr.*, Vol. 10, 2002, pp. 79–92.

[3] Shim, J. and Karin, M. The control of mRNA stability in response to extracellular stimuli. *Mol. Cells*, Vol. 14, 2002, pp. 323–31

[4] Kozak, M. Regulation of translation in eukaryotic systems. Annu. Rev. *Cell Biol.,* Vol. 8, 1992, pp. 197–225.

[5] Preiss, T. and Hentze, M. W. Starting the protein synthesis machine: eukaryotic translation initiation. *Bioessays*, Vol. 25, 2003, pp. 1201–11.

[6] Davidson, E.H., et al. A genomic regulatory network for development. *Science*, Vol. 295, 2003, pp.1669–1678

[7] Clyde, D.E., et.al. A self-organizing system of repressor gradients establishes segmental complexity in Drosophila. *Nature*, Vol. 426, 2003, pp. 849–853.

[8] Margulies, M., et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, Vol. 437, 2005, 376–380.

[9] Shendure, J., et. al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, Vol. 309, 2005, 1728–1732

[10] Bass BL. RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem.*Vol. 71, 2002, pp. 817-46.

[11] Nishikura K. Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nat Rev Mol Cell Biol.* Vol. 7(12), 2006, pp. 919-31.

[12] Dawson, TR., et al. Structure and sequence determinants required for the RNA editing of ADAR2 substrates. *J Biol Chem*. Vol. 279(6), 2004, pp. 4941-4951

[13] Kel, A. E. et. al. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research* Vol. 31 2003, pp. 3576-9

[14] Li JB, et al. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science*. Vol. 324(5931), 2009, pp. 1210-3.