Poster H-21

n-Nucleotide Frequency Analysis for Detecting Conserved String Pattern in Mus musculus Genome

Authors:
Fumikazu KONISHI (*RIKEN GSC*)
Aki Hasegawa (*RIKEN GSC*)
Akihiko KONAGAYA (*RIKEN GSC*)

Short Abstract: We have developed a method of detecting conserved string pattern using a correlation coefficient of the progression of the recursive association of the substring frequencies among all the possible nucleotide sequence patterns within the specific size toward significant short nucleotide sequence detections.

Long Abstract:
Genome wide detection of significant sequences by composition is a fundamental tool in post genome biology. We show how a number of sequences detected tasks, including the significant genomic regions, can be performed efficiently without known biological facts. Our approach is based on a correlation coefficient of the recursive association of the substring frequencies among all the possible nucleotide sequence patterns within the specific size toward significant short nucleotide sequence detections. Though DNA sequence is composed by four nucleotide of ATGC and there is bias in the probability in each genome, an appearance of frequency usually lowers when nucleotide sequences get long. And, a combination number of the pattern is the theoretical numbers of L power of 4 as for the nucleotide of the length L. Therefore, the pattern which any genome has used can be supposed selectively on the whole combinations. And the frequency shows a clue about whether the combination of that arrangement should be necessary.
The length which is enough to specify a position on a genome is known with 25mer from 15na at present. And, the sequence makes use of a primer which is a complemental strand of polymerase chain reaction (PCR) to specify target gene in genome. Therefore, application of shortest unique sequence to biological problems has been an active aria of PCR research. In this research, it is not a purpose to find the shortest sequence for PCR, but it is a purpose here to give it new evaluation value to the sequence for finding significant pattern. When the sequence S of a length L (target sequence) is given to it, the frequencies of the substrings which that arrangement composes are arranged by the progression of the length, and the evaluation value is calculated as a coefficient of correlation that expressed how much bias there is in comparison with the theoretical progression value.
Our method of the evaluation value is based on the idea of estimating a curve fitting of a polynomial equation: y=ax+b with substings in a target sequence. Consider for example the sequence S = ACTAGT. It contains six proper prefixes, of which the following six are : {A,AC,ACT,ACTA,ACTAG,ACTAGT}. We determine for every position i in S the length L of the sub-stings S[1..L-1]. We have only considered the forward strand of a given sequence. As for S, the frequency of the substring from S[1..1] to the S[1..L] is arranged, and transform into the values which make the binary logarithm. If the sequence has a feature of randomness, the fitting curve is a line with slope 2 theoretically. However, theory and practice do not always coincide in the value of the measured frequencies from genome

sequence. It is because an organism conserves the sequence which has an important function and it is kept as that reason. Therefore, when it is an indispensable sequence, it has bias between the random sequence in the use frequencies of the substrings, and it can be detected. The detected conserved sequence can be made a catalog with annotating with well-known facts.

We demonstrate the utility of a conserved sequence by applying them to three steps: (i) Generating all the possible sequence patterns from a length of 1 to L. (2) Counting each frequency of the sequence on a target DNA sequence. (3) Transform a sequence into a series of the frequency with every sub-string. (4) Estimate the fitting curve by regression, and calculate a correlation coefficient.

In case of S1 = ACTAGT and S2= GCCGGC for Mus musculus genome, S1 has a series of the frequency as follows. (table.1) The ratio of S[1..x]/S[1..x+1] can be estimated at roughly 1/4 because DNA sequence is composed by four nuclear acid. S1 can read almost on the expectation. In the case of S2, there is significant difference in the ratio between the S1 and the S2. The S2[1..4] is very low value in comparison with other prefix cases.

And, all of the sequences can make a series of frequency for estimating the fitting curve. Finally, the parameter of curve and a coefficient of correlation can be calculated from each line of which a table of series of frequency is. (table.2) And the sequence pattern that a coefficient of correlation and inclination are away from the theoretical value will become a catalog as a candidate of the sequence that an organism conserved through the degree of agreement between theoretical and measurement values. When the sequence of the size of L=10 mer is carried out as a target, as for us, we have a combination number of sigma (n power of 4) [n = 1,..L]. And, the amount of combinations can be reduced by removing the pattern which appeared no more than once. However, if the distributed processing was not used, the counting substrings frequency may be difficult to perform without grid computing when L is bigger than 20. We have built the system which could carry out large-scale calculation. In this poster, we will report about the result that Mus musculus genome was carried out about L=20.

Table 1.

S Prefix frequency ratio (S[1..x]/S[1..x+1])

S1[1..1] A 722255467 -

S1[1..2] AC 132153585 0.18

S1[1..3] ACT 40375876 0.30

S1[1..4] ACTA   8069834 0.20

S1[1..5] ACTAG 1698326 0.21

S1[1..6] ACTAGT 337775 0.20

S2[1..1] G 517767088 -

S2[1..2] GC 101635199 0.20

S2[1..3] GCC 26411224 0.26

S2[1..4] GCCG   1234567 0.05 *

S2[1..5] GCCGG 356441 0.29

S2[1..6] GCCGGC 83646 0.23

table 2

Prefix, n, slope, intercept, r, two-tailed prob

ACTAGT 6 2.185140 18.483492 0.999188 0.000001 0.150475

GCCGGC 6 2.624692 15.976950 0.992489 0.000084 0.552513