VISUALIZATION OF DNA SEQUENCES BY COLOR CUBE TRANSFORMATION

[1]Cheremushkin E., [1,2]Kel A., [3]Lobiv I.V., [3]Murzin F., [3]Polovinko O.
[1]Institute of Cytology and Genetics SB RAS, Novosibirsk
[2]BIOBASE GmbH, Wolfenbuettel
[3]Institute of Informatics Systems SB RAS, Novosibirsk

[1]Cheremushkin E.
[1,2]Kel A.
[3]Lobiv I.V.
[3]Murzin F.
[3]Polovinko O.
[1]Institute of Cytology and Genetics SB RAS
[2]BIOBASE GmbH
[3]Institute of Informatics Systems SB RAS

We have developed several methods of visualization of the primary structure of DNA sequences using a color cube transformation. Using this methods we have analyzed promoters of liver enriched genes. The problem of comparison promoter structures become actual in the course of functional annotation of genomes (Cheremushkin & Kel, 2002). The created color images of the alignment of these promoters help us to identify sub-regions of elevated similarity that may correspond to specific functionally important signals in these promoters.

Methods

Visualization of DNA sequences by color cube transformation.

Let $\overline{S}$ be a sequence in an alphabet consisting of four letters A, C, G, T. The k - th element of the sequence will be denoted by s_k, and M be a length of the sequence.

Suppose a positive number N is given. Denote by $BL_N[i]$ a subsequence of $\overline{S}$ having a length equal to N and beginning of i - th position, i.e. $BL_N[i] = s_i \ldots s_{N+i-1}$.

Let $n_A[i,N]$, $n_C[i,N]$, $n_G[i,N]$, $n_T[i,N]$ be the numbers of letters A, C, G, T in the considered subsequence $BL_N[i]$ respectively. If i, N are fixed we will write for brevity n_A, n_C, n_G, n_T.

It is easy to see that $n_T = N - (n_A + n_C + n_G)$. It means that it is sufficient to study only three components. From here frequencies $p_A = n_A / N$, $p_C = n_C / N$, $p_G = n_G / N$ can be calculated.

There exists the natural function from an interval [0,1] of real numbers onto a set of integers $\{i : 0 \le i \le 255\}$ defined by the formula $f(x) = \mathrm{int}(255 \times x)$. Let us introduce $\overline{p}_A = f(p_A)$, $\overline{p}_C = f(p_C)$, $\overline{p}_G = f(p_G)$. Than a triple $\langle \overline{p}_A, \overline{p}_C, \overline{p}_G \rangle$ may be considered as the components of colors $\langle R, G, B \rangle$ respectively.

The color image may be defined by three matrices $S = (S_R, S_G, S_B)$, $S_R = S_R(i,j), S_G = S_G(i,j)$ $S_B = S_B(i,j)$, $0 \le i \le n-1$, $0 \le j \le m-1$. Usually the values of $S_R(i,j), S_G(i,j), S_B(i,j)$ change from 0 to 255. Where n and m correspond to the width and the height of the image. A set of triples $\{ (r,g,b) : 0 \le r,g,b < 255 \}$ is called the color cube. We make a transformation of the frequencies of nucleotides into colors of the color cube.

The first algorithm of visualization

Let us fix the width and height of the image: *n* and *m*. Suppose that two positions i_1, i_2 on a sequence $\overline{S}$ are given, $i_1 - i_2 \le n \cdot m$ and $i_1 \le k \le i_2$. Let us consider a window $BL_N[k]$ of the size *N* that is moving along the given sequence $\overline{S}$. Then, we compute the corresponding triple $\langle \overline{p}_A, \overline{p}_C, \overline{p}_G \rangle$ for every position k.

Therefore we write

$$\langle \overline{p}_A, \overline{p}_C, \overline{p}_G \rangle = \langle \overline{p}_A(k), \overline{p}_C(k), \overline{p}_G(k) \rangle = \langle R(k), G(k), B(k) \rangle .$$

Now we can construct the following image

$$S_R(i,j)=\begin{cases}R(i_1+m\cdot i+j-1) & if\ i_1+m\cdot i+j-1\le n\cdot m;\\ 0 \quad otherwise;\end{cases}$$
$$S_G(i,j)=\begin{cases}G(i_1+m\cdot i+j-1) & if\ i_1+m\cdot i+j-1\le n\cdot m;\\ 0 \quad otherwise;\end{cases} \qquad (1)$$
$$S_B(i,j)=\begin{cases}B(i_1+m\cdot i+j-1) & if\ i_1+m\cdot i+j-1\le n\cdot m;\\ 0 \quad otherwise.\end{cases}$$

It means that we fill the pixels in the image in a process of obtaining the components $\langle R,G,B\rangle$. At the beginning we are filling the upper row, i.e. $i=0$. Then we are filling the first row and so on. If the number of necessary components of color is not sufficient, then we fill $\langle 0,0,0\rangle$, i.e. the black color.

Used algorithms are described in the enlarged report.

Results

The developed technique can be used for analysis of the sets of aligned regulatory sequences to present visually the information about similarity of local regions, their nucleotide compositions and overall sequence structure.

In future the interactive techniques of using such images for navigating through the sets of aligned sequences and for finding potential signals in local regions will be developed. Several image analysis and image recognition approaches can be applied for comparison of sets of sequences.