**CONTEXT-FREE METHOD OF PATTERN RECOGNITION IN THE GENETIC TEXTS.**

AKBEROVA N., KOLPAKOV A., LEONTIEV A.

Kazan State University, Nucleic acids biochemistry lab.

Address: 420008 Kazan, Kremlevskaya, 18

Telephone: (8432) 7679 2; (8432) 429138

e-mail: aleontie@ksu.ru, nakberov@ksu.ru

Keywords: pattern recognition, genetic text, replication origin, object-oriented language

**Abstract**

A computer system for pattern recognition in DNA sequences has been elaborated and tested on a set of prokaryotic replication origins. The image of the origin is being constructed on the basis of the predicate calculus using special object-oriented langua ge. Experimentally established features of the origins' sequences were used to build an image which is called "experimental". Symmetry properties of the sequences under consideration made the basis for the "formal" context-free image of origins. The im ages of two types were compared in order to reveal those symmetry properties of the sequences which are important for pattern recognition in genetic texts. The predictive value of "formal" image has been evaluated in comparison to "experimental" image an d other methods of pattern recognition.

**Image construction**

Both "experimental" and "formal" images are considered as a combination of elements - subsequences having some biological or formal sense - and positional relationships between elements. For the description of images following relationships between elemen ts were used: 1) s1 is a subsequence of s2, 2) s1 is n nucleotides upstream (downstream) to s2, 3) s1 overlaps s2, 4) s has a length of n nucleotides. Three 13-mers (R -gatctcttattag, M - gatctgttctatt, and L - gatctatttattt) and the DnaA binding site (t twtncaca) were considered as elements of "experimental" image. Some restrictions on their positions known from literature were also used. The "formal" image is being built using the same possible relationships between subsequences of the origin text. As t he elements of this image the symmetrical structures in the genetic text have been considered. The symmetries of the text which were used for construction of the "formal" image were described earlier in [1]. They are called "colored" symmetries according to the crystallographic tradition [2]. The symmetries taken into consideration are based on the following transformations of the DNA alphabet: {a <- t, c <- g}, {a <- g, c <- t}, {a <- c, g <- t}. Along with the translation (in crystallographic sense of the word) and inversion these transformations form 8 types of symmetry structures. For a string 'aagct' these structures look like :

|  |  |
| --- | --- |
| Direct | Inverted |
| 1) aagct...aagct | 2) aagct...tcgaa |
| 3) aagct...ttcga | 4) aagct...agctt |
| 5) aagct...ggatc | 6) aagct...ctagg |
| 7) aagct...cctag | 8) aagct...gatcc |

**Discussion**

Sets of image elements and relationships were found to be sufficient to built an effective recognition system.

The comparison of "experimental" and "formal" images reveals the fact that some of non trivial symmetry structures (not only direct repeats) overlap with the biologically meaningful elements of the replication origin. Though the biological sense of such structures is yet unknown, one can suppose that they may play some role in protein-DNA interactions.

Though it is difficult to prove the statistical value of our results, we believe that there is significant difference in frequences of occurance of different types of symmetries. Direct and inverted repeats of the common type, as well as the symmetries involving the transformation {a <- t, c <- g} are the most frequent in replication origin sequences. This fact needs further investigation involving studies of real process of protein-DNA interactions.

**Conclusions**

We believe that the context-free nature of symmetry properties of the genetic text may be regarded as an advantage of our method of image construction because it doesn't require statistical estimation of the text properties. The second advantage is comparatively small number of elementary comparisons which are necessary for pattern recognition in the text in comparison to methods using multiple allignment.

We hope that the "formal" image of replication origins will be used to generate experimentally origin sequences. Testing them in vivo will show the real significance of our results.

**References**

1. A.Leontiev, The symmetry of the single strand DNA molecules. Biophysica. 1992, v.37, No.5, pp .874-878
2. B.K.Vainstein, The symmetry of the crystals. Nauka. 1979