COMBINATORICS OF CIRCULAR CODES

Aleksandr N. Serdiukov
Supervisor: Andrei V, Smolensky, PhD in Physic- and mathematical sciences, Math-Mech. Faculty, St.Petersburg State University
Laboratory of Continuous Mathematical Education, St. Petersburg, Russian Federation, serdyukov-alexander@yandex.ru

1. Introduction

My work is devoted to studying the combinatorics of 216 maximal C3 circular codes – a mathematical objects (sets of words) which come from biology. It could be one of the mechanisms of protection against the reading frame shift.

Circularity is very special property of a set of the words. It means that if form a new word from this set's elements and write it on the circle, it will have at most one decomposition to the elements of this set. In biology this means that after reading 13 nucleotides we will retrieve the reading frame.

In 1996 Arquès and Michel found that the subset of trinucleotides which occurs mostly at the zero frame forms a so-called circular code [1]. Surprisingly, although that sets differ for different organisms, they were circular codes too [3].

Current definition of circular code is quite 'verificative' and 'unconstructive' – it allows us to check whether the given set of words is circular code or not, but does not say anything about their structure and how are they built. One of the latest studies of circular codes that was done in 2014 showed the basic structure of each trinucleotide circular code [2].

2. Research methods

I used following methods:

- Theoretical proofs in language theory
- Theoretical proofs in graph theory
- Programming

I decided to explore the structure of all these codes together in a whole new way using a graph theory. So, I built a special graph which I called "the variant codes graph". Its nodes are circular codes, and edge between two nodes show that intersection of these two sets of words is maximal (18 words).

To work with this graph, I wrote a program on python language using networkx library. This program was mainly used to build the graph, to study its properties and to check whether my assumptions are correct.

Using a graph, I succeeded in finding a new property which was not discovered before: degrees of D4-orbits in the factor-graph coincide with degrees of their members in source graph. It occurred that this property is very special and does not hold for an arbitrary graph or even multigraph.

Then, I have tried to find an invariant which would help to differ these D4-orbits.

3. Results and conclusion

I found that the maximal group of letter mappings, which preserve the circularity property, is D4 over the genetic alphabet and found its interpretation in a graph theory – it is connected with the subgroup of automorphisms of this graph.

Also, I found another special property of a circular codes which became visible in a graph: degrees of D4-orbits in the factor-graph coincide with degrees of their members in source graph.

Moreover, I found an invariant - multiset of the neighbors' degrees, which differs 20 out of 27 D4-orbits. This result is another significant step in describing the structure of all trinucleotide circular codes and finding a way to work with them. It showed that it is very effective to work with them using graph theory.

4. References

[1] Arquès, D.G., Michel, C.J., 1996. A complementary circular code in the protein coding genes. J. Theor. Biol. 182.

[2] Fimmel, E., Giannerini, S., Gonzalez, D.L., Strüngmann, L., 2014. Circular codes, symmetries and transformations. J. Math. Biol. 70.

[3] Michel, C.J., 2015. The maximal C3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. J. Theor. Biol. 380.

[4] Michel, C.J., Pirillo, G., Pirillo, M.A., 2012. A classification of 20-trinucleotide circular codes. J. Inf. Comput. 212.

[5] Michel, C.J., Pirillo, G., 2010. Identification of all trinucleotide circular codes. J. Comput. Biol. Chem. 34(2).

[6] Bussoli, L., Michel, C.J., Pirillo, G., 2012. On conjugation partitions of sets of trinucleotides. J. Appl Math 3.

[7] Michel, C.J., Pirillo, G., Pirillo, M.A., 2008. A relation between trinucleotide commafree codes and trinucleotide circular codes. J. Theor. Comput. Sci. 401.

5. Acknowledgements

The author expresses his gratitude to his supervisor Andrei V. Smolensky and Ilya Alexandrovich Chistyakov for his general supervision of their research work in The Laboratory of Continuous Mathematical Education.