Review of Habilitation Thesis of Dan Faltýnek

The habilitation thesis "Lingvistika genetického textu" is an interesting mixture of linguistics and sequence analysis of genetic texts. It represents the author's diverse professional interests. It is based on individual papers, which are interconnected into a surprisingly functional whole. It is difficult to find a field of linguistics that would not be reflected in the work – there are areas using phonetic analogies to describe DNA, there are studies in grammar (again in analogy to "DNA texts"), I also found reflections on universal grammar, and especially there is the whole bunch of quantitative-linguistic applications in relation to the analysis of "genetic texts" (in forensic linguistics, but also for the needs of discourse analysis – speeches of Czech presidential candidates etc.). There is also a reflection on etymology and a number of boundary linguistic disciplines (especially in relation to the origin of cabbage).

Actually, I was in doubt regarding being a reviewer of the habilitation thesis of Dr. Faltynek. There were two reasons to feel an inappropriate person to write this review: 1. I am not a fluent speaker of Czech and 2. I am not a linguist, but a biologist. The first reason is not so serious, because the time I spent reading the submitted work was, of course, inversely proportional to the level of my knowledge of Czech (with one book, however, I spent even more time, with Hašek's "Osudy dobrého vojáka Švejka"). The second reason seemed stronger to me, but given that there are more reviews (and, according to my information, the authors of them are linguists) and because I helped to establish something what is called 'DNA linguistics', it is probably justifiable, that I am writing this review.

When reading, I was happy about two things in particular: 1. the author is not afraid to cross traditional boundaries between disciplines - it is necessary in contemporary science, I can judge it as a person who was educated in mathematics and computer science, who wrote his master thesis in kind of computational linguistics and doctoral thesis on structural and computational biology. 2. The author feels assured that quantitative linguistic methods are the most appropriate methods in this new field of DNA linguistics from one side, and that he must follow the current development of statistical tools appearing in the field of data analysis (whether in Life Sciences, Data Sciences or Digital Humanities) from another side.

Therefore, I evaluate the work very positively, the connection between computational biology and linguistics is outlined very correctly and with a visionary ability. The author has a great deal of erudition in computational linguistic methods (and linguistics, as far as I can judge) and above all he has the ability to very correctly grasp functional analogies between research areas and thus arrive at a functioning computational-linguistic models that are useful in computational biology. This is in direct opposition to a number of contemporary boundary disciplines that claim a link to biology (e.g. biosemiotics, to a lesser extent also code biology), but at the same time form only anthropomorphic analogies that are not useful to biology.

Of course, I also have some objections that however do not destroy my overall positive impression of this work. First of all, it can be seen that author has acquired quantitative computational tools gradually (albeit quickly) and also in a linguistic context, which is reflected in a certain obsolescence of the presented procedures. The author himself is aware of this (e.g. the original models of "DNA phonetics" are conducted by a simple Ngram analysis, while new research of "DNA texts" of cabbage and procedures for text attribution (Nezval) already use Bag-of -Words methods). This degree of obsolescence is understandable, because computational biology is evolving much faster than computational and quantitative linguistics (colleagues will forgive me). Having in mind to develop a

linguistic approach to the analysis of genomic sequences, one can ask, "What can be called a "word" of a genetic text?" The simplest way is to call by a term "word" every Ngram of the DNA text. If so, a popular approach in DNA linguistics would be based on calculation of frequent or rare words. If an Ngram occurs considerably more (or less) frequently than expected, then it becomes a potential "signal," and the question arises as to the "biological" meaning of this word. Let us take as an example a notion of DNA-protein binding site (Bolshoy, Towards an Encyclopedia of Sequence Biology).

A DNA binding site is a region of chemical bond formation with a specific protein. Such protein usually has a number of effective binding sites, which are rather similar but different. For example, consider the following description of a set of recognition sites: only A is always found in the first position; either C or T may be found in the second position, all four letters appear in the third position, and any base except A appears in the fourth position. It should be noted that the description above does not give any indication of the relative frequencies of the bases A, C, G, and T at any position except the first position. There are much more sophisticated ways to describe binding sites, in particular, and textual patterns, in general. Similarly, to words of a natural language that may appear in different forms along texts, "morphemes" of DNA texts are not identical but correspond to "grammatical rules generating different forms of one and the same morpheme". By analogy with synthetic languages different versions of a particular binding site may be viewed as derivatives of the same word, constructed using different morphemes. Therefore, a "word" of the DNA language may be defined as a consensus sequence for a particular binding site. Another way to define a "word" of the DNA language is "profile". In general, a "word" of the DNA language is a "pattern", certainly not any pattern but defined by some rules.

DNA binding sites can be defined as short DNA sequences (typically 4 to 30 base pairs long, but up to 200 bp for recombination sites) that are specifically bound by one or more DNA-binding proteins (DBP) or protein complexes (DBPC). For example, a DBS associated with transcription factors is called a transcription factor binding site (TFBS). While one DBS of the certain factor is a short string over a four-letter alphabet, to present all or at least representative majority of DBSs for this factor we would need a model. Currently, the state-of-the art method for representing DBSs is via position-specific scoring matrices (PSSMs), also called position weight matrices (PWMs) and, sometimes called profiles. These PSSMs are normalized representations of the position-specific log-likelihoods of a nucleotide's probability to occur at each position of the DBS. There are more advanced models of binding processes.

Another modest objection is that the issue of statistical treatment of the results of quantitative analyzes, which is practically absent in the presented thesis of Dan Faltynek, is also somewhat debatable. Sometimes, this absence disturbs my mind. However, one can assume that the author did not want to overwhelm the text with it and that these treatments in many cases occurred in the original publications and papers.

I also have one more conceptual objection. In most cases, the author analyzes the "DNA text" as if it were another kind of "text" like human texts written in some natural or artificial human language and therefore looks for text patterns there. Actually, Dr. Faltynek looks after patterns related to the so-called "genetic code" that can be found in biology (and which is related to the famous central dogma of molecular biology frequently formulated as DNA → RNA → protein). But this is certainly not fully correct, contemporary biology has perhaps a dozen different types of sequence templates and biases that were described as "second" genetic codes and that go beyond the original simplistic vision of the genetic code published by James Watson in the first edition of The Molecular Biology of the Gene (1965). The author himself knows this because he refers to Edward Trifonov (pp. 25, 49),

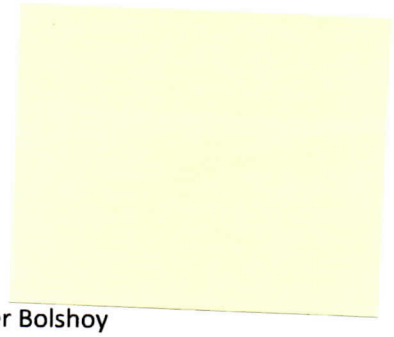
who contributed with his own discoveries to many of these new codes. In 1989 E. N. Trifonov published his highly-cited article „The Multiple Codes of Nucleotide Sequences ". In that pioneer paper Trifonov has introduced his own definition of the term "sequence code": "Code is a sequence pattern instructive for one or another specific molecular (multimolecular) interaction or process". As wrote Dr. Barbieri (Code Biology. A New Science of Life. 2015): "In 80-ies, Edward Trifonov ... started a life-long campaign in favor of the idea that the nucleotide sequences of the genomes carry several messages simultaneously, and not just the message of the classic triplet code. He concluded that there are many overlapping codes in the genome, and gave them the collective name of sequence codes". After thirty years of research in the field of "DNA linguistics" I have proposed my definition: "Genetic code is a set of sequence patterns responsible for the same specific biological function together with a set of rules defining these patterns".

I therefore ask whether the analyzes of DNA as strings of symbols can really give us some new insight into the structure of information in computational biology. What else could be left out? In what respect could the manifestations of Zipf's and Menzerath-Altmann's law be interesting in the case of DNA (and possibly protein) sequences? I proposed the following definition: Sequence Biology is concerned with the statistical or rule-based modeling of genetic sequences from a computational perspective, as well as the study of appropriate computational approaches to subjecting a DNA, RNA or peptide sequence to its intrinsic features, biological function or macromolecular structure encoded in it. Sequence Biology is concerned with description, classification and analysis of Genetic Codes (Bolshoy, Towards an Encyclopedia of Sequence Biology).

I believe that new computational tools will need to be launched in further development of Sequence Biology, which will make it possible to examine "higher levels" of codes than the basic DNA code. Close collaboration of bioinformaticians and linguists is must. Mgr. Dan Faltýnek, Ph.D. is an excellent candidate for such a collaboration. I would strengthen: the fact that the doctor of the humanities can master both formal and computational methods of biology to such an extent is absolutely admirable and I do not know anyone like him who would come from the humanities and acquire exact scientific competencies to such an extent (the reverse is certainly more common).

Therefore, once again I express a positive evaluation of the submitted text and I recommend that Mgr. Dan Faltýnek, Ph.D. awarded the degree of associate professor (or *docent*) in the field of General Linguistics.

June 18, 2020, Ramat Gan



Alexander Bolshoy