Linguistic Frontiers • 1(1) • 65-73 • 2018
DOI: 10.2478/lf-2018-0003

Linguistic **Frontiers**

Towards an Encyclopaedia of Sequence Biology

Review Article

Alexander Bolshoy[1,2,3,*]

[1]*Department of Evolutionary and Environmental Biology, University of Haifa, Haifa, Israel*
[2]*Max Planck Institute for Molecular Genetics, Ihnestraße 63-67, 14195 Berlin, Germany*
[3]*Department of Business Administration, College of Law & Business, Ramat-Gan, Israel*

Received July 30, 2018; Accepted August 23, 2018

Abstract: In this review, I have presented several topics relevant to the present state and to the future state of the scientific field that I propose to call sequence biology (SB). In some pertinent publications, this field was called DNA linguistics. At the heart of SB lies a concept of a sequence code. In this review, I discussed three concepts: a concept of SB, a concept of encyclopaedia of genetic codes, and a concept of a corpus DNA linguistics..

Keywords: *code • DNA • linguistics morphology • pattern • corpus linguistics • bioinformatics*

1. DNA Linguistics

In 80s, Trifonov and Brendel published pioneer works in the field that they proposed to call DNA Linguistics [1, 2]. Actually, the proposed name is a shortened version of *DNA linguistics' morphology*. Indeed, in linguistics, morphology is the study of words, how they are formed, and their relationship to other words in the same language. To justify this name one should explain what he or she would call DNA language, DNA words, and which methods would be used in this new field. Pevzner and other researchers joined Trifonov and his coworkers [3] to establish this new interdisciplinary field that became a part of computational biology. DNA linguistics is concerned with words of DNA texts.

1.1. DNA Texts

Mathematically saying, text is a string over the alphabet of {A, C, G, T}. There are a lot of different meanings of the term "text". Our definition of the term "DNA text" is as follows: "A DNA text is an array of symbols designating DNA basic elements, nucleotides. The symbols arranged in groups provide definite and recognizable instructions". These instructions are destined to rather different "reading devices", such as ribosome, polymerases, and binding proteins [4]. For example, we may say metaphorically that a ribosome gets an instruction to translate a fragment of a DNA text into a string over the alphabet of 20 amino acids. Actually, before this instruction to perform translation, a ribosome gets another instruction that defines the symbol where the translation should start; in other words, ribosome finds the start of translation of a fragment to be translated. Such instructions are sometimes called genetic codes [5].

1.2. DNA Language

Hence, DNA linguistics deals with DNA texts. Would we say that these texts are "written" in DNA Language? To answer this question, we must clarify what do we mean using the term "language" with respect to DNA texts. In early 70s, some linguists defined language as the transmission of information via non-iconic arbitrary signs. Such definitions, probably, would allow us to come with a positive answer: yes; the claim that "DNA texts form a corpus of DNA language" is a legitimate statement. However, even among linguists, any definition of the term "language" that does not mention syntax" cannot be a valid one. In mathematics, computer science, and linguistics, a formal language is a set of strings of symbols from the certain alphabet together with a set of rules that are specific to it.

* *Corresponding author: Alexander Bolshoy, E-mail: bolshoy@evo.haifa.ac.il*

The strings formed from this alphabet are called words, and the words that belong to a particular formal language are sometimes called well-formed words or well-formed formulas. A formal language is often defined by means of a formal grammar such as a regular grammar or context-free grammar, also called its formation rule. In computer science, formal languages are used among others as the basis for defining the grammar of programming languages. Our conclusion is that without any rules governing stream of DNA words in DNA texts, an application of the term language to DNA texts is a metaphor only.

1.3. Words of DNA Language

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 substring of the DNA text [6]. If so, a popular approach in DNA linguistics would be based on the assumption that frequent or rare words may correspond to signals in DNA. If a word 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 [1]. However, even "contrast words" [1, 2] or those "surprisingly" over- or underrepresented substrings cannot serve as appropriate morphemes in DNA morphology.

In the DNA language, one type of a word/morpheme analogue may be a binding site [4]. A DNA-binding site (DBS) is a region of chemical bond formation with a specific protein. Such a 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".

1.4. From Words to Textual Patterns

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.

In computational biology, a problem of enumerating all possible patterns and choosing the most frequent or the fittest among them was treated practically from the beginning of this field [7]. In the frame of this approach, the fitness measures vary from estimates of the statistical significance of discovered signals to the information content of the fragments that altogether belong to the same group. Bioinformaticians tend to use statistical significance of extracted sequence patterns but this is not an exclusively possible attitude. Let us formulate the problem of DBS description once more.

DBSs 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 (DBPs) or DNA-binding protein complexes (DBPCs). 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. The simplest and the most popular way in molecular biology and bioinformatics is to characterize the set of all binding sites of a particular protein by its consensus sequence. The consensus sequence is obtained by aligning all known "versions" of the recognition site and is defined as the idealized sequence that contains the predominant base at each position. All the actual versions should not differ from the consensus by more than a few substitutions. 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 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. A part of it may be incorporating neighbouring base thermodynamic information [8] to align the binding sites searching for the lowest thermodynamic alignment conserving specificity of the binding site. What information should be necessarily included to this underlying model? Here are several questions to be answered by a model designer:

- To consider or not possible interdependencies of contacts between protein residues and DNA nucleotides?
- To consider all site positions as equal or to consider some positions as more important than others? In which way scoring measures should be adjusted to such in many cases very incomplete knowledge?
- To allow or to forbid gaps? An answer to this question depends on our knowledge regarding structure or function of respected DBPCs.
- If gaps are allowed how to calculate penalty function? (Here, the nature of penalty is not evolutionary but rather biophysical and biochemical.)

2. Sequence Codes

Molecular biology researchers in their publications use various names for the patterns with function when they describe such patterns hidden in nucleotide sequences: functional templates, biological signals, etc. As Barbieri [9] wrote: “In 80-ies, Edward Trifonov from the Weizmann Institute of Science, Israel started a life-long campaign in favour 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 [1, 2, 10, 11]”. In this manuscript, I would use terms “sequence code”, “DNA code”, and “genetic code” as synonyms.

2.1. Definition of the Term Sequence Code

The Free Dictionary site gives us quite a few definitions of the term “code”:

- In communications – a **set** of symbols and rules for their manipulation by which the symbols can be made to carry information;
- In communications and information processing – a **system** of rules to convert information in one form into another form or representation;
- In computer science – a **set** of machine symbols that represent data or instructions;
- In information theory and computer science – an **algorithm** which uniquely represents symbols from some source alphabet, by *encoded* strings, which may be in some other target alphabet;
- In molecular biology – the **way** in which a sequence of 20 amino acids is determined by a sequence of four nucleotides [12];
- In law – a **set** of legal rules ¼; and
- In semiotics – a **set** of practices ¼.

One can see that each code definition contains either “a set of symbols” or a term from the group consisting of similar notions such as “a rule”, “a way”, “a practice”, and “an algorithm”. Our definition of the term “sequence code” that I am going to propose satisfies two essential conditions: the definition covers the triplet code, which is often simply named the genetic code, and is suitable for sequence codes mentioned by the researchers using this term [2, 5, 11–15]. Actually, I believe that our definition is a natural extension of the definition introduced earlier by Trifonov [5].

2.2. Genetic Code: How the Term “Code” Was Coined

Obviously, to the popularity of the term “codes” applied to messages carried simultaneously in DNA greatly contributed the common acknowledgement of the term “genetic code”. How this term was invented? As early as in 40s, the Nobel Prize-winning physicist Erwin Schrödinger claimed that the genetic material must contain a “code-script” that determined “the entire pattern of the individual’s future development and of its functioning in the mature state” [16]. This was the first clear suggestion that genes contained some kind of “code”. Both James D. Watson and Francis Crick credited Schrödinger’s book and each independently acknowledged the book as a source of inspiration for their initial researches [17, 18]. (By the way, “although some of the notions in the book have been superseded by modern science, this remains a classic, written with great insight and modesty (Schrödinger downplays his potential as a biologist), and is worth the read if only as a portal in to the minds of those luminary workers” [19].)

John von Neumann, another great scientist, probably, was the first person to formulate hereditary properties using the term “information”. In 1948, von Neumann described a gene as a “tape” that could programme the organism – like the “universal Turing machine” described in 1936 by Alan Turing. A few years later in 1950, geneticist

Hans Kalmus deliberately applied cybernetic thinking to the problem of heredity and suggested that a gene was a "message". Such an atmosphere certainly prepared Watson and Crick to use "cybernetics" terms [20].

Ten years after Schrödinger's brilliant insight, Watson and Crick's second 1953 article on the structure of DNA provided the world with the key to the secret of life, casually employing the new concepts that had been created by cybernetics and propelling biology into the modern age with the words: "it therefore seems likely that the precise sequence of the bases is the code which carries the genetical information".

2.3. Multiple DNA Codes – Trifonov in 1980s

In 70s, Edward N. Trifonov published papers showing that genetic sequences carry several messages simultaneously and not just the message of the classic triplet code. He started to use a notion of "DNA overlapping codes". When in 1989 Trifonov published his review "The Multiple Codes of Nucleotide Sequences" [5], he used the following definition: "¼ the term 'code' is meant as a sequence pattern instructive for one or another specific molecular (multimolecular) interaction or process. It is useful to differentiate between general and particular codes, depending on how common and widespread they are in various organisms. The classical triplet code is the first and best example of a general code. Other, rather modest, examples are the mRNA polyadenylation signal, TATA-boxes of prokaryotic and eukaryotic promoters, gene splicing signals, etc". Ten years later [11], he gave a similar definition: "Genetic code means any sequence pattern or bias responsible for specific biological or biomolecular function". I propose the following 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.

2.4. Sequence Biology (SB) – What Is It?

Taking a wiki definition of computational linguistics as a foundation for further development, I propose the following definition:

> SB is concerned with the statistical or rule-based modelling 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. SB is concerned with description, classification, and analysis of genetic codes.

I have described in the following an example of the code, namely, initiation-of-translation code in bacteria.

3. Encyclopaedia of Sequence Biology (ESB) – What Is It?

Following the introduced definition of SB, the ESB and the encyclopaedia of genetic codes are the two names for the same project. The proposed encyclopaedia would be concerned with genetic codes, so we would like to present our vision of design of the encyclopaedia of genetic codes. I gave earlier our definition of a genetic code. Here I try to describe how entries of the encyclopaedia of DNA codes should be organized from my point of view.

The ESB will be a compendium providing summaries of information regarding sequence codes. ESB will consist of entries that are arranged either alphabetically by code names or thematically following corresponding biological processes and taxonomy. Probably, before or in parallel to compiling, the encyclopaedia dictionaries of genetic codes should appear. The entries in encyclopaedia should be longer and more detailed than those in the dictionaries of sequence codes.

3.1. Vocabularies of Contrast Words

Statistical methods have been successfully used from the early 1980s to extract information from sequences of DNA. In particular, identifying overrepresented motifs might point out unknown biological information [1–6]. From this perspective, these results allowed a biologist to compile Gnomic - a dictionary of contrast words in genetic texts [2]. Interestingly, Gnomic was defined as a dictionary of codes (!), and to the best of my knowledge, neither a vocabulary nor a dictionary of contrast words ever has been compiled.

3.2. Vocabulary of DNA Codes

Vocabularies here are lists of the codes appearing in related dictionaries. For example, vocabulary of prokaryotic codes is a list of those codes that act in prokaryotes, and we can somehow define them to include further in a dictionary of prokaryotic codes. I propose to design and construct vocabularies, dictionaries, and the encyclopaedia of genetic codes in parallel.

3.3. Dictionaries of Sequence Codes

A dictionary is a collection of words, which may include information on definitions, usage, etymologies, pronunciations, translation, etc. Interestingly, in 1986, both a notion of a word of a DNA text and a sequence code were introduced by E. N. Trifonov. To the best of my knowledge, a dictionary of the contrast words has never been published, even a one-species dictionary.

In 1986, Edward N. Trifonov and his student Volker Brendel published "Gnomic: Dictionary of Genetic Codes" – the first and the only dictionary of this kind. The short description of the book says that "this dictionary is the first compilation of all the elements of DNA sequences of known biological meaning; it lists alphabetically and describes some 700 nucleotide sequence 'words'' based on a four-character alphabet (A, C, G, T) used to represent the hereditary information carried by DNA molecules ¼" [2]. In his short review of the book [21], McGeoch mentioned, "Trifonov and Brendel set out their ideas on the analysis of sequence information in an idiosyntactic, even whimsical, manner. The book title is a play of 'genomic' and a Greek root relating to 'knowing'. (By the way, Russian word 'гном' [*gnom*, from *lat*. gēnomos — living under the surface, or from ancient Greek Γnώsh — knowledge] — means fairy-tales' creatures that are known in different languages as «dvergar», «zwerge», «dwarfs», «krasnolud», «trpaslík», etc.] — and it explains extensive usage of parenthetic line drawings of bearded dwarfs along the book.) The book largely consists of lists of short sequences".

Actually, here I invite biologists and linguists to join me and start "**Gnomic – 2.0** Project: New Dictionaries of Genetic Codes". I use here plural instead of single mainly because we are aware today that practically all genetic codes are pretty different for different domains. (In biological taxonomy, a domain, also superkingdom or empire, is the highest taxonomic rank of organisms in the three-domain system: archaea, bacteria, and eukarya.) Even the codes related to the biological process named "translation" are rather different for bacteria and eukarya, all the more so for transcription.

One of the goals is to agree on construction of a genetic-codes defining vocabulary. A defining vocabulary is a list of words that be used to write dictionary definitions. We should design a restricted list of terms that can be used for producing simple definitions of any word in the dictionary. Several words that definitely must be included in the defining vocabulary are prokaryotic, eukaryotic, bacterial, archaeal, viral; transcription, translation, replication, splicing; initiation, termination, prolongation; etc.

3.4. Example of an Entry in ESB

I believe that the most natural way to arrange Encyclopaedia's entries is thematically following molecular biology issues. ESB will consist of the sections named as corresponding biological processes. Sub-subsections will follow taxonomy, and entries will be related to the phases of the process. For example, in the section "Translation", the following sub-subsections are planned to be constructed: triplet code, translation initiation, translation termination, translation pausing, translation framing, and translation recoding. Triplet code is (probably, the only!) universal code and is common to all subsections of the section "Translation". Let me describe one example of an entry construction that is related to the biological process = TRANSLATION; the phase = INITIATION; and the taxonomy selection: DOMAIN = BACTERIA. Translation initiation bacterial code (TIBC) is the name of the entry. Here, we describe a well-studied TIBC for *Escherichia coli*. To the best of my knowledge, a complete TIBC awaits to be defined properly.

There is more than one way to ribosome to start translation and, consequently, TIBC is a set of more or less frequent sequence patterns responsible for initiation of translation in bacteria. In this example of the code entry, I do not pretend neither to completeness nor to absolute accuracy of the descriptions of these patterns. One can find almost all information regarding this code in numerous publications of Kozak ([22] and references therein).

3.4.1. Shine-Dalgarno (SD) containing code (TIBC/SDC)

The SD sequence is a ribosomal binding site in mRNA, generally located around position -7 to -4 of the translational start codon, and it has the sequence complementary to part of the 3¢ end of 16S rRNA. In the case of *E. coli*, the

3¢ end of 16S rRNA is ¼ GAUCACCUCCUUA-3¢ and SD sequence is uaAGGAGGug. The SD sequence does function, albeit with reduced efficiency, when it resides as far as 13 nt from the start codon, but no unequivocal evidence supports larger distances. The SD interaction consists of three to nine contiguous bases that have standard complementarity (not including G·U) to some of 13 bases at the 3¢ end of 16S rRNA (Figure 1).

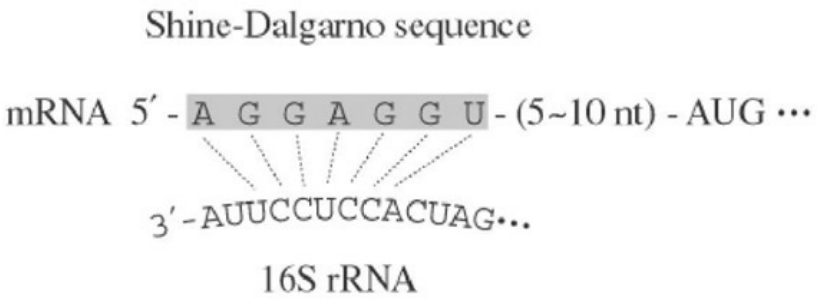

Figure 1. Shine-Dalgarno (SD) sequence of *Escherichia coli*.

Owing to absence of the defining vocabulary (see above), I have used a freestyle definition of the pattern. I assume that in the process of accurate compilation of ESB entries, the defining vocabulary will be processed as well. We can see that there are three elements of this partial code definition of TIBC/SDC: SD, a start codon, and a distance between a start codon and SD. The strengths of these three elements are compensatory: if the start codon is AUG (it is the strongest start codon AUG>CUG>UUG), the SD interaction may be weaker (three to five contiguous bases); a larger distance requires stronger SD interaction, etc. [23].

3.4.1.1. TIBC/SDC with AUG start codon (TIBC/SDC+AUG)

SD elements in natural mRNAs often consist of only three or four bases, despite the availability in 16S rRNA of nine bases. A longer than four bases SD sequence presents if the mRNA contains secondary structure that limits ribosomal access to the AUG codon. The requirement for at least a three base pair SD interaction is almost universal. The three elements of this partial code definition of TIBC/SDC+AUG are as follows: SD length is three to six bases, a start codon is AUG, and a distance between a start codon and SD is from -4 to -13.

3.4.1.2. TIBC/SDC with non-AUG start codon (TIBC/SDC-AUG)

If the start codon has a suboptimal form (GUG or UUG), then the SD sequence is longer than four bases. (By the way, it has been shown that GC content affects the frequency of genes starting with GUG compared with AUG [24]. Furthermore, in some bacteria, such as Bacillus, UUG is more prevalent and leads to higher levels of protein production than GUG [25]. The three elements of this partial code definition of TIBC/SDC-AUG are as follows: SD length is five to nine bases, a start codon is GUG or UUG, and a distance between a start codon and SD is from -4 to -6.

3.4.2. Leaderless-mRNA gene (TIBC/L)

In this case, the mRNA lacks a 5¢-untranslated region (UTR) and hence has no SD sequence in it; thus, the start codon itself serves as the most important signal for the translation initiation. There were ever propositions that signals downstream of the start codon called “downstream boxes” may bind with the 16S rRNA and help translation initiation of leaderless genes, but these suggestions were then refuted by experimental evidences [26]. Unlike AUG, a weaker codon such as UUG or GUG cannot support initiation when positioned exactly at the 5¢ end of the mRNA. TIBC/L as a DNA code requires precise knowledge of the location of the corresponding initiation of the transcription site.

4. Corpus DNA Linguistics

Corpus linguistics may become a significant part of SB. Studies in this discipline would deal with corpora (bodies) of genomic sequences. As it is well-known to linguists, “text corpus is a large and unstructured set of texts (nowadays usually electronically stored and processed)”. Somebody can say that NCBI databases are this kind of corpora, but

I do not think so. Let us check the goals and structure of genomic corpora. Text corpora are used to do statistical analysis and hypothesis testing, and till now, NCBI databases have fulfilled this role, but we expect from genomic corpora more. They should be used for getting, improving, or validating rules defining genetic codes. The text-corpus method is a digestive approach that derives or improves a set of DNA-code rules relevant to taxonomy or another approach governing selection of the texts to the corpus. For example, the corpus of Gram-positive bacteria may be used for derivation of the termination-of-transcription code of Gram-positive bacteria.

Sequence-code grammar (SCG) is a model for describing the syntactic environments of individual sequence codes, derived from studying their occurrences in authentic genomic corpora. Each code is a set of patterns assigned to it. SCG would describe typical contexts in which they are used. Often, these contexts are distinct for different environments. So, different types of contexts have to be analysed, in particular, interrelations with other relevant codes. For example, starts of translation, as a rule, are related to starts of transcription. Respectively, initiation-of-translation code is related to initiation-of-transcription code, especially in prokaryotes. SCG rules for such relations may be rather sophisticated.

In corpus linguistics, a collocation is a sequence of words or terms that co-occur more often than would be expected by chance. In SB, we may develop this approach as well. For example, I expect to apply this attitude and propose a “factor collocation”. The factor collocation is a sequence of binding sites that co-occur more often than would be expected. Let us bring here an example well-known to the linguists. In phraseology, collocation is a subtype of phraseme. An example of a phraseological collocation is the expression *strong tea*. While the same meaning could be conveyed by the roughly equivalent powerful tea, this expression is considered excessive and awkward by English speakers. Conversely, the corresponding expression in technology *powerful computer* is preferred over strong computer. Strikingly similar examples may be brought from analysis of TFBs. Basically, the synonyms in the case of TFs would be TFs of one family. Sometimes, different synonyms are used differently in different combinations with other words (A. Kel, personal communication).

4.1. Construction of Text Corpora

I assumed that GenBank should be at the basis of our corpora compilation. “The National Center for Biotechnology Information (NCBI) provides a large suite of online resources for biological information and data, including the GenBankÒ nucleic acid sequence database. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. Database Genome is part of GenBank and contains sequence and map data from the whole genomes of over 1000 organisms. The genomes represent both completely sequenced organisms and those for which sequencing is in progress. All three main domains of life (bacteria, archaea, and eukaryota) are represented, as well as many viruses, phages, viroids, plasmids, and organelles.” Somebody could imagine that the database Genome [27] is a kind of text corpora, but it is not.

- Genomes are annotated in a way dissimilar to corpora annotations.
- A choice of genomes to be sequenced is very biased, while texts in a corpus must be carefully selected.
- Some species are represented by many strains, and others exist in only one reference genome.

Principles of genomic corpus should be discussed and defined, and here, it is my primary contribution to these issues. Let us exploit a linguistic view on corpora: “The advantage of publishing an annotated corpus is that other users can then perform experiments on the corpus. Linguists with other interests and differing perspectives than the originators’ can exploit this work. By sharing data, corpus linguists are able to treat the corpus as a locus of linguistic debate, rather than as an exhaustive fount of knowledge”. So, we would like to have an access to genomic corpora with texts from GenBank but annotated following the principles formulated earlier and accompanied with toolset helping researchers to analyse corpora by different means.

4.2. From Vocabularies of DNA Codes to a 3A Sequence-Biology Retrieval Database

In the previous chapters, I introduced my vision of genetic-code notion. Here, I would like to present my vision of a pass from vocabularies, dictionaries, and other compilations of genetic codes to a newly organized database of annotated genetic texts. These texts, organized in a way dictated by corpus-driven approach, would use entries of the encyclopaedia of DNA codes (ESB entries are described above in Section 3.4) to produce annotations of genomic texts, to pass from a general scheme of a code to its particular realization in the given text. Moreover, such a database assumes feedback as well: as a result of the application of a genetic-code scheme, following analysis of

the application will allow users to suggest improvement and refinement of the patterns and rules defining the applied code. Corpus linguistics would inspire us to develop new approaches in genomic-text analyses.

Corpus linguistics has generated a number of research methods, for example, the 3A perspective: annotation, abstraction, and analysis. Let us try to adapt these linguistic methods for biological purposes. I will start with reformulation of the definitions of the 3A.

- Annotation consists of the application of a scheme to texts. For example, if the scheme is PWM, an appropriate tool (for example, PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix [28]) may be used to scan texts. As a result, annotations may include genome positions, a forecasted affinity, similarity to a general pattern, and numerous other representations.
- Abstraction consists of the translation (mapping) of terms in the scheme to terms in a theoretically motivated model. My interpretation of the above linguistic definition of abstraction in sequence biology suggests translation of terms in the genetic-code scheme to biophysical or biochemical or other biological terms.
- Analysis consists of statistically probing, manipulating, and generalizing from the dataset. Analysis might include statistical evaluations, machine learning, or knowledge discovery methods.

So many corpus-driven research methods may be applied to genomic databases that I shall better bring examples.

4.3. Examples of Analyses That May Be Applied to Such a 3A Database

Here are several examples. They are the first ones that come to my mind, but a reader can easily extend this list of examples.

- My impression is that search for collocations may outcome in very interesting results. Co-locations of TFBSs, enhancers and silencers, stop codons, and terminators – these and other sequence elements – mapped using genetic-code schemes may assist in discovery of collocations.
- Distribution of distances between corresponding starts of transcription and translation in different prokaryotic taxa. Today, it is a well-known fact that sequences of the UTRs of mRNAs play important roles, but the connection between UTR length and protein expression is not clear.
- Variance of SD sites mapped using a corresponding scheme in different prokaryotic taxa.
- Search for potential non-AUG start codons. It was done in the past but still there is a potential for further improvements. See, for example [29]: "¼ motif analysis revealed that 1) with the right sequence context, certain non-AUG start codons can generate expression comparable to that of AUG start codons, 2) sequence context affects each non-AUG start codon differently, and 3) initiation at non-AUG start codons is highly sensitive to changes in the flanking sequences. Complete motif analysis has the potential to be a key tool for experimental and diagnostic genomics": "It is thus becoming increasingly clear that start codon selection is regulated by ¼ sequence/structural elements within messenger RNAs and that non-AUG translation has a profound impact on cellular states" [23, 30].

This is only a very preliminary proposal of application of corpus-driven research methods to genomic databases.

References

[1] Brendel, V., Beckman, J.S., Trifonov, E.N., 1986. Linguistics of nucleotide sequences: Morphology and comparison of vocabularies. *Journal of Biomolecular Structure and Dynamics*, 4, 11–21.

[2] Trifonov, E.N., Brendel, V., 1987. *Gnomic: Dictionary of genetic codes*. Rehovot: Balaban Publishers, 1986; Wiley-VCH Verlag GmbH.

[3] Pevzner, P.A., Borodovsky, M.Y., Mironov, A.A., 1989. Linguistics of nucleotide sequences. I: The significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *Journal of Biomolecular Structure and Dynamics*, 6, 1013–1026.

[4] Bolshoy, A., Volkovich, Z., Kirzhner, V., et al., 2010. *Genome clustering: From linguistics models to classification of genetic texts*, Studies in Computational Intelligence, Berlin: Springer-Verlag.

[5] Trifonov, E.N., 1989. The multiple codes of nucleotide sequences. *Bulletin of Mathematical Biology*, 51, 417–432.

[6] Pevzner, P., 2000. *Computational molecular biology: An algorithmic approach*. Cambridge, MA: MIT Press.

[7] Brazma, A., Jonassen, I., Eidhammer, I., et al., 1998. Approaches to the automatic discovery of

patterns in biosequences. *Journal of Computational Biology*, 5, 279–305.
[8] Salama, R.A., Stekel, D.J., 2013. A non-independent energy-based multiple sequence alignment improves prediction of transcription factor binding sites. *Bioinformatics*, 29, 2699–2704.
[9] Barbieri, M., 2015. *Code biology. A new science of life*. Dordrecht: Springer.
[10] Trifonov, E.N., 1987. Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16s rRNA nucleotide sequence. *Journal of Molecular Biology*, 194, 643–652.
[11] Trifonov, E.N., 1999. Elucidating sequence codes: Three codes for evolution. *Annals of the New York Academy of Sciences*, 870, 330–338.
[12] Crick, F.H., 1963. On the genetic code. *Science*, 139, 461–464.
[13] Turner, B.M., 2000. Histone acetylation and an epigenetic code. *Bioessays*, 22, 836–845.
[14] Turner, B.M., 2007. Defining an epigenetic code. *Nature Cell Biology*, 9, 2–6.
[15] Barbieri, M., 2012. Code biology – a new science of life. *Biosemiotics*, 5, 411–437.
[16] Schrödinger, E., 1944. *What is life? The physical aspect of the living cell*. Cambridge, UK: Cambridge University Press.
[17] Crick, F., 1989. *What mad pursuit*. London: Penguin.
[18] Watson, J. 1981. *The double helix*. London: Weidenfeld and Nicholson.
[19] Derry, J.F., 2004. Review of what is life? by Erwin Schrödinger. *Human Nature Review*, 4, 124–125.
[20] Cobb, M. 2015. *Life's greatest secret: The race to crack the genetic code*. London: Profile Books.
[21] McGeoch, D.J., 1987. Books in brief, *TiBS*, 12, 165.
[22] Kozak, M., 1999. Initiation of translation in prokaryotes. *Gene*, 234, 187–208.
[23] Belinky, F., Rogozin, I.B., Koonin, E.V., 2017. Selection on start codons in prokaryotes and potential compensatory nucleotide substitutions. *Scientific Reports*, 7(1), 12422. doi: 10.1038/s41598-017-12619-6.
[24] Villegas, A., Kropinski, A.M., 2008. An analysis of initiation codon utilization in the domain bacteria–concerns about the quality of bacterial genome annotation. *Microbiology*, 154, 2559–2661.
[25] Rocha, E. P., Danchin, A., Viari, A., 1999. Translation in Bacillus subtilis: Roles and trends of initiation and termination, insights from a genome analysis. *Nucleic Acids Research*, 27, 3567–3576.
[26] Moll, I., Grill, S., Gualerzi, C.O., et al., 2002. Leaderless mRNAs in bacteria: Surprises in ribosomal recruitment and translational control. *Molecular Microbiology*, 43 (1), 239–246.
[27] Genome database, available at: <https://www.ncbi.nlm.nih.gov/genome>.
[28] Ambrosini, G., Groux, R., Bucher, P., 2018. PWMScan: A fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics*, 34, 2483–2484.
[29] Diaz de Arce, A.J., Noderer, W.L., Wang, C.L., 2018. Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. *Nucleic Acids Research*, 46(2):985–994.
[30] Kearse, M.G., Wilusz, J.E., 2017. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Development*, 31(17), 1717–1731. doi: 10.1101/gad.305250.117.