The Linguistics of DNA: Words, Sentences, Grammar, Phonetics, and Semantics

SUNGCHUL JI

Department of Pharmacology and Toxicology, Rutgers University, Piscataway, New Jersey 08855, USA

There are theoretical reasons to believe that biologic systems and processes cannot be fully accounted for in terms of the principles and laws of physics and chemistry alone, but they require in addition the principles of semiotics—the science of symbols and signs, including linguistics.\textsuperscript{1-3} For convenience, we may refer to the belief, common among contemporary molecular biologists, that the laws of physics and chemistry are necessary and sufficient to account for life as the PC (physics and chemistry) paradigm, while the alternative view that principles of semiotics are additionally \textit{absolutely} required for a complete understanding of living systems and processes as the PCS (physics, chemistry, and semiotics) paradigm.

It was von Neumann who first recognized the necessity for symbolic self-representation of organisms as a prerequisite for efficient self-replication.\textsuperscript{4} In view of the fundamental importance of this insight for biology, we may refer to this notion as the \textit{von Neumann doctrine}. This doctrine was further elaborated and developed by Patté into what may be called the theory of \textit{matter-symbol complementarity}.\textsuperscript{5,6} The linguistic theory of DNA presented here can be viewed as a natural extension to the structure and function of DNA of the \textit{von Neumann doctrine} and Patté's theory of \textit{matter-symbol complementarity}. (See Note Added in Proof.)

Since the discovery of the DNA double helix in 1953, many biologists have employed language as a useful metaphor to describe certain aspects of molecular biologic phenomena.\textsuperscript{7-9} But recently it was postulated that language is more than just a metaphor and that linguistics provides a fundamental principle to account for the structure and function of the cell. This conclusion is supported by the facts (1) that cells use a language, called cell language or cellese, defined as \textquotedblleft a self-organizing system of molecules, some of which encode, act as signs for, or trigger, gene-directed cell processes,	extquotedblright and (2) that cell language has molecular counterparts to 10 of the 13 design features of human language (\textit{humanese}) characterized by Hockett and Lyon, thus suggesting an isomorphism between cellese and humanese.\textsuperscript{10,11} Because cellese must be transmitted from one generation to the next, it must be encoded in DNA. Therefore, the main objective of this communication is to characterize the structure and function of DNA based on linguistic principles.

\textbf{ISOMORPHISM BETWEEN CELL AND HUMAN LANGUAGES}

Both human and cell languages can be treated as a 6-tuple \{L, W, S, G, P, M\}, where L is the alphabet (i.e., a set of basic symbols called \textit{protosemata})\textsuperscript{15}, W is the vocabulary or lexicon (i.e., a set of words), S is an arbitrary set of sentences, G is a set of the rules governing the formation of sentences from words (the \textit{first articulation}) as well as the forma-
tion of words from letters (the second articulation), $P$ is a set of physical mechanisms realizing and implementing a language, and finally $M$ is a set of objects (both symbolic and material) or processes referred to by words and sentences. Table 1 summarizes a comparison between sound-based and visual signal-based human language and molecule-based cell language with respect to these categories of linguistic features. The table is self-explanatory, and newly appearing terms are explained in the accompanying footnotes. The isomorphism between cell and human languages evident in Table 1 suggests the existence of three distinct categories of genetic information in DNA here called the lexical, syntactic, and semantic. To visualize the relation among these three categories of genetic infor-

<table>
<thead>
<tr>
<th>Human Language</th>
<th>Cell Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Alphabet (L)</td>
<td>Letters</td>
</tr>
<tr>
<td>2. Lexicon (W)</td>
<td>Words</td>
</tr>
<tr>
<td>3. Sentences (S)</td>
<td>Strings of words</td>
</tr>
<tr>
<td>4. Grammar (G)</td>
<td>Rules of sentence formation</td>
</tr>
<tr>
<td>5. Phonetics (P)</td>
<td>Physiologic structures and processes underlying phonation, audition, and interpretation</td>
</tr>
<tr>
<td>6. Semantics (M)</td>
<td>Meaning of words and sentences</td>
</tr>
<tr>
<td>7. First Articulation</td>
<td>Formation of sentences from words</td>
</tr>
<tr>
<td>8. Second Articulation</td>
<td>Formation of words from letters</td>
</tr>
</tbody>
</table>

$^6$Genes that control the spatiotemporal evolution of the expression of structural genes by regulating the time- and space-dependent folding patterns of chromosomes.$^9$

$^8$Conformational strains of biopolymers that carry free energy (to do work) and information (to control work).$^9$

$^9$Dissipative structures of Prigogine (or attractors) localized within the cell.$^9$

$^9$Molecular interactions that do not implicate any breaking or forming of covalent bonds.

$^9$Molecular interactions that involve changes in covalent bonds, namely, alterations in valence electronic configurations.
mation, it is convenient to use a loaded carousel as a metaphor for DNA with the alignment as shown in Table 2.

Just as a grammar constrains mentally the word order in sentences, so a carousel constrains physically the positioning of slides into a linear array, any linear array. The genetic analog of this constraint is referred to as the syntactic genetic code identified with the physicochemical constraints of nucleic acids that control the folding patterns of chromatin in response to microenvironmental conditions such as the presence of transcription factors, pH, ions, and mechanical stresses of nuclear scaffolding. Please note that there are a large number (i.e., n!) of arranging n slides into n slots in a carousel. But only one, or at most a few, of these linear arrays will actually be utilized by a given speaker. The information needed (log₂ n! bits) to select these few arrangements out of the large possible arrangements derives from the brain of the speaker. But in the case of DNA, the information determining the temporal order in which a set of genes is expressed must be encoded in DNA itself (in the form of semantic genetic code)—in regions that were previously called spatiotemporal genes and postulated to be located in noncoding DNA.⁹¹⁰ In the human genome, structural genes account for approximately 3% of the total DNA mass, whereas the remaining 97% of DNA is noncoding and was once thought to be without any biologic function. But impressive amounts of empirical data were recently accumulated in the literature, indicating that noncoding regions, particularly “repetitive sequences,” play an important role in genetic control processes.¹³ Consistent with these developments, it is postulated here that these noncoding regions regulate the spatiotemporal evolution of the expression of structural genes and thus contain genetic information analogous to the semantic information of sentences. The genetic information that determines the spatiotemporal organization of gene expression is referred to as “semantic genetic code.” It is thought that semantic genetic information is a subset of syntactic genetic information, just as semantically meaningful sentences constitute but a small subset of grammatically correct sentences in human language. The syntactic genetic information is distributed over the whole DNA molecule in that every aspect of the physics and chemistry of DNA affects the dynamics of DNA. Therefore, the sum of all the genetic information encoded in DNA is 200% (Table 2). This makes sense only if we can assume that DNA structures encode more than one kind of information within identical sequences and that different kinds of

<table>
<thead>
<tr>
<th>Model</th>
<th>Molecular Representation</th>
<th>Genetic Code (DNA mass involved)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slides</td>
<td>Structural genes in coding DNA</td>
<td>Lexical genetic code (3% in humans)</td>
</tr>
<tr>
<td>Carousel</td>
<td>Sugar-phosphate backbone, Watson-Crick base pairing, chemistry and physics of DNA</td>
<td>Syntactic genetic code (100%)</td>
</tr>
<tr>
<td>Order of slides</td>
<td>Space- and time-dependent gene expression, made possible by space- and time-dependent foldings of chromatins exposing right genes at right times, all regulated by spatiotemporal genes located in noncoding DNA</td>
<td>Semantic genetic code (97%)</td>
</tr>
</tbody>
</table>

TABLE 2. The "Loaded Carousel" Model of DNA Structure and Function
genetic information can overlap in DNA, in agreement with the multiple genetic code hypothesis of Trifonov. The present result is also consilient with the view that DNA possesses dual or complementary aspects—dynamic and semiotic, or material and symbolic. The syntactic genetic code represents the dynamic or material aspect of DNA obeying the laws of physics and chemistry, while lexical and semantic genetic codes constitute the symbolic (or sign) aspect that obeys the rules forged by biologic evolution. This interpretation fits nicely with the notion of the matter-symbol complementarity (or more generally matter-sign complementarity; see Note Added in Proof) as the most fundamental distinguishing feature of biology vis-à-vis physics and chemistry.

Indirect evidence for the existence of spatiotemporal genes (carrying semantic genetic code) was recently provided by Amano et al. Their data from Figure 8 can be replotted in a graph of the percentage of noncoding bases per genome versus the relative amount of structural genes in the form of transcription factors per genome to obtain two lines, one with a zero slope passing through five species of unicellular organisms (Mycoplasma genitalium, Haemophilus influenzae, Methanococcus jannaschii, Synechocystis sp., and Escherichia coli) and the other with a slope of about 20 passing through three species (Saccharomyces cerevisiae, Caenorhabditis elegans, and Homo sapiens), two of which are multicellular organisms. Interestingly, these lines intersect in the neighborhood of E. coli and S. cerevisiae. Two conclusions may be drawn from this plot: (1) The amount of noncoding DNA increases abruptly with the multicellularity of organisms (most likely due to the fact that noncoding regions act as “spatiotemporal genes” regulating the development of multicellular organisms), and (2) Of the two mechanisms for regulating gene expression—trans mechanisms mediated by transcription factors and cis mechanisms mediated by noncoding regions, the latter contributing to a greater extent (due to the slope being greater than 1) than the former as the complexity of multicellular organisms increases.

The role of noncoding DNA strongly suggested by these data is difficult to be accommodated by the traditional view that the final referents or meaning of genes are polypeptides. However, the data are consistent with the so-called DNA-polypeptide-IDS hypothesis which claims that the final products of genes (i.e., structural genes under the control of spatiotemporal genes) are not polypeptides but dynamic processes collectively called intracellular dissipative structures (IDSs) whose generation is catalyzed by enzymes encoded in structural gene. IDSs include ionic gradients in the cytosol or across biomembranes, and mechanical stress gradients in biopolymers including cytoskeletons and DNA, all of which together act as the proximal or immediate causes for cell functions. (see also Fig. 1).

According to some linguists, the phenomenon of double articulation or duality (see seventh and eighth rows in Table 1) is the most fundamental aspect of all human languages. The cell language theory is based on the basic assumption that the cell-linguistic counterpart of double articulation is the duality of covalent and non-covalent interactions in the cell. Just as the first and second articulations are both essential in human language, so it is postulated that both covalent and conformational interactions are fundamental in cell language (enabling intercellular communication and signal transduction). This postulate appears to provide the first explicit rationale for the fundamental role of conformational interactions in molecular biology, as observed in ligand-protein interactions, protein foldings, and chromatin reorganizations during the cell cycle.
FIGURE 1. The Bhopalator, a molecular model of the living cell. The cell is a molecular machine in that its moving parts are made out of molecules, some of which act as molecular motors driven by conformational strains (called conformants) generated from chemical reactions or ligand-binding reactions. The final form of expression of genes is not polypeptides, as usually thought, but dissipative structures of Prigogine (or attractors) (see the rectangle), namely gradients of chemical concentrations and mechanical stresses in the cell. These dissipative structures act as the direct causes for all cell functions. Solid arrows indicate the direction of information flow—from DNA to mRNA, to proteins, to dissipative structures of Prigogine (also called intracellular dissipative structures, or IDSs), and back to DNA. Dotted arrows indicate feedback interactions. The cell receives inputs from its surrounding (Step 19) and process it according to the genetic information stored in DNA (Steps 5–11 and 1–4) and outputs the result (Step 20). Because IDSs can influence the rate of mutations and recombinations of DNA (Step 10), DNA can guide its own evolution. That is, DNA is a self-evolving molecule driven by conformons and IDSs. For more information, see reference 9, particularly p. 178.

THE BHOPALATOR: A MOLECULAR MACHINE THAT ACCEPTS CELL LANGUAGE

Since the founding of the cell theory in the mid-nineteenth century, there had been no rigorous and comprehensive theoretic model of the living cell available in the literature until 1983, when the Bhopalator model of the cell was proposed in a meeting held in Bho-
pal, India. The name of the model reflects the convention that mechanisms of \textit{self-organizing chemical reaction diffusion systems} are named as “X-atoms,” where X is the name of a city connected in some way with the model. Two concepts are novel in this model—(1) \textit{conformomers}, sequence-specific conformational strains of biopolymers that provide free energy and control information for driving all molecular motors in the cell, and (2) \textit{intracellular dissipative structures (IDSs)}, gradients of chemical concentrations and mechanical stresses in the cell that mediate information transfer from the nucleus to the cytosol and from the cytosol to the extracellular space. With conformomers and IDSs (synonymous with “attractors”), the cell cannot only “read” genetic messages encoded in DNA but also “implement” and “reify” these messages into molecular processes and actions constituting cell functions. In other words, the Bhopalator can be viewed as a molecular machine that accepts cell language encoded in DNA, just as the Turing machine acts as an abstract machine that accepts and defines a formal language encoded on a tape. It is to be noted that these molecular entities that drive the cell, namely conformomers and IDS’s, can be viewed as the microscopic embodiments of the matter-symbol complementarity discussed by Pattee.

\section*{Predictions}

1. The cell language theory predicts that DNA of higher eukaryotes contains two kinds of genes: structural genes located in coding regions (accounting for \~3\% of the human genomic mass) and spatiotemporal genes located in noncoding regions (\~7\% of the human genomic mass).

2. Spatiotemporal genes encode the information controlling the timing of gene expression.

3. The timing information encoded in spatiotemporal genes is retrieved through space- and time-dependent chromatin folding and unfolding processes driven by ATP-dependent topoisomerases and free energy-releasing binding interactions between transcription factors and DNA.

\section*{Conclusion}

The cell language theory and the Bhopalator model of the living cell provide the first comprehensive and rigorous theoretic framework for molecular and cell biology. As such, they may find important applications in functional genomics in the coming decades.

\textit{[Note added in proof: It was recently suggested elsewhere (S. Ji, “The cell as the smallest DNA-based molecular computer” BioSystems, in press) that the ideas of J. von Neumann (i.e., the necessity of self-representation for self-reproduction) and H. Pattee (namely, the matter-symbol complementarity as an essential feature of all self-reproducing systems) be combined into what may be called the “\textit{von Neumann-Pattee principle of matter-cancel complementarity}.” The term ‘symbol’ is replaced with the more general term, ‘sign,’ since according to C.S. Peirce (1839–1914), signs include symbols along with icons and indexes (J. J. Liske, “A General Introduction to the Semeiotic of Charles Sanders Peirce, Indiana University Press, Bloomington, 1996). The essential content of the von}
Neumann-Pattee principle of matter-sign complementarity is that all self-reproducing systems embody two complementary aspects—the physical law-governed material/energetic aspect and the evolutionary rule-governed sign aspects. The dual role of DNA revealed in Table 2, namely the fact that \([\text{syntactic} (100\%)] + [\text{lexical} (3\%) + \text{semantic} (97\%)] = 200\%\), finds a rigorous theoretical rationale in the von Neumann-Pattee principle of the matter (syntactic)-sign (lexical & semantic) complementarity.

REFERENCES