Many years ago I heard a lecture by the mathematical linguist Solomon Marcus with the memorable title ‘Linguistics as a pilot science’. Linguistics is a pilot science in the sense that concepts and theoretical frameworks developed in linguistics tend to be borrowed as models in other sciences. Most of us are familiar with the influence linguistic ideas have had in such fields as anthropology, sociology, cognitive science, philosophy, literary analysis, musicology, semiotics, and social psychology. But to fully appreciate the pilot science status of linguistics these days, one has to turn to molecular biology and biochemistry, and theories of the structure of DNA.

The metaphor of DNA as a language has become very prominent in DNA research. DNA sequences are viewed as texts written in a complex and unfamiliar language. For help in decoding these texts, some biologists are said to be using methods of linguistic analysis. E.N. Trifonov and V. Brendel have produced the first dictionary of this genomic language, which they call ‘Gnomic’. More recently, the biologist Robert Pollack has written a book which explores the metaphor of DNA as a language and urges that biology adopt a new paradigm which owes less to physics and more to the study of signs. When we look more closely at how this metaphor is being applied, however, we find that current thinking about the structure of DNA is roughly where thinking about human language was decades, even centuries, ago.

The DNA molecule consists of sequences of four bases, adenine, cytosine, guanine, and thymine. These bases, abbreviated A, C, G, and T, respectively, are arrayed along two linked strands, where A on one strand is always linked to T on the other, and C is always linked to G. These four letters form ‘words’ (known as domains); words make up ‘sentences’ (genes). One problem in identifying words is that there are no spaces between them, rather like spoken Dutch.

Trifonov and Brendel seek to identify ‘contrast words’, defined as $n$-strings (a) whose $n-1$ derivative strings appear significantly more frequently as part of the $n$-string than elsewhere, and (b) which are not themselves part of a statistically distinguishable $n+1$ string (1986: 13). For example, the English string eyon occurs almost exclusively as part of beyond; any further extension will almost certainly belong to a neighbouring word (e.g. beyondt as in beyond the...).

When this method is construed as a theory of grammar, it becomes a finite-state machine, the ‘elementary linguistic theory’ which Chomsky showed to be descriptively inadequate in Syntactic Structures. However useful such a model may be in identifying consistent substrings in an unfamiliar text, it hardly represents the state of the art of linguistic analysis. Similarly, Trifonov & Brendel’s Gnomic dictionary resembles modern dictionaries less than it does the random vocabulary lists of missionaries and explorers. The following is a sample entry (p. 42):

\begin{enumerate}
  \item \texttt{AATATAATAAAA}: sequence frequently found 5 to 15 bases upstream from coding regions of Dictyostelium actin genes.
\end{enumerate}

What is striking to a linguist is that the ‘sentences’ of DNA, though impressively long, lack—so we are to believe—all the features that make human languages several orders of magnitude more complex than finite-state machines: hierarchy, unbounded dependencies, empty elements,
underlying forms, the distinction between A and A-bar positions – in short, everything that theoretical linguists work on.

Could it be that DNA simply lacks all these things? This is hardly credible. Recall that human DNA encodes instructions for, among other things, language. All of UG is in there (pace Piaget): X-bar theory,\(^8\) the prosodic hierarchy,\(^9\) the S-R-E structure of tense,\(^10\) the general principle that bars successive a-phrases (ruling out \(*Juan hizo afeitar a Pedro a María*),\(^11\) the Free-V constraint (whereby word-final vowels must not be parsed in the nominative),\(^12\) and, of course, the Null Subject parameter.\(^13\) Consider also that DNA has been evolving for billions of years, whereas UG is not older than the human species (pace Koko).

Moreover, there is empirical evidence that there is more to DNA than linear strings of genes. It has been observed that interspersed among the identified genes are many sequences that appear to have no genetic meaning, called ‘junk DNA’. Surprisingly, the great majority of DNA base sequences appear to be junk, 95–97\% in some estimates.\(^14\) Junk DNA appears to be highly structured, with the same one or two letters repeated many times, and is unexpectedly stable, suggesting it does something worthwhile. It displays nonlocal dependencies.\(^15\) According to R. J. Britten, some junk sequences may ‘serve as subtle modulators of the genes they are near’; moreover, these sequences ‘are just where one would expect them to be if they were to act as docking sites for the proteins that flick on genes or turn them up to high volume’.\(^16\)

Subtle modulators? Docking sites? In the words of the baseball manager Casey Stengel, ‘Can’t anybody here play this game?’ These may be novel concepts in DNA research, but theoretical linguistics is concerned with little else. Consider, for example, the Hebrew sentence in (2):

\begin{center}
(2) \textit{rina amra le-ran \textit{še-hicliax} ba-bxina}  \\
Rina said to-Ran that-succeeded in.the-test.
\end{center}

‘Rina told Ran that he succeeded in the test.’

For the portion of this sentence that consists of the single (phonological) word \textit{še-hicliax} , Ritter (1995: 440) posits the structure in (3).\(^17\)

![Diagram of the sentence structure](attachment:diagram.png)
Preceding (or dominating) the single node to which we can attach a clear meaning (the verb *hicliax* ‘succeeded’) are 17 junk nodes whose relation to the external world is somewhat more indirect. These nodes, called ‘functional categories’, are nevertheless not meaningless, but serve as subtle modulators of the word they are near, and act as docking sites for various constituents. Notice also that these nodes amount to 94.4% of the structure in (3), about the same as the percentage of junk DNA. The parallels are too striking to be coincidental.

Pollack (1994: 12) proposes that DNA researchers should draw inspiration from semiotics, the science of signs. Semiotics is unlikely to shed much light on the structure of DNA, though, because the one well-understood sign system (traffic lights) is much simpler than both human language and DNA, and the more ambitious sign systems studied by semioticians (Renaissance painting, the language of film noir) have so far proved to be even more elusive than DNA (pace Eco). No, for the future of DNA research, look to linguistics, the pilot science.

Notes

2. M. Schrage, Learning to speak the language of life. (DNA as language is one of modern science’s most powerful metaphors). *Los Angeles Times*, April 22, 1993, D1 col. 2.
12. A. Prince & P. Smolensky, Optimality Theory: constraint interaction in generative grammar (Technical report #2 of the Rutgers Center for Cognitive Science). Rutgers University, 1993, p. 101. This constraint is ranked high enough in Lardil for its effects to be observed there.
14. N. Angier, Keys emerge to mystery of ‘junk’ DNA; most DNA does not code for genes; biologists have some new answers to this puzzle (*Science Times*). *The New York Times*, June 28, 1994, C1 col. 5.
15. P. Yam, Talking trash; linguistic patterns show up in junk DNA. *Scientific American* 272, March 1995, 12. The ‘linguistic patterns’ refer to Zipf’s law.