文献紹介

David B. Searls : Investigating the Linguistics of DNA with Definite Clause Grammars, *Logic Programming Proc. of North American Conference*, pp. 189-208 (1989).

DCG (Definite Clause Grammar) とは，自然言語処理にはなじみのパーサである．これはいわばPrologのマクロであり，静的には文法記述を与えるのと同時に，動的には文構造を決定するためのパーサとして用いられる．すなわち，文とは主語と述語からなるものであり，おのおのは名詞句と動詞句であるなどという文脈自由文法 (CFG) 風の記述がそのままプログラムとしても動くものと思ってよい．DCG自体は自然言語処理という観点からはすでに一定の評価を得たものと言ってよく，現時点では高速化以外の興味は持たれていない．つまり，自然言語の構造記述能力という点では，解決できない問題も多いままでいる．

ところが，このパーサを自然言語以外のところでシンタックス解析に用いようという試みが現れた．生物の遺伝子構造，すなわちDNAの解析に用いようというものである．

DNAというのはわずか4種類のヌクレオチドと呼ばれる塩基からできた暗号文である．すなわち，このヌクレオチドの配列が生物の遺伝情報のすべてを担っている．この4種類は2組の相補的な対をなす．この配列は長い1本のストリングの構造をなすわけであるが，それぞれのヌクレオチドに対になるヌクレオチドを配列させたもう1本別のストリングが存在し，この2本がからまりあって有名ならせん構造をなす．

本論文では，まずこの配列を形式言語とみなすと，この言語がチョムスキー階層のどのクラスに属するかという議論から始まる．まず，DNAは文脈自由 (CFL) ではない．DNAは任意回，複数のパターンの繰返しを持つからである．ここで憶測されるのがDNAは文脈自由と文脈依存 (CSL) の間にあるインデックス文法 (indexed grammar) ではないか，ということである．インデックス文法とは各非終端記号にインデックスをつけておいてインデックスにスタック操作を行うことにより，$a^n b^n c^n$のような文を生成できる文法のクラスである．DCGにとってインデックス文法をインプリメントするのは容易である．各非終端記号がインデックス情報をPrologのリストの形で持って回れるようにしてやればよい．

さて，論文の次の目的は，DNA上に現れる特有の言語パターンに着目し，そのパターンを容易に記述できるように文法表現を拡張していくことである．

DNA上の情報というのは，mRNAにコピーされて出ていき，リボゾームに入力情報として読み込まれる．一方，各アミノ酸はタグとしてtRNAを持っている．mRNA上のヌクレオチドは3個1組で一つのアミノ酸を指示していて，この組と相補的なtRNAを持つアミノ酸がリボゾームによって引き合わさせ，アミノ酸列（ポリペプチド）を生成する．この過程において情報となるのはヌクレオチド配列といった一次構造だけではない．すなわち，RNAの持つ配列がストリングの他の部分と結合して決まる二次構造も重要な情報源になる．

本論文で導入されるさまざまな文法表現というのは，この二次構造をも容易に記述できることを意図したものである．例えば，RNAは1本のストリングであるが，そのうちのある部分は他の離れた相補的な部分と結びついて複雑なループを作ったりする．まずあるストリングXに対して，それと相補的なヌクレオチド配列からなるストリングを$\widetilde{X}$と書く．このようなストリングを表す変数をストリング変数と呼ぶ．すると，ギャップを含んだ相補的なパターンの繰返しというのは，

inverted_repeat --> X, _, $\widetilde{X}$.

などと書くことができる．また細菌の遺伝子の中に発見される抑制因子の構造は以下のように書くことができる．

attenuator --> X, _, $\widetilde{X}$, _, X, _, $\widetilde{X}$.

以下，同様にしてDNA, RNA上のさまざまな現象，例えば，重ね合わせ，突然変異，遺伝子組み替えなどを扱えるよう順次便利な記法を導入していく．“@”はそれが最初に呼ばれた位置をポイントする．“−”はそれに続く変数のストリングに結びつくものの，そのストリングは未消化のままにしておく．そして“&”は両側のストリングを同時に満たすようなストリングを指す，などなどである．

DCGをDNAの解析に運用しようというのは単に

応用例としておもしろいというだけではない．DNAの長さというのは数十億とも言われる．自然言語の一文の音列の比ではない．並列アーキテクチャの進歩によるパーサの高速化という事態に最もよく応える実験材料として，これはずいぶん興味を引く問題ではなかろうか．

〔東条　敏（三菱総合研究所 情報技術開発部）〕