Vol.37 No.10　　情 報 処 理　　Oct. 1996

解 説

ゲノム情報

5. 計算論的言語理論とDNA計算†

横森 貴††　小林 聡††

1. はじめに

生物のすべての遺伝情報を含んでいるDNA配列はランダムに並べられた記号列などではけっしてなく，生命情報を伝えかつ達成するために，ある明確な規則に基づいて構成されている暗号文である．そこで，形式言語理論の数理的な枠組みを援用して，文法によってDNA配列をモデル化したり予測したりすることが考えられる．本稿では，特にRNAの二次構造予測に関する最近の試みについて考察する．さらに，最近注目を集めているDNA計算(すなわち，DNAの組換え規則に基づく計算)の理論と実験に関して，その動向を概説する．

2. 木文法によるRNA二次構造のモデル化と予測

近年，RNAがもつ機能およびその構造に関する研究の進歩によりRNA配列の構造が遺伝子の発現制御や蛋白質への翻訳などにおいて非常に重要な役割を果たしていることがわかりつつある[1]．しかしながら，RNAの構造をX線解析やNMRにより決定するのは，困難であるのが現状であり，計算機によりRNAの(二次)構造を予測する手法の開発が望まれている．

RNA二次構造の予測問題は，組合せ最適化問題として定式化され，動的計画法やニューラルネットワークなどの手法が適用されている[2]が，一方で，文法を利用したRNA二次構造の予測・学習手法が最近注目されている．本章では，特に，この文法に基づくアプローチについて解説する．

2.1 RNA二次構造と文法

RNAは，アデニン(A)，ウラシル(U)，グアニン(G)，シトシン(C)という4種類の塩基が糖，リン酸と結合してできたリボヌクレオチドを基本単位とし，それが多数連結してできた一本鎖状の高分子である．4種類の塩基は，AとU，およびGとCの間において，水素結合により塩基対を構成し，とくに，RNA分子の内部に逆繰り返し構造が存在する場合には，その部分が相補的に水素結合してステム構造(図-1(a)参照)とよばれるエネルギー的に安定した構造をとることが多い．RNA分子において塩基対を構成している部位を予測する問題が，RNA二次構造予測である．

まず，図-1(a)のステム構造を持ち得る塩基配列の集合を考える．以下本稿においては，有限アルファベットΣ上の文字列の集合を言語とよび，空文字列をλ，Σ上の文字列すべてからなる集合をΣ^*で表す．また，$\Sigma^+ = \Sigma^* - \{\lambda\}$とする．

アルファベット$\Sigma = \{\mathbf{a},\mathbf{u},\mathbf{g},\mathbf{c}\}$によって4種類の塩基を表し，各塩基記号上の上線によって，Watson-Crickの塩基対を表すことにする．また，任意の語(塩基配列)$w = x_1 x_2 \cdots x_n (x_i \in \Sigma, 1 \leq i \leq n)$に対し，$w^R = x_n x_{n-1} \cdots x_1$とする．たとえば，$\overline{acggaucg}^R = cgauccgu$となる．このとき，ステム構造をもつ塩基配列の集合は，$\{uxv\bar{x}^R t \mid u,v,t \in S^*, x \in S^+\}$で表され，文脈自由文法で表現できることが容易にわかる．

しかしながら，文脈自由文法は図-1(b)のシュードノット構造☆を表現できないため，RNA二次構造の表現手段としては，必ずしも十分ではない[1]．そこで，シュードノット構造を含むような二次構造を柔軟に表現できる文法が必要になる．

† Computational Linguistics and DNA Computation by Takashi YOKOMORI and Satoshi KOBAYASHI (Dept. of Computer Science and Information Mathematics, Univ. of Electro-Communications).

†† 電気通信大学情報工学科

☆ Psuedo-knot(擬似結び目)の形をとるRNA二次構造の一例

2.2 木文法による RNA 二次構造のモデル化と予測

本節では，RNA 二次構造の表現手段として著者らが提案している木文法 TAG^2_{RNA} を概説する．詳しい定義などは文献3)を参照されたい．

TAG^2_{RNA} は，初期木とよばれる木に順次 adjunct tree とよばれる木を接合しながら木を生成するシステムと考えられる．木の接合は，タグ(*)がつけられている頂点に対して常に行われ，同じ adjunct tree によって同時に導入された葉に現れる塩基対は，水素結合をしているものと解釈される．たとえば，シュードノット構造をもつ塩基配列の導出の様子が図-1(c)に示されている．これは，文脈自由文法では表現できない構造である．

この文法には，n を塩基配列の長さとしたとき，$O(n^5)$(文法にある制約を加えると $O(n^4)$)時間の構文解析アルゴリズムが存在する[3]．この構文解析アルゴリズムにより，与えられた配列が特定の二次構造をもつか否かを判断できる．

この木文法による RNA 二次構造予測は，各 adjunct tree に自由エネルギー[4]に相当するペナルティを与え，ペナルティの総和が極小となる構文解析木を求めることで実行することができる．これにより，TAG^2_{RNA} で表現できるような二次構造の中で，最適な二次構造を探索することが可能となる．実験結果によると，交差する依存性をもつ二次構造を含んでいても，既知の構造とよく一致する結果が得られている[3]．図-2 に 16S-rRNA(500 番目～569 番目の塩基)の二次構造予測結果が示されている．実線は既知の二次構造であり，点線は TAG^2_{RNA} による予測結果である．ごく短い(長さ2の)ステム構造が余分に予測される以外は，ほぼ一致する結果となっている．

さらに，RNA 二次構造の文法による学習問題をとり扱った研究例として，tRNA の同定・予測に確率文脈自由文法の学習アルゴリズムを用いた榊原らの研究[5]があげられる．また，興味深い他のアプローチとして，シュードノットを表現するために2つの確率文脈自由言語の共通集合を用いる手法が提案されている[6]．

文法を用いて遺伝子配列やアミノ酸配列を言語理論的あるいは計算論的な観点から解析する試みは，Searls[7]に詳しく，多くの文献が参照されているので参考にされたい．

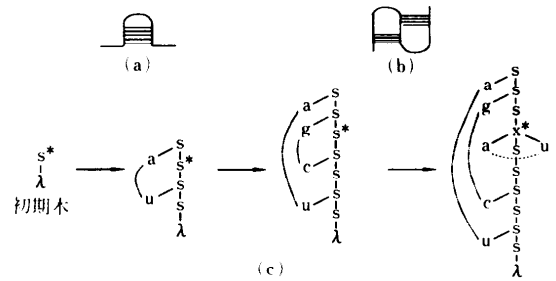

図-1 RNA 二次構造と TAG^2_{RNA} による導出

3. DNA 計算―その理論と実験

本章では，最近注目を集めている DNA 分子の組換え機構に基づく新しい計算モデルについて概説する．DNA 分子は，ある特定の記号列によって切断され，また連結(結合)されることにより複雑な構造の1次元配列へと成長しうる．この組換えの機構を利用した新たな計算モデルを提案する研究が進展している．一方これとは独立に，DNA 分子を大量に増幅(コピー)させる技術などを利用して，並列処理を行わせる考えを基本にして，現実的な時間内では解けないであろうと予想されているいわゆる NP 完全問題を解く試みが進行している．ここでは，これらの動向を概説する．

3.1 DNA 計算の形式モデル―スプライシング・システム

A-T および C-G のいわゆる相補性塩基の性質から，DNA 配列はある特定の配列を認識する制限酵素によって切断され，またその断面はリガーゼとよばれる酵素により連結されることがよく知られている．たとえば，制限酵素 *Eco*RI は配列

$$\begin{matrix}X_1GAATTCX_2\\ X_1CTTAAGX_2\end{matrix} \text{ を } \begin{matrix}X_1G\\ X_1CTTAA\end{matrix} \text{ と } \begin{matrix}AATTCX_2\\ GX_2\end{matrix}$$

に，また配列

$$\begin{matrix}Y_1GAATTCY_2\\ Y_1CTTAAGY_2\end{matrix} \text{ を } \begin{matrix}Y_1G\\ Y_1CTTAA\end{matrix} \text{ と } \begin{matrix}AATTCY_2\\ GY_2\end{matrix}$$

とに切断 (splicing) する．ここに，X'は任意の配列Xの相補配列を表す．これらは，リガーゼによって，たとえば，

$$\begin{matrix}X_1GAATTCY_2\\ X_1CTTAAGY_2\end{matrix} \text{ と } \begin{matrix}Y_1GAATTCX_2\\ Y_1CTTAAGX_2\end{matrix}$$

のように2つの新たな配列を生成し得る．

Head[13]はこの現象を数学的に解析するために，以下のようなモデルを提案した．スプライシング・ルール $r=u_1 \# u_2 \$ u_3 \# u_4$（各 u_i は有限アルファベット V 上の記号列，#，$ は V に含まれない特殊記号）と V 上の記号列 x,y,w,z に対してスプライシング操作 $\vdash_r$ を

$$(x,y) \vdash_r (w,z) \text{ iff}$$
$$x=x_1u_1u_2x_2, y=y_1u_3u_4y_2$$
$$w=x_1u_1u_4y_2, z=y_1u_3u_2x_2$$
$$(x_1,x_2,y_1,y_2 \in V^*)$$

で定義する．x,y は各々サイト u_1u_2, u_3u_4 においてスプライスされたという．（サイト u_1u_2 と u_3u_4 は上記の例では GAATTC に対応している．）さて，V とその部分アルファベット Σ が与えられたとき，$H=(V,\Sigma,A,R)$ をスプライシング・システム（H-システム）とよぶ．ここに $A \subseteq V^*$, $R \subseteq V^* \# V^* \$ V^* \# V^*$ である．H によって生成される言語 L(H) を

$$L(H)=\sum_{i \geq 0} S^i(A) \cap \Sigma^*$$

ここで

$$S(A)=\{w \in V^* \mid (x,y) \vdash_r (w,z), x,y \in A, r \in R\}$$
$$S^0(A)=A,$$
$$S^i(A)=S^{i-1}(A) \cup S(S^{i-1}(A))(i \geq 1)$$

とする．たとえば，A を初期（原始）段階における単純な DNA の配列集合，R を制限酵素の集合，$V=\Sigma=\{A,C,G,T\}$ とする $H=(V,\Sigma,A,R)$ を考えると，L(H) は制限酵素とリガーゼとのスプライシングによって A から生成される（一本鎖）DNA 配列の集合となる．これにより，分子生物学的な興味から DNA が単純な配列からどのようにして複雑な配列へと進化したのかが形式言語理論の数学モデルで解析できることになり，また計算機科学的な見地からはスプライシング操作による計算モデルとしての能力が研究対象となる．前者の立場からの結果として，文献20）があり，後者の興味からの研究は以下のように進展している．与えられた H-システム (V,Σ,A,R) において R は $V^* \# V^* \$ V^* \# V^*$ の部分集合であることに留意しよう．また，言語理論的な終端アルファベット Σ の導入に関する分子生物学的な動機として，進化の過程で特定の性質（配列）のみが優性として生き残ることが考えられるが，（3.2 節で後述するように）DNA 計算の実行という見地からするとこの

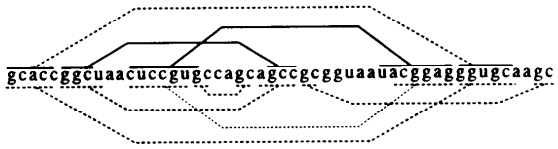

図-2 TAG^2_{RNA} による二次構造予測結果

“フィルター機能”は分子生物学的な実験技術によって実現できることに留意されたい．A,R を各々言語族 F_1,F_2 から選ぶとき生成される言語 L(H) 全体の族を $EH(F_1,F_2)$，その特殊な場合として $V=\Sigma$ のとき生成される言語族を $H(F_1,F_2)$ と表すことにする．興味深い結果として

定理1

$H(Fin,Fin) \subset REG = EH(Fin,Fin) \subset EH(Fin,REG) = RE$ が成り立つ．ただし，Fin は有限言語の族，REG は正則言語族，RE は帰納的可算言語族である．

R は一般に無限集合となりえるが，スプライシング・ルールが無限個ゆるされるという状況は非現実的な設定であるので可能ならば有限としたい．この条件のもとで任意の帰納的可算言語を生成させる（すなわち，チューリング機械の能力と等価である）ためいくつかのモデルが提案されている．まず，スプライシングを受ける記号列にある種の“寿命”をマルチセットという形で定式化したとき，生成される言語族を一般に $EH(mF_1,F_2)$ で表す．このとき，$EH(mFin,Fin) = RE$ であることが知られている[19]．また，環状 DNA を想定した環状スプライシング・システムが提案され，それが A,R とも有限言語族のとき生成する言語族が RE と一致することが示されている[21]．これらは，その言語生成能力がチューリング機械と等価であることのみならず，（万能チューリング機械と同様の意味で）各々万能のスプライシング・システムが構成できることも示していて，プログラミング可能な DNA コンピュータへの理論的な計算モデルでもあるという点で興味深い．今後の課題として，これらの計算モデルが分子生物学的にどのように実現可能であるかが論じられることになろう．

3.2 DNAコンピュータへ向けて—いくつかの実験的試み

ことの起こりは，Adleman☆が分子生物実験的手法により有向ハミルトンパス問題(DHP)の解法を提示したことに始まる[8]．DHPとは，有向グラフにおいて，指定された始点から終点へ至るパスで他のどの点もちょうど一度だけ通るようなものが存在するか？ を決定する問題であり，NP完全問題の1つとしてよく知られている．分子生物学における基本的実験技術を習得したのち，彼は頂点数7のDHPを以下の手順で解くことを試みた：

1. 4文字のDNAアルファベット上の記号列としてユニークに定まるように各頂点と有向辺とを長さ20の(一本鎖)DNAに符号化する．
2. 符号化されたDNAの集まりを混ぜ合わせ，可能なパス(DNA)の集合を生成する．(ここでは二本鎖DNAになる．)
3. 始点が符号化されたDNAで始まり終点が符号化されたDNAで終わる長さ140(＝20×7)のパスのみを選別する．
4. さらに各頂点を一度しか含まないパスのみを選別する．
5. もし，この時点で残っているパスが存在するならばそれは解であり，そのようなパスがなければ解はないと判定する．

問題をDNA配列へ符号化するという奇抜なアイディアもさることながら，ここで重要なのはNP完全問題を解く手段として，DNA分子のもつ指数的並列性に目をつけたという点である．すなわち，2.において大量の可能なパスを小さな試験管内に小量の溶液として生成させることにより，並列ランダム探索を容易に実現していることである．実験においては，ある特定のDNA配列を指数的に増幅させる(コピーを作る)ことを可能にするPCR(ポリメラーゼ連鎖反応)とよばれる分子生物実験における技法がいたるところで有効に用いられている．(たとえば，3.および4.における特定の性質をもつ配列の選別に，また5.においてミクロの世界での計算結果を認識可能にするため，解のDNAパスを増幅させるのに利用されている．)また，3.においては始点と終点を正しくもつパスのみを選択するのに，DNAの“相補配列性”が重要な役割を果たしている．

これらの分子生物学的な実験の詳細は参考文献に委ねるとして，計算機科学の視点からAdlemanの仕事はどのようにとらえられるであろうか．Kaplan[14]，有田[22]らの追実験結果から，現時点であきらかな問題点は以下のものがある：

(1) 実験結果が誤りを含み易く解の信頼性に乏しい．
(2) Adlemanの手法で実際的なサイズの問題(たとえば，サイズ200のDHP)を解くとすると，地球の重量以上の化学材料が必要となる．(Hartmanis[12]の批評)
(3) DNAによる解法は汎用性がないのではという点．

(2)は重要な点であるが，これに関する決定的な解決法はまだ知られていない．

(1)と(3)に関しては，汎用DNAコンピュータは実現可能か？ という視点からの研究が現在進行している．たとえば，Beaver[10]は，チューリング機械の動作をDNA上で模倣し，超並列性のもとでPSPACE問題のクラスが多項式時間でDNA計算可能であることを示している．一般に，DNA計算における分子生物学的操作では，1)分離：ある特定の部分配列を含む配列集合(配列の入った試験管)のみを分離する，2)マージ：2つの配列集合を混ぜる，3)検知：配列集合が空か否かを調べる，4)増幅：配列集合のコピーを作る，が考えられる．Adleman[9],Lipton[16]は1),2),3)のみを用いてNP問題が直接的に効率よく解けることを主張している．また，Amosら[17]は，これらに基づく誤り確率の少ないDNAコンピュータの実現法を提案している．(なお，有用な情報源として以下のサイトがある：http://www.cs.princeton.edu/dabo/biocomp.html)

生物コンピュータの発想は，1950年代おわりのFeynmanのmicroscopic computerの構想にまでさかのぼることができるが，DNAコンピュータは現在のコンピュータに対する補完的なものと考えるのがよいのかもしれない．いずれにせよ，DNAコンピュータの可能性に結論をだすのにはまだ時期早尚であろう．真空管の怪物もどき人類最初のコンピュータが作られた半世紀前を思い起

☆理論計算機科学者で(RSA系とよばれる)公開鍵暗号系の提案者の1人として知られている．

こしてみるとき，それを今日のラップトップにまで育てあげたような，長期的な視野にたって暖かく見守りかつ育て上げる人間の英知を，この“試験管コンピュータ”に対しても期待してみてはどうであろうか．

謝辞 3.2節の素稿に目を通し有益なコメントとともに貴重な資料をいただいた有田正規氏(東京大学)に感謝いたします．

参考文献

1) Dam, E., Pleij, K. and Draper, D. : Structural and Functional Aspects of RNA Pseudoknots,Biochemistry, Vol.31, No. 47, pp.11665-11676(1992).

2) Steeg, E.W. : Neural Networks, Adaptive Optimization, and RNA Secondary Structure Prediction,In Artificial Intelligence and Molecular Biology,edited by L. Hunter,AAAI Press/MIT Press, pp.121-160(1993).

3) Uemura, Y., Hasegawa, A., Kobayashi, S. and Yokomori, T. : Grammatically Modeling and Predicting RNA Secondary Structures, In Proc. of Genome Informatics Workshop VI, Universal Academy Press, pp.67-76 (1995).

4) Turner, D. H., Sugimoto, N., Jaeger, J. A., Longfellow, C. E.,Freier, S. M. and Kierzek, R. : Improved Parameters for Prediction of RNA Structure,Cold Spring Harb. Symp. Quant. Biol., Vol. 52, pp.123-133 (1987).

5) Sakakibara, Y., Brown, M., Underwood, R. C., Mian, I. S. and Haussler, D. : Stochastic Context-free Grammars for tRNA Modeling, Nucleic Asids Res., Vol.22, pp.5112-5120 (1994).

6) Brown, M. and Wilson, C. : RNA Pseudoknot Modeling Using Intersections of Stochastic Context Free Grammars with Applications to Database Search, Proc. of 1st Pacific Symposium on Biocomputing (Eds. L.Hunter and T.Kein), World Scientific Publishing Co. (1996).

7) Searls, D.B. : The Computational Linguistics of Biological Sequences,In Artificial Intelligence and Molecular Biology,edited by L. Hunter,AAAI Press/MIT Press, pp.47-120 (1993).

8) Adleman, L. : Molecular Computation of Solutions to Combinatorial Problems,Science, Vol.266, pp.1021-1024 (1994).

9) Adleman, L. : On Constructing A Molecular Computer, manuscript (1995).

10) Beaver, D. : A Universal Molecular Computer, Penn State Univ. Tech. Report CSE-95-001 (1995).

11) Ferretti, C. and Kobayashi, S. : DNA Splicing Systems and Post Systems, Proc. of 1st Pacific Symposium on Biocomputing (Eds. L.Hunter and T.Kein), World Scientific Publishing Co. (1996).

12) Hartmanis, J. : On the Weight of Computations, EATCS Bulletin, vol.55,pp.136-138 (1995).

13) Head, T. : Formal Language Theory and DNA: An Analysis of the Generative Capacity of Specific Recombinant Behaviors, Bull. Math. Biology, Vol.49, pp.737-759 (1987).

14) Kaplan, P., Cecchi, G. and Libchaber, A. : Molecular Computation : Adleman's Experiment Repeated, Tech.Rep., NEC Res.Inst. (1995).

15) Lipton, R. J. : DNA Solution of Hard Computational Problem, Science, Vol.268, pp.542-545 (1995).

16) Lipton, R. J. : Speeding Up Computations via Molecular Biology, Manuscript (1995).

17) Amos, M., Gibbons, A. and Hodgson, D. : Error-resistant Implementation of DNA Computations, Proc. of 2nd Annual Meeting on DNA Based Computers, June, Princeton, pp.87-101(1996).

18) Paun, G. : Computationally Universal Distributed Systems Based on the Splicing Operations, Manuscript (1995).

19) Paun, G. : A Challenge for Formal Language Theorists, EATCS Bulletin, Vol.57, pp.183-194 (1995).

20) Yokomori, T. and Kobayashi, S. : DNA Evolutionary Linguistics and RNA Structure Modeling : A Computational Approach, Proc. of INBS-IEEE Conference, Washington D.C., pp.38-45 (May 1995).

21) Yokomori, T., Kobayashi, S. and Ferretti, C. : On the Power of Circular Splicing Systems and DNA Computability, Report CSIM 95-01, Univ. of Electro-Communications, Tokyo (1995).

22)有田正規，渡辺真理：DNAコンピューティング，情報処理学会プログラミング研究会研究報告(May 1996).

(平成8年6月26日受付)

横森 貴（正会員）

1951年生．1974年東京大学理学部数学科卒業．1979年同大学理学系大学院博士課程修了．理学博士．同年産業能率大学助手．1983年富士通（株）入社．1989年電気通信大学情報工学科助教授．1981-82年カナダ・マクマスタ大学，1982-83年米国ペンシルベニア大学ポストドクトラルフェロー，1995-96年カナダ・ウォータールー大学客員助教授．形式言語理論，計算論的学習理論，ゲノムサイエンスの研究に従事．電子情報通信学会，人工知能学会，EATCS，LA，IEEE各会員．

小林 聡（正会員）

1988年東京大学工学部航空学科卒業，1993年同大学院工学系研究科博士課程修了．同年電気通信大学情報工学科助手．博士（工学）．計算論的学習理論，形式言語理論，遺伝子情報処理の研究に従事．人工知能学会，電子情報通信学会，EATCS各会員．