様式 C-19

科学研究費補助金研究成果報告書

平成２１年　５月２０日現在

研究種目：　基盤研究（C）
研究期間：　2007 年度～2008 年度
課題番号：19560374
研究課題名（和文）　情報理論的な手法を用いた DNA 計算の漸近解析に関する研究

研究課題名（英文）　Asymptotic Analysis of DNA Computation using Information Theoretic Method

研究代表者　鎌部　浩（KAMABE HIROSHI）
岐阜大学・工学部・准教授
研究者番号　80169614

研究成果の概要：
DNA 系列は４記号の系列と考えることができるが，工学的な処理の容易さや構造の安定性などを考慮すると，いくつかの制約を満たしている DNA 系列だけが工学的に利用できると考えることができる. DNA 計算では計算の入力や途中の計算結果を DNA 系列の集合として保存するが，これらも前述の制約を満たしていることが要請される．したがって DNA 計算などで系列を利用するためには，そもそもそうした制約を満たす系列がどれくらい存在するのかを知る必要がある．これは情報理論における「入力制約を持つ通信路の容量」に対応する．しかしながら，DNA 系列が満たすべき制約はこれまで情報理論が対象としてきた制約のクラスには入らない．したがって，本研究では計算機実験などの方法を試みたがうまくいかず，いろいろ調査した結果，この問題を解くためにうまく応用できる理論があることがわかった．その方法と計算機実験を組みわせることによって，原理的には容量を計算できることがわかった．この方法を用いてより現実的な制約の容量を計算した.

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,000,000	300,000	1,300,000
2008 年度	500,000	150,000	650,000
年度			
年度			
年度			
総　計	1,500,000	450,000	1,950,000

研究分野：　工学
科研費の分科・細目：　　電気電子工学・　通信・ネットワーク工学
キーワード：　DNA 計算，情報理論，通信路容量，母関数

１．研究開始当初の背景

DNA 系列を大量に複製する方法として PCR 法が開発された. Adleman はその方法を用いて計算ができることを，実際に DNA を使って計算を行うことで示した．その後 DNA 系列は計算だけではなく，データ記録などでも利用することが検討さるようになってきた．DNA 系列は原理的には４文字の系列であると考えることができるが，工学的に利用するときには，その利用目的に応じていくつかの制約を満たす系列だけを使うことになると予想される．このため，これまではこれらの制

約を満たすような系列を生成する方法が中心的に研究されていた．しかし，研究を進めるにつれて符号の良さを評価する際に，制約を満たす系列の数の漸近的な振る舞いなどの情報理論的な量を求める必要性を感じるようになってきた．

2. 研究の目的

DNA 計算や DNA を用いた記録などで DNA 系列を利用する場合には，利用される系列は一般にはある制約を満たさなければならない．しかしながら，これらの制約は，計算を進めていくために必要な性質を列挙しているだけであるため，それらの条件を満たす系列が，実際に計算を進めるために十分なほどたくさんあるかどうかは自明ではない．制約を満たす系列の長さに対する指数関数的な増加の割合は，情報理論的には入力制約のある通信路の通信路容量に対応する．しかしながら，DNA 系列が満たすべき制約は，情報理論でこれまで研究されてきた制約クラスよりも広いクラスに含まれる制約であり，一般にはこれまでの手法が適用できない．そこで，DNA 系列の通信路容量を計算する方法について研究し，その結果を DNA 計算のための符号構成の問題に適用することを目的とした．

本研究で対象とした制約は以下通りである．DNA 系列は 4 文字からなる系列であるとみなすことができる．しかし均一な温度で反応させたり，所望の構造を持つような系列は，以下に述べるような制約 (のいくつか) を満していなければならないことが予想される．

1. C(Cytosine) と G(Guanine) が含まれている数が，反応温度に影響する．したがって，均一は反応を期待しようとすると，C と G とが含まれている割り合いが同じ系列だけを集めて利用したいと考えられる．
2. DNA の一重鎖で「ヘアピン構造」と呼ばれる構造を実現するためには，C と G のペア，及び T(Thymine) と A(Adenine) のペアが，括弧の対応が取れるのと同じように，あるところを中心として対称になっている必要がある．
3. 前項と同じ状況で，ヘアピンの構造には直接には関与しない冗長な分子が間にはさまることがある．このような分子は，ヘアピンの構造には関係せず,「宙に浮いている」ことになる．
4. 前項までの制約に，工学でよく出てくる (d,k)-制約を重畳した制約を考える．

これらの制約の容量 (制約を満す系列は長さとともに指数関数的に増加していくが，その指数関数的な増加の係数) を求めることが目的である．

3. 研究の方法

以下の二つの方法で通信路容量を求めることを試みた.（1）計算機によるシミュレーション（2）母関数を用いた解析的な方法と計算機を使った計算とを組み合わせる．

4. 研究成果

以下のステップで，DNA 系列のための制約の容量を計算できることがわかった．

- 制約を文脈自由文法として表現する．
- DSV 法によってその文法が生成する有限系列の個数（制約を満たす有限系列の個数）に関する母関数を求める．
- その母関数の極を求め，それから容量を求める．

最後のステップは，実際的な制約に対しては，計算機によって数値的に求めることになる場合が多い．　上記の方法を用いて，DNA 系列の制約と工学的な制約とを組み合わせた場合の制約の容量を数値的に求めた．

以下にその具体例を示す．ただし，定式化と結果のみを述べ，詳細な計算と定式化の正しさの証明は省略する．

4.1 ヘアピン構造の制約

C と G 及び T と A の二組によるヘアピン構造は，Dyck シフトと呼ばれる数学的な構造に対応する．Krieger は，この制約の容量 (数学的には位相的エントロピー) が $\log 3$ であることを解析的に示している．しかし Krieger はその後，他の著者の論文へのコメントとして，DSV 法と呼ばれる方法でも計算できると述べている．その概略を，上に述べたステップに沿って示す．

まず制約を文脈自由文法に従って書き下す．それは以下のようになる．

$$D \rightarrow \sum_{i=1}^{n} \alpha_i \cdot D \cdot \beta_i \cdot D + \varepsilon \quad (1)$$

$$D_L \rightarrow \sum_{i=1}^{n} D \cdot \beta_i \cdot D + \sum_{i=1}^{n} D_L \cdot \beta_i \cdot D \quad (2)$$

$$D_R \rightarrow \sum_{i=1}^{n} D \cdot \alpha_i \cdot D + \sum_{i=1}^{n} D \cdot \alpha_i \cdot D_R \quad (3)$$

$$D_B \rightarrow \sum_{i=1}^{n} D_L \cdot \alpha_i \cdot D + \sum_{i=1}^{n} D_B \cdot \alpha_i \cdot D \quad (4)$$

これらすべてが，求める系列を表現する変数である．

D，D_L，D_R，D_B の各々の母関数 $D(t)$，$D_L(t)$，$D_R(t)$，$D_B(t)$ は，DSV 法によって求めることができる．実際，計算によって以下を得ることができる．

$$\begin{aligned} D(t) &= \frac{2}{1+\sqrt{1-4nt^2}} \\ &= \frac{1-\sqrt{1-4nt^2}}{2nt^2}, \end{aligned} \tag{5}$$

$$D_L(t) = D_R(t) = \frac{tD(t)^2}{1-tnD(t)}, \tag{6}$$

$$D_B(t) = \frac{t^2D(t)^3}{(1-ntD(t))^2}. \tag{7}$$

さらに $\bar{t} = 1/(1+n)$ のとき，上式の分母は 0 になる．したがって，$\widetilde{D}(t) = D(t)+D_L(t)+D_R(t)+D_B(t)$ と置くと，$\widetilde{D}(t)$ は $t = 1/(1+n)$ のとき極を持つことがわかる．さらによく調べると以下がわかる．

$$\frac{1}{n+1} = \sup\{r \geq 0 : \widetilde{D}(t) \text{ は } z\ (|z| < r) \text{ で解析的 }\} \tag{8}$$

母関数に関する理論から上の文脈自由文法で与えられる言語のエントロピーは $\log(n+1)$ となる．

4.2 中性的な記号のあるヘアピン構造の制約

C-G や T-A のようなペアを組むという制限がない場合の制約は，Motzkin シフトと呼ばれる数学的な対象 (記号力学系) に対応する．

この場合の制約は以下のように表現できる．

$$M \rightarrow \varepsilon + \sum_{i=1}^{n} \alpha_i \cdot M \cdot \beta_i \cdot M + \sum_{j=1}^{m} 1_j \cdot M, \tag{9}$$

$$M_R \rightarrow \sum_{i=1}^{n} M \cdot \alpha_i \cdot M_R + \sum_{i=1}^{n} M \cdot \alpha_i \cdot M, \tag{10}$$

$$M_L \rightarrow \sum_{i=1}^{n} M_L \cdot \beta_i \cdot M + \sum_{i=1}^{n} M \cdot \beta_i \cdot M, \tag{11}$$

$$M_B \rightarrow \sum_{i=1}^{n} M_L \cdot \alpha_i \cdot M + \sum_{i=1}^{n} M_B \cdot \alpha_i \cdot M. \tag{12}$$

M に対応する母関数は以下のように与えられる．

$$M(t) = \frac{1-mt-\sqrt{(mt-1)^2-4nt^2}}{2nt^2}. \tag{13}$$

他の母関数は以下のように与えられる．

$$M_L(t) = M_R(t) = \frac{ntM(t)^2}{1-tnM(t)}. \tag{14}$$

$$M_B(t) = \frac{nM(t)M_L(t)}{1-tnM(t)} = \frac{n^2tM(t)^3}{(1-tnM(t))^2} \tag{15}$$

また，計算によって以下を導くことができる．

$$\sup\{r \geq 0 : \widetilde{M}(t) \text{ is analytic at } z \text{ with } |z| < r\}$$

$$= \frac{1}{n+m+1} \tag{16}$$

ただし，$\widetilde{M}(t) = M(t)+M_L(t)+M_R+M_B(t)$ である．よって，母関数に関する理論から容量は $\log(n+m+1)$ となることがわかる．

4.4 ヘアピン構造と (d,k)-制約との組合せ

DNA 系列を工学的に計算や記録に利用しようとすると，工学的な制約も同時に満足するような系列が必要になると考えられる．そこで，ここでは (d,k)-制約とヘアピン構造との組合せについて考える．

DNA 系列に情報を記録すると，それを読み出す仕組みが必要になる．そこで，前述の制約に加えて

[[と]]

の二つが出てこないような制約を課した場合について考察した．このような場合には，原理的には前述の方法で容量を計算できる．

まず，制約を記述する文法は以下のように記述できる．

$$D \rightarrow \varepsilon + (\cdot D \cdot) \cdot D + [\cdot E \cdot] \cdot D, \tag{17}$$

$$E \rightarrow (\cdot D \cdot) + (\cdot D \cdot) \cdot D \cdot (\cdot D \cdot) + \varepsilon. \tag{18}$$

対応する母関数は以下のように与えられる．

$$D(t) = 1 + t^2D(t)^2 + t^2E(t)D(t), \tag{19}$$

$$E(t) = t^2D(t) + t^4D(t)^3 + 1. \tag{20}$$

$D(t)$ は，0.3816 に極を持つことがわかる．

次に，容量を求めるために，前述の文法で記述された単語の語頭部分と語尾部分の文法を記述する．さらに中央部分の語の文法も記述すると以下のようになる．

$$D_R \rightarrow D \cdot \widetilde{D}_R \tag{21}$$

$$\widetilde{D}_R \rightarrow (\cdot P + [\cdot B \tag{22}$$

$$\widetilde{D} \rightarrow (\cdot D \cdot) + [\cdot E \cdot] \cdot D \tag{23}$$

$$\widetilde{E} \rightarrow (\cdot D \cdot) + (\cdot D \cdot) \cdot D \cdot (\cdot D \cdot) \tag{24}$$

$$P \rightarrow (\cdot P + [\cdot B + \widetilde{D} + \widetilde{D} \cdot \widetilde{D}_R \tag{25}$$

$$B \rightarrow (\cdot P + \widetilde{E} + \widetilde{E} \cdot \widetilde{D}_R \tag{26}$$

その母関数は以下のようになる．

$$\widetilde{D}_R(t) = tP(t) + tB, \tag{27}$$

$$P(t) = tP(t) + tB(t) + \widetilde{D}(t) + \widetilde{D}(t)\widetilde{D}_R(t), \tag{28}$$

$$B(t) = tP(t) + \widetilde{E}(t) + \widetilde{E}(t)\widetilde{D}_R(t). \tag{29}$$

$\widetilde{D}_R(t)$ は以下の方程式の解として与えられる．

$$\begin{pmatrix} 1-t(1+\widetilde{D}(t)) & -t(1+\widetilde{D}(t)) \\ -t(1+\widetilde{E}(t)) & 1-t\widetilde{E}(t) \end{pmatrix} \begin{pmatrix} P(t) \\ B(t) \end{pmatrix} = \begin{pmatrix} \widetilde{D}(t) \\ \widetilde{E}(t) \end{pmatrix}. \quad (30)$$

必要なのは $D_B(t)$ であり，これは $D_B(t) = D(t)D_R(t)^2$ で与えらる．前述の議論を用いると，$D_B(t)$ からエントロピーは次のように求まる．

$$\log_2 1/0.377997313629 = 1.40355211342485.$$

次に D_2 と (1,2) RLL 制約との組合せについて考える．これを $D^{RLL}_{1,2}$ と書く．この制約のための文法は以下のようになる．

$$\begin{aligned} U &\rightarrow \varepsilon + [\cdot U_1 \cdot] \cdot U + (\cdot V_1 \cdot) \cdot U && (31)\\ U_0 &\rightarrow (\cdot U_1 \cdot) + (\cdot U_1 \cdot) \cdot U \cdot (\cdot U_1 \cdot) + \varepsilon && (32)\\ U_1 &\rightarrow (\cdot V_0 \cdot) + (\cdot V_0 \cdot) \cdot U \cdot (\cdot V_0 \cdot) + [\cdot V_1 \cdot] \\ &\quad + [\cdot V_1 \cdot] \cdot U \cdot [\cdot V_1 \cdot] + \varepsilon && (33)\\ V_0 &\rightarrow [\cdot V_1 \cdot] + [\cdot V_1 \cdot] \cdot U \cdot [\cdot V_1 \cdot] + \varepsilon && (34)\\ V_1 &\rightarrow [\cdot U_0 \cdot] + [\cdot U_0 \cdot] \cdot U \cdot [\cdot U_0 \cdot] + (\cdot U_1 \cdot) \\ &\quad + (\cdot U_1 \cdot) \cdot U \cdot (\cdot U_1 \cdot) + \varepsilon && (35) \end{aligned}$$

さらに，これらに対応して，以下の多項式方程式を考える．

$$\begin{aligned} U(t) &= 1 + t^2 A_1(t) U(t) + t^2 A_1(t) U(t) && (36)\\ A_0(t) &= t^2 A_1(t) + t^4 A_1(t)^2 U(t) + 1 && (37)\\ A_1(t) &= t^2 A_0(t) + + t^4 A_0(t)^2 U(t) + t^2 A_1(t) \\ &\quad + t^4 A_1(t)^2 U(t) + 1 && (38) \end{aligned}$$

エントロピーを計算するためには，これらの規則で生成される語の語頭部分，語尾部分，中央部分の規則を与える必要がある．それは以下のように与えられる．ここで，S_R は U の語頭部分である．

$$\begin{aligned} S_R &\rightarrow [\cdot P_1 + (\cdot B_1 && (39)\\ S_0 &\rightarrow (\cdot V_0 \cdot) + (\cdot V_0 \cdot) \cdot U \cdot (\cdot V_0 \cdot) && (40)\\ S_1 &\rightarrow (\cdot V_0 \cdot) + (\cdot V_0 \cdot) \cdot U \cdot (\cdot V_0 \cdot) + [\cdot V_1 \cdot] \\ &\quad + [\cdot V_1 \cdot] \cdot U \cdot [\cdot V_1 \cdot] && (41)\\ E_0 &\rightarrow [\cdot U_0 \cdot] + [\cdot U_0 \cdot] \cdot U \cdot [\cdot U_0 \cdot] && (42)\\ E_1 &\rightarrow [\cdot U_0 \cdot] + [\cdot U_0 \cdot] \cdot U \cdot [\cdot U_0 \cdot] && (43)\\ &\quad + (\cdot U_1 \cdot) + (\cdot U_1 \cdot) \cdot U \cdot (\cdot U_1 \cdot) && (44)\\ P_1 &\rightarrow (\cdot B_0 + [\cdot B_1 + S_1 + S_1 \cdot S_R && (45)\\ P_0 &\rightarrow (\cdot P_1 + S_0 + S_0 \cdot S_R && (46)\\ B_0 &\rightarrow [\cdot B_1 + E_0 + E_0 \cdot S_R && (47)\\ B_1 &\rightarrow [\cdot P_0 + (\cdot P_1 + E_1 + E_1 \cdot S_R && (48)\\ S_R(t) &= 2tP_1(t) && (49)\\ S_0(t) &= t^2 A_0(t) + t^4 A_0(t)^2 U(t) && (50)\\ S_1(t) &= t^2 A_0(t) + t^4 A_0(t)^2 U(t) + t^2 A_1(t) \\ &\quad + t^4 A_1(t)^2 U(t) && (51)\\ P_1(t) &= \frac{S_1(t)}{1 - t(1 + t + 2S_1(t))} && (52) \end{aligned}$$

これらから，母関数を求め，その分母の零点を求めることによって，容量は以下のようなることが示せる．

$$\limsup_{n\to\infty} \frac{1}{n} \log[t^n] G(t) = \log 0.383000231 \quad (53)$$

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

(1) Hiroshi Kamabe, Context-Free Shifts and Shifts of Finite Type, The 2009 International Conference on Bioinformatics and Computational Biology(CD-ROM), July 14-17, Las Vegas, USA, 2008.

〔学会発表〕（計２件）

(1) 鎌部浩，文脈自由文法による制約と有限タイプの制約の組み合わせの表現，電子情報通信学会技術研究報告，IT－１０８（２０２），pp.19-24，２００８年９月，沖縄.

(2) 鎌部浩，RNA 系列に関する制約の容量，第３０回情報理論とその応用シンポジウム，１１月２７～３０日，沖縄，２００７．

〔図書〕（計０件）

なし

〔産業財産権〕

○出願状況（計０件）

なし

○取得状況（計０件）

なし

〔その他〕

なし

６．研究組織

(1) 研究代表者

鎌部浩 (KAMABE HIROSHI)
岐阜大学・工学部・准教授
研究者番号 80169614

(2) 研究分担者

なし

(3) 連携研究者

なし