様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年　5 月 18 日現在

機関番号：13701
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：21560394
研究課題名（和文）　DNA 系列が持つ制約の特徴量の計算とその符号構成への応用
研究課題名（英文）　Calculation of Characteristics of Constraints for DNA Sequences and its Applications to Code Construction
研究代表者
鎌部　浩（KAMABE HIROSHI）
岐阜大学・工学部・教授
研究者番号：８０１６９６１４

研究成果の概要（和文）：
本研究では、DNA 計算のために必要な系列の制約の容量の計算と、符号化について研究した。まず、文脈自由文法の形式で記述される DNA 系列の制約の容量を求めた。次に、分子プログラムが字句解析可能であるための制約の容量を、情報理論および記号力学系的な手法で求めた。これによって、符号の構成のための指針を得た。

研究成果の概要（英文）：
In this research we have investigated the calculation of the capacity of constraints for DNA computation and coding scheme for the constraints. We have obtained two kinds of results. One of them is how to calculate the capacity of constraint described as context free grammar. The second is how to calculate the capacity of constraints for parsing DNA programs uqiquely. From the last result we have obtained a guideline for constructing codes for the constraints.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２１年度	900,000	270,000	1170,000
２２年度	800,000	240,000	1040,000
２３年度	500,000	150,000	650,000
年度			
年度			
総　計	2,200,000	660,000	2,860,000

研究分野：情報理論
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：情報理論、通信路容量、計算論的言語理論、DNA コンピュータ

1. 研究開始当初の状況

Adleman によって DNA 系列を計算に利用できる事が示されて依頼，DNA 系列を用いた計算を行うための様々な計算方法が提案されてきた．当初は，DNA 系列を記号列としてとらえ，記号列の操作で計算を進める方法が数多く提案されていた．ところが，DNA を操作するために必要となる物理的および科学的な反応を円滑に進めるためには，DNA 系列に制約を課す必要があることが認識されるようになってきた．そのために様々な聖夜が提案されていた．しか

し，その制約を課すことによって，計算を進めていくのに十分な DNA 系列があるかどうかが問題になる．そこで，制約を満す系列の豊富さを表現する「 容量」の計算が重要な問題になってくる．DNA 計算が計算機科学の分野の研究課題であったため，これまでの研究では，多くが計算機科学的な方法で調べられていた．

2. 研究の目的

本研究では，これまで計算機科学の視点から研究されてきた，DNA のための制約符号の理論を，情報理論，符号理論，記号力学系の理論の視点から考察することである．さらに，それらの結果を用いて，具体的な符号の構成を与えることである．

3. 研究の方法

二つの方法で研究を行った．

(1) 一般に系列の制約は，記号力学系として定式化できる．そこで，これまで計算機科学的な方法で解析されてきた制約を記号力学系の理論を用いて，解析し容量を求める．

(2) DNA 系列は，A, T, G, C の四つの記号から成る系列であると考えることができる．系列の中の G と C の含有割合は，化学的な反応温度に密接に関係している．これが，信号が含む周波数成分の解析と密接に関係している．そこで，その理論を援用して理論を構築する．

4. 研究成果

DNA 系列は四つの記号 A(アデニン)，T(チミン)，G(グアニン)，C(シトシン) から成る系列であると考えることができる．A と T，G と C は対になって，二重鎖を構成する．このため，DNA 符号を考えるときには，これらの対についても考察しておくことが必要である．

θ をこれらの対の関係を表現する関数とする．つまり，$\theta(A) = T$，$\theta(T) = A$，$\theta(G) = C$，$\theta(C) = G$ とする．

$\Sigma = \{A, T, G, C\}$ と置く．Σ^+ を，Σ の要素からなる，長さが 1 以上の系列の全体とする．L を，有限系列の集合とする．L が以下の条件を満すときに「 prefix-free 制約を満たす」という．

$$\phi(L)\Sigma^+ \cap L = \emptyset.$$

以下の条件を満すときに「 suffix-free 制約を満す」という．

$$\Sigma^+\theta(L) \cap L = \emptyset.$$

以下の条件を満すときに「 comma-free 制約を満す」という．

$$\Sigma^+\theta(L)\Sigma^+ \cap L^2 = \emptyset.$$

以下の条件を満すときに，言語 L は「 outfix-free 制約を満す」という: $u \in L$ に対して，$\theta(u_1)x\theta(u_2) \in L$ であり $\theta(u) = \theta(u_1)\theta(u_2)$ であれば，x は空系列でなければならない．この制約は，DNA 二重鎖を構成するときに，片方の一重鎖の一部が「 浮いて」しまうことがないことを保証している．

$m \geq 1$ に対して，

$$L^{m+1} \cap \Sigma^+\theta(L^m)\Sigma^+ = \emptyset$$

であるとき，L は「 intercode 制約を満す」と言う．

S を入力制約を満す系列全体の集合とする．$\mathcal{B}_n(S)$ を，S の長さが n の部分系列の集合とする．以下で定義される量を容量と呼ぶ．

$$h(S) = C(S) = \lim_{n\to\infty} \frac{\log|\mathcal{B}_n(S)}{n}$$

これは，通信路の誤りなし通信路容量に対応している．

comma-free 制約に対しては，その制約を満す言語と，容量が 1 に漸近する系列の構成方法はわかっていた．このことは，この制約を課したとしても，本質的には DNA 系列の豊富さは減少しない，ということを意味している．

(1) prefix-free 制約と suffix-free 制約について

prefix-free 制約と suffix-free 制約を満す言語 (有限系列の集り) についても，ほぼ同様のことがわかっていた．ただし，その容量 h は，DNA 系列の場合には

$$\log 3 < h < \log 4$$

ということまでしかわかっていなかった．

本研究では，suffix-free 制約と prefix-free 制約を満す言語を具体的に構成し，情報理論的および記号力学系的な手法を用いて符号の長さを長くすれば，容量が 1 に漸近することを示した．

より具体的には，この言語は以下のように構成される．$\mathcal{F}_n$ を以下のように定義する．

$$\mathcal{F}_n = \{AC^n\}.$$

そして，$\mathcal{F}_n$ の要素が出てくるような系列を排除することによって，記号力学系 S_n を定義する．この力学

系は SFT(Shift of Finite Type) と呼ばれるクラスの記号力学系になるので，有限有向グラフによって表現することができる．そのグラフを G_n とする．

G_n に対して，AC^nAC^n を生成するパスを付加する．これによって出来るグラフを G'_n とする．さらに，G'_n によって生成される記号力学系を T'_n とする．すると以下のことがわかる．

T'_n によって生成される言語は prefix-free であり，かつ，suffix-free である．そして，その容量は n が大きくなると，1 に漸近する．

(2) outfix-free 制約について

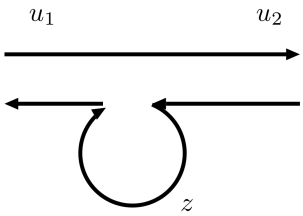
prefix-free 制約, outfix-free 制約，outfix-free 制約すべてを満す言語が実際に存在することはわかっていた．

本研究では以下の集合の要素が有限であることを示した．

$$S_L = \{\theta(u_1)\theta(u_2) : u \in L \text{ and } \theta(u_1)\theta(u_2) \in L \text{ but } \theta(u_1)x\phi(u_2) \notin L \text{ for } x \neq \epsilon\}.$$

outfix-free 制約は，もともと，二重鎖を構成するときに，二重鎖が構成されない部分が残ることがないように課された制約である．

例えば，以下の図のような (不完全な) 二重鎖ができてしまうと，計算がうまく進まなくなってしまう．



このような可能性を，系列を設計する段階で排除しようとするのが，outfix 制約である．

今，outfix 制約を満すある言語 L を考えるとする．その言語は，$\theta(u_1)\theta(u_2) \in L$ であるけれども，$\theta(u_1)x\phi(u_2) \notin L$ となるような，空でない x を持つかも知れない．

ところで，

このような条件を満す x は有限個しかない

ことを証明できる．このことは，outix 制約が非常に制限的な制約であることを物語っている．

(3) intercode 制約について

intercode 制約は，ある符号語系列が，長さが異なる別の情報の符号化系列と同じになることがないことを保証している．これは，符号理論で言うところの，一意復号可能性を保証するための制約である．

この制約については，以下の式を満すような，言語が存在することが知られていた．

$$\log\left(4^{\frac{n-1}{n}}\right) < h(X) < \log((4^{n-1}+1)^{\frac{1}{n}})$$

これに対して本研究では

$\lim_{n\to\infty} h(X_n) \to 1$

を満すような制約があることを，情報理論および記号力学系の理論を使って示した．つまり，本質的にはこの制約は制限を課していない，ということがわかる．

(4) GC 成分の制約

ある系列 u の GC 成分は以下のように定義される．

$$GC(u) = \frac{G(u) + C(u)}{n}$$

ここで，$G(u)$ と $C(u)$ はそれぞれ，u の中の G の数と C の数である．

この値は，符号理論で用いられる $RDS(u)$ と同じような役割りを果す．$RDS(u)$ は，系列 u の各記号を数値として足し合わせたものである．

信号の周波数成分が記号周波数 (記号を送出する周波数) を持たないための条件は，ある定数が存在して，任意の部分系列の RDS の絶対値が，その値以下であることが知られている．

また，符号が有限状態機械で生成されるのであれば，その機械 (有限有向グラフ) が満すべき条件も知られている．

これまでにも，GC 成分を制限する制約については，様々な特徴付けが示されていたが，情報理論，符号理論の成果に基いたものはなかった．

本研究では，coboundary 条件と呼ばれる条件とよく似た条件が，GC 成分の制限に強く関係していることを示した．また，この成果は，符号を構成する方法も与えている．つまり，その条件を満すようにグラフを構成し，記号力学系の理論に基いて符号を構成することが原理的に可能である．

5. 主な発表論文等

[雑誌論文]（ 計2 件）

1. H. Kamabe, Outfix-free and intercode constraints for DNA squences, Proceeding of 2011 IEEE International Symposium on Information Theory, 2011, pp.1574–1578, Saint Petersberg, 2011.(査読あり)

2. H. Kamabe, Constraints for DNA sequences by formal languages and its capacity, Proceeding of Nature and Biologically Inspired Computing 2010, pp. 622–627, 2010, Kitakyusyu.(査読あり)

[学会発表](計4 件)

1. H. Kamabe, Outfix-free and intercode constraints for DNA squences, 2011 IEEE International Symposium on Information Theory, Saint Petersberg, July 31–August 5, 2011.

2. H. Kamabe, Constraints for DNA sequences by formal languages and its capacity, Nature and Biologically Inspired Computing 2010, Kitakyushu, December 15–17, 2010.

3. H. Kamabe, Constraints for DNA sequences by formal languages and its capacity, 情報理論とその応用シンポジウム 2010，11 月 30– 12 月 3 日，信州.

4. 鎌部 浩，DNA 系列の制約について，電子情報通信学会情報理論研究会技術研究報告，2010 年 9 月 22 日，東北学院大学.

6．研究組織

(1) 研究代表者

鎌部 浩 (HIROSHI KAMABE)

岐阜大学・工学部・教授

研究者番号 80169614