3Z-02

マルコフ連鎖に基づく1本鎖DNA中の語の出現数の対称性の解析*

楠　瑞哉 †　　水田　智史 ‡
弘前大学理工学研究科　　弘前大学理工学研究科

1　イントロダクション

シャルガフの第1パリティ則 [1] によれば, 2本鎖DNA中の相補的な塩基 ($A \Leftrightarrow T$, $G \Leftrightarrow C$) の出現数は一致する. これはDNAの2本鎖が相補的な塩基によって結合していることから明らかである. また多くの生物種の1本鎖DNAにおいても, 相補的な塩基の出現数がおよそ一致する ($A \approx T$, $G \approx$C) ことがわかっており, これはシャルガフの第2パリティ則 [2] と呼ばれている. この法則の明白な成立要因はわかっていない. さらに第2パリティ則を拡張して, 1本鎖中における, 複数の塩基単位である語と, それを相補的な塩基に置き換えて逆順にした相補的な語の出現数がほぼ等しいという報告があり [3], この関係性を本研究では**語の数の対称性**と呼ぶ. 第2パリティ則および語の数の対称性の成立要因は, ゲノム進化に対して何らかの制限を与えていると考えられ, それを調べることで, ゲノム進化の新たな手がかりを得ることが期待できる.

本研究では, 語の数の対称性から統計的に成立する対称性の影響を取り除いた場合について詳細な解析を行うことで, 第2パリティ則および語の数の対称性の成立要因を探る.

2　方法

ゲノム配列中において統計的に生じる対称性の影響を取り除くために, 配列中の語の出現数とその期待値とで差をとり, これについて, 語の数の対称性の有無および残された対称性の要因について調べていく.

2.1　語の出現数とその期待値

DNA配列中の語の出現数の測定は, 配列の先頭から末尾に向けて, 読み取りウィンドウを1塩基ずつずらし, 出現した語をカウントすることで行う.

語の出現数の期待値の算出は以下のように行う. ゲノム配列モデルとして, 配列中の塩基の連続をマルコフ連鎖とみなしたモデル [4] を使用する. このモデルを用いると, 語 w 中の i 番目の塩基 $X_i(= A, T, G, C)$ と, 長さ M の語 $w^M = X_1 \ldots X_M$ に対して, その次に現れる塩基 X_{M+1} の条件付き出現確率は,

$$P(X_{M+1}|w^M) = F(w^{M+1})/F(w^M) \tag{1}$$

と表せる. ここで $P(A|B)$ はBが生じたときにAが生じる条件付き確率, $F(w)$ はDNA配列中の語 w の出現数である.

またマルコフモデル中での語 $w^N = X_1 \ldots X_N$ の条件付き出現確率は, w^N 中の長さ $M(< N)$ の部分塩基列 $w_i^{M+i-1} = X_i \ldots X_{M+i-1}$ と, 式 (1) を用いて

$$\begin{aligned} P(w^N|w_1^M) =& P(X_{M+1}|w_1^M)P(X_{M+2}|w_2^{M+1}) \\ & \ldots P(X_N|w_{N-M}^{N-1}) \end{aligned} \tag{2}$$

と表される. さらに語 w^N の出現数の期待値は $E(w^N) = F(w_1^M)P(w^N|w_1^M)$ とかける. この式よりDNA配列中の語 w^M の影響による, 語 w^N の出現数の期待値を求めることができる.

多くのゲノム配列は, その局所的な領域ごとに塩基の組成や出現確率が異なるため, 配列を50kbの断片に分割し, 各領域で語の出現数の期待値を計算する. それらの合計を, ゲノム全体の語の出現数の期待値とする.

2.2　対称性の評価

以上のようにして得られた語の出現数の期待値 $E(w)$ と実測値 $F(w)$ との差 $d_w = F(w) - E(w)$ に対して, 相関係数を用いて語の数の対称性を評価した. 本研究では, 相関係数 r を

$$r = \frac{2}{4^N} \frac{\sum_{w,\hat{w} \in WP} (d_w - \bar{d})(d_{\hat{w}} - \bar{d})}{Var} \tag{3}$$

によって算出した. ここで 4^N は語長 N に対する全ての語の種類, $\sum_{w,\hat{w} \in WP}$ は語 w と相補的な語 $\hat{w}$ の各ペア

---

* Markov Chain Based Analysis of Symmetric Word Frequencies in Single-Stranded DNA

† Mizuya Kusunoki, Graduate School of Science and Technology, Hirosaki University

‡ Satoshi Mizuta, Graduate School of Science and Technology, Hirosaki University

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.

についての和, $\bar{d}$ は差 d_w の語全体での平均, Var は分散である.

2.3 使用した生物種

バクテリアから 53 種, アーキアから 52 種, 真核生物から 25 種の合計 130 種の生物種を, ゲノムの長さや GC 含量といった特徴が散らばるように選んだ. 染色体を複数持つ生物種については, 各染色体での測定結果を合計したものをゲノム全体の結果として扱う.

3 結果

断片の長さ 50kb, 語長 $N = 9$, 期待値の算出に用いた語長 $M = 4$ で, 生物種ごとに測定を行った. 得られた相関係数 r の度数分布を図 1 に示す. 期待値との差をとる前の場合では, 多くの生物種が高い相関係数を示していることが確認できる. また差をとった場合では, 対称性の要因を取り除いたことによって, 全体的な相関係数が低下しているが, 多くの生物種が相関を保っていることが確認できる. このことから, 語の数の対称性は統計的に成立しているものだけではなく, 他の要因が影響していると考えられる.

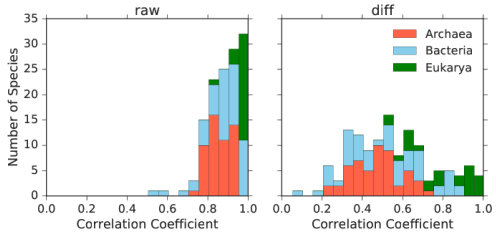

図 1 左：差をとる前の語の数に対する相関係数の度数分布. 右：差に対する相関係数の度数分布.

次に, 語の数の差についての相関係数 r とゲノム配列長とをプロットしたものを図 2 に示す. ここでは視認性のために 50Mb のゲノムサイズまで表示していることに注意する. 図より, 全ての生物種のドメインで配列長が長くなるほど相関係数も大きくなる傾向が確認できる.

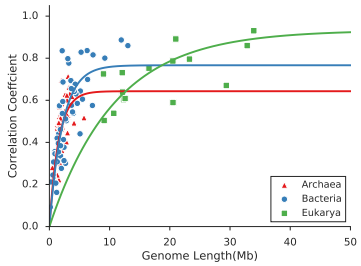

図 2 相関係数とゲノム配列の長さの関係

また, 配列長が相関係数を高める要因であるかを調べるために, ゲノム配列の各断片, および断片をつなげて配列を伸ばした場合の相関係数を調べた (図 3). ここでは, バクテリアの代表的なモデル生物である *Escherichia Coli* K12 を対象に行った. 各断片では, 差をとることで相関係数が 0 に近い値をとっている (Diff) 一方, それら断片の語の数の差 d_w を加算していった場合には, 配列長が長くなるほど相関係数が大きくなる傾向が見られる (Diff-Sum). このことから, 配列長と対称性を高める要因との間には何らかの関係があると考えられる.

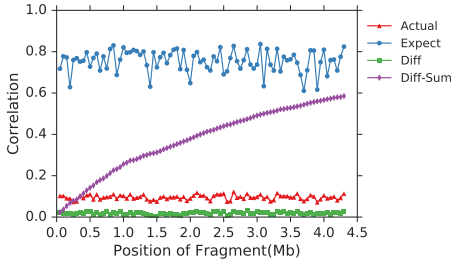

図 3 断片の位置ごとの相関係数

4 まとめ

DNA 配列中の語の出現数とマルコフモデルを用いた語の出現数の期待値との差について, 生物種ごとに語の数の対称性を調べた. その結果, 多くの生物種が差をとっても相関が残ることが確認できた. また, 配列長が長いほど語の数の対称性も高くなる傾向が見られた.

参考文献

[1] Chargaff, E. (1950). Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, **6**(6), 201-209.

[2] Rudner, R., Karkas, J. D., & Chargaff, E. (1968). Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. *Proc. Natl. Acad. Sci. USA*, **60**(3), 921-2.

[3] Prabhu, V. V. (1993). Symmetry observations in long nucleotide sequences. *Nucleic Acids Research*, **21**(12), 2797-2800.

[4] Baisnee, PF., Hampson, S., and Baldi, P. (2002). Why are complemntary DNA strands symmetric? *Bioinformatics*, **18**(8), 1021-33.

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.