P16

DNA 上の翻訳開始点付近の GC 含量の分布と生物種との相関性
The Distribution Profiles of GC Content surrounding the Translation Initiation Site in Different Species

京都大学化学研究所　水野　政彦　金久　實
Masahiko MIZUNO, Minoru KANEHISA
Institute for Chemical Research, Kyoto University
E-mail: mizuno@kuicr.kyoto-u.ac.jp

ABSTRACT

We analyze the distribution of GC content around the translation initiation site in genomic DNA sequences of different species using a DNA data bank, GenBank (release 72, 73). We select the data sets of entries with no homologous sequences whose homology is more than certain thresholds in each species. A window length of 100 bases is shifted each time by 50 bases along the nucleotide sequences of these entries from the -500 base position to the +500 base position. (The first nucleotide at the initiator codon is designated +1, with positive and negative integers proceeding 3' and 5', respectively.) The average of GC content in each region surrounding the translation initiation site is calculated from 10 and more windows located at the same region in the same species. The comparison of the GC content patterns of different species suggests that the distribution profiles of GC content around the translation initiation site depend on the kind of organism groups to which the species belong and that these GC content profiles are characteristic of organism groups.

概要

核酸配列データベースである GenBank (release 72、73) を用い、生物種ごとに DNA 上の翻訳開始点付近の GC 含量の分布を調べた。同一生物種内で配列の類似した重複データを除くために、その遺伝子がコードするアミノ酸配列が一定の homology 以下のものだけを選び、データとした。翻訳開始点を座標 +1 とし、-500 base から +500 base までの領域内で、大きさ 100 base の ウィンドウを 50 base ずつずらし、各ウィンドウの GC 含量の平均を同一生物種の複数の遺伝子より求めた。その結果、同一生物群に属する複数の生物種の比較から、対応する各ウィンドウの GC 含量の絶対差が大きいにもかかわらず、塩基配列に沿った GC 含量の変化パターンは同一生物群内で類似しており、さらにそれらは各生物群に特徴的な傾向であることがわかった。

I. はじめに

ヒト染色体上では、バンド構造に対応して mega base オーダーの GC または AT に富む領域が存在することが示唆されている[1, 2]。また、温血脊椎動物では、isochore と呼ばれる 300 kilo base 以上におよぶ GC 含量の異なる領域がゲノム上に存在することが報告されている[3]。このような GC 含量の異なるモザイク構造は大腸菌では存在しないとされている[4]。さらに、GC に富む isochore は単子葉植物にはあるが、双子葉植物には見られないことが指摘されている[5]。一方、タンパク質をコードする領域は非コード領域より GC に富む傾向があることも知られている。

このように生物種により、マクロなレベル、ミクロなレベルにおいて、DNAの塩基組成には偏りが存在する。本研究では、染色体のバンド構造やisochoreより小さい、100 base長単位に、いくつかの生物種のタンパク質構造遺伝子の翻訳開始点のまわりのGC含量の分布を調べた。

II. データの選択

核酸配列データベースであるGenBank (release 72、73) より、以下の条件を満たすデータを用いた。

1. ソースによる選択

- genomic DNAから決定された配列である (mRNAを除く)。
- mitochondrion、chloroplast、kinetoplast、plasmid由来でない。
- 原則として、完全なタンパク質をコードしている。
- エントリー名の初めの3文字より、同一の生物種である。

2. アミノ酸配列レベルでの選択

1. ソースによる選択によって選ばれたエントリー中には同一または類似配列が含まれているので、これを除くために以下に示すように核酸配列の翻訳領域をアミノ酸配列に翻訳し、アミノ酸配列が類似するものは除いた。

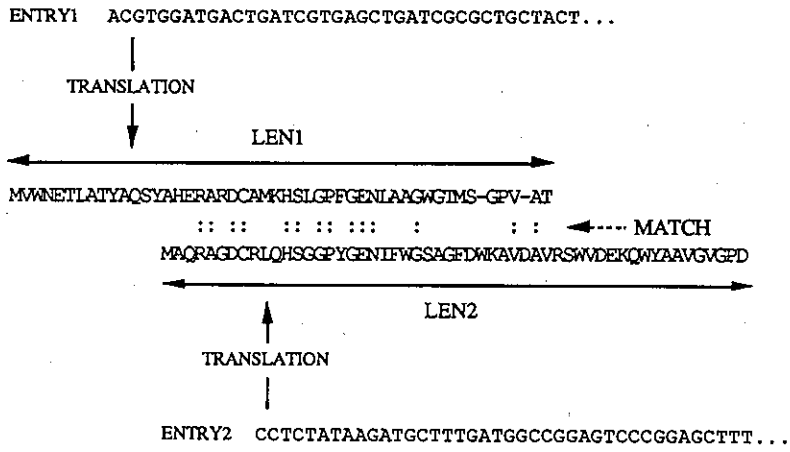

Fig.1 Data selection by amino acid sequence homology. Entry1 and entry2 are genes derived from the samespecies. LEN1 and LEN2 are the length of amino acid sequences. Two amino acid sequences are aligned using Dayhoff score matrix. If inequality (a) is true, a entry which has more windows in the nucleotide sequence is selected.

まず、翻訳されたアミノ酸配列の残基長が20残基未満のエントリーを除き、Fig.1のように、Dayhoff matrixを用いたアライメントをする。

$$\frac{\text{MATCH}}{\min(\text{LEN1}, \text{LEN2})} \times 100 \geq \text{THRESHOLD} \qquad \text{(a)}$$

THRESHOLD: 20, 25, 50, 75(%)

の左辺で表わされるホモロジーが、右辺のしきい値以上の場合には、2つのエントリーを類似配列と見なし、これらのアミノ酸をコードするもとの核酸配列上にとれる長さ100 baseのウィンドウの数が多い方のエントリーを選ぶようにした。

III. GC含量の測定法

測定の範囲は翻訳開始点を座標+1として、-500から+500 baseまでとした。ただし、3'端については、翻訳領域が+500 base未満で終わっている場合には翻訳領域の3'端までとした。長さ100 baseのウィンドウを核酸配列に沿って50 baseずつずらし、各ウィンドウ中のGC含量をパーセントで算出する。ただし、5'端、3'端を含むウィンドウで100 baseに満たない核酸配列を含むウィンドウは除外した。翻訳開始点付近の各領域におけるGC含量を、同一生物種の複数の遺伝子のウィンドウのGC含量の平均値として求めた。

IV. 結果

結果をFig.2に示す。1つの生物種について、5通りのしきい値をとった場合のGC含量をグラフで示した。いずれのしきい値についても、各領域におけるGC含量の値はほとんど変化していない。そして、同一生物群内ではGC含量の絶対差が大きいにもかかわらず、その変化パターンは類似している。

Fig.2に示した3つの生物群、哺乳動物・植物・原核生物を比較すると、生物群によって翻訳開始点の直前と直後とのGC含量の差の大きさが異なっている。また、-100から+150 baseの領域において、各生物群に特徴的なGC含量の変化パターンが見られる。この領域において、哺乳動物ではGC含量の変化がほとんど見られないのに対し、植物では大きく変化している。さらに原核生物では変化は見られるが、GC含量のピークはいずれも植物より下流へ50 baseずれている。これらのことは、DNA上の翻訳開始点付近のGC含量の分布と生物種・生物群との間に相関性があることを示唆している。

以上の結果に加えて、他の生物種、及びmRNAの場合についても報告したい。

本研究は文部省科学研究費重点領域研究「ゲノム情報」の助成を受けています。また京都大学化学研究所スーパーコンピュータラボラトリより計算機のCPU timeの提供を受けました。

参考文献

[1] S. Aota, T. Ikemura: *Nucleic Acids Res.*, 14, 6345 & 8702 (1986).
[2] T. Ikemura, K. Wada: *Nucleic Acids Res.*, 19, 4333 (1991).
[3] G. Bernardi, G. Bernardi: *J. Mol. Evol.*, 24, 1 (1986).
[4] 池村淑道: RNAの世界 (大澤省三、志村令郎編)、205 (1990).
[5] J. Salinas, G. Matassi, LM. Montero, G. Bernardi: *Nucleic Acids Res.*, 16, 4269 (1988).

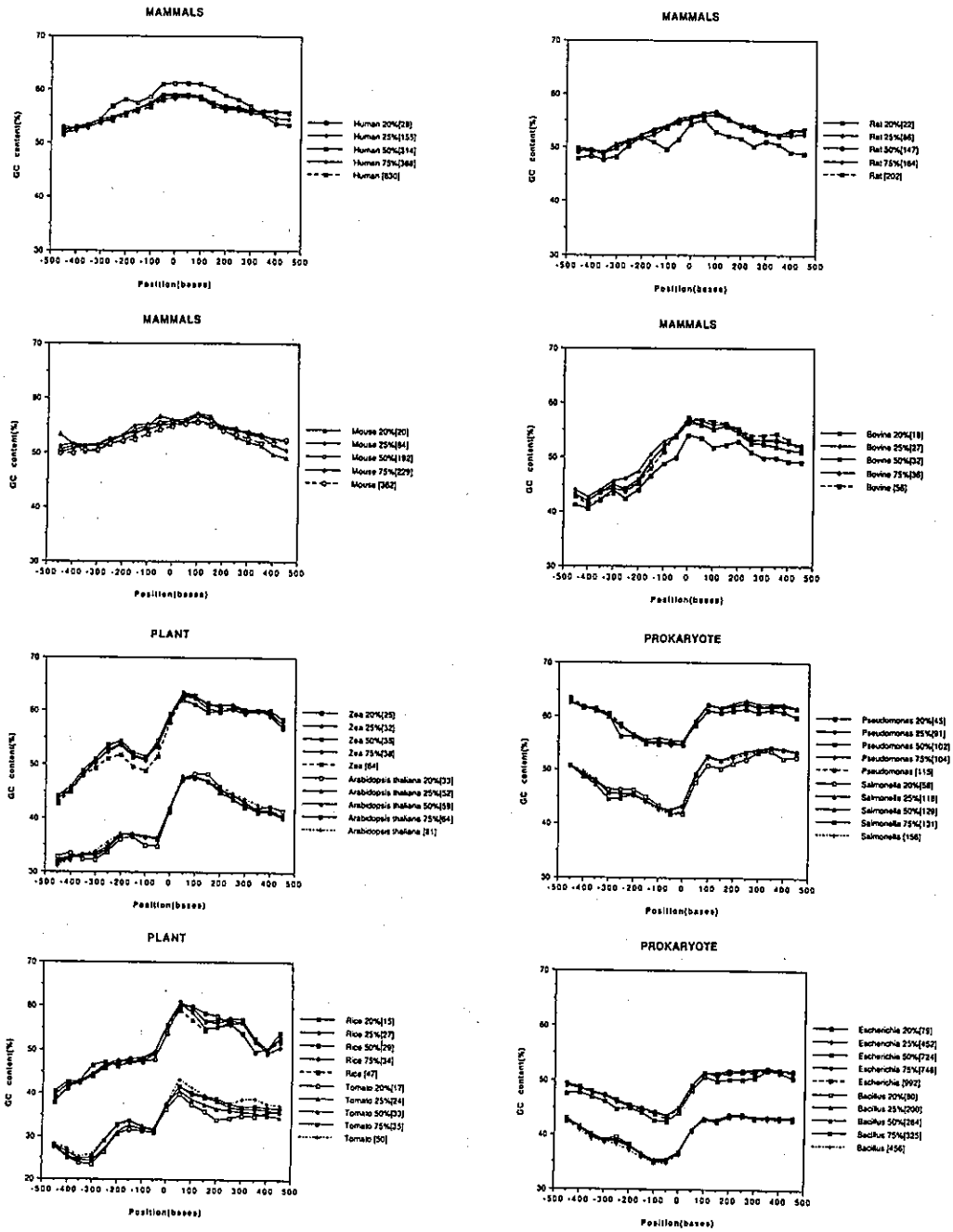

Fig.2 The distribution profiles of GC content around the translation initiation site in different species. Each species is classified into one of three organism groups: mammals, plant, and prokaryote. A window length of 100 bases is shifted each time by 50 bases along nucleotide sequences from the -500 base position to the +500 base position. (The first nucleotide at the initiator codon is designated +1, with positive and negative integers proceeding 3' and 5', respectively.) The average of GC contents in each region surrounding the translation initiation site is calculated from 10 and more windows located at the same region in the same species. In the legends, percentages mean thresholds of amino acid sequence homology shown in inequality (a). The value in bracket is the number of entries (genes).