核酸塩基配列の基礎的解析

正 員 宮 腰 隆 (富 山 大 学)

非会員 野 田 徹 ((株)富士通関西システムズ)

正 員 畠 山 豊 正 (富 山 大 学)

Basic Analysis for Sequence of Nucleic Acid

Miyagoshi Takashi, Member (Toyama University), Noda Toru, Non-member (Ltd. Fujitsu Kansai Systems), Hatakeyama Toyomasa, Member (Toyama University)

We have analyzed sequence of human T-cell receptor. A histogram showing the numbers of the neighboring base pairs was made. It indicates the following characteristics: (1) the numbers of the same base pair are larger than these of the theoretical values. (2) the numbers of C → G pair are much less than their theoretical values. (3) The sequence holds properties based on the complementarity of them.

Moreover, we have obtained the theoretical values of the interval in that a given base does not appear, and compared with the actually measured value. We would like to clarify statistical characteristics of the sequence by the information theory approach.

キーワード：核酸塩基配列，塩基数解析，期待値，隣接塩基対数ヒストグラム

1. はじめに

ヒトゲノムの解読が進み，今後は，遺伝子の機能を解明し革新的な医療技術や遺伝子産業への発展へと期待される中で，ホモロジー解析，相同性検索，マルチプル・アライメントなどの計算機による支援ソフトも開発されてきている⁽¹⁾。一方，GenBankでは1999年までに約535万LOCUS，総計46億bpの塩基配列が報告されており，膨大なデータ解析結果の表現法も重要な課題となっている⁽²⁾。また，核酸の物理的性質を調べるための基礎となる解析は，GC含量など各塩基数の割合以外はあまり調べられていないようである。

我々は，ヒトのT-cell receptorの塩基配列⁽³⁾を対象にして，(1)隣接塩基対数解析(どの塩基の隣にどの塩基が現れ易いのか)の特性が一見して認識できるヒストグラムを作成した。また，(2)同一塩基の不出現間隔解析(ある塩基が連続して現れない塩基の数はいくつか)の理論値を求め実測値と比較検討した。

2. DNAの成分と構造⁽⁴⁾

DNA(デオキシリボ核酸)はゲノムの分子の実体であり，ヌクレオチドと呼ばれる分子が結合してできた高分子である。ヌクレオチドはそれぞれ1個の糖，1個のリン酸，1個の塩基から成っている。ヌクレオチドには4種のものがあるが，塩基の部分だけが異なっており，それぞれアデニン(A)，チミン(T)，グアニン(G)，シトシン(C)であり，AとGはプリン(pu)，TとCはピリミジン(py)と呼ばれている。従って，DNAは糖とリン酸でできた主鎖に沿って特

定の配列で塩基が並んだものであり，このときの塩基配列が特定のタンパク質，ひいては生物を作る設計図となっている。

3. 塩基数解析

〈3・1〉隣接塩基対数解析 ある核酸の塩基配列すべてに対して，5'から3'へ向かって，当該塩基の隣が何塩基であるかを調べ，菱形の4隅に隣接塩基が同じ場合，すなわち，A → A，C → C，T → T，G → Gをこの順に右回りにとり，それらの出現数を棒グラフで示す。また，各辺と対角線を三等分して，例えば，上辺ではA → AよりにはA → CをC → CよりにはC → Aの出現数を棒グラフで示す。このことによって，左端の列から右端の列に向かってpu-pu，pu-py，py-pu，py-pyの順に並ぶので，1つのヒストグラムで当該核酸の隣接塩基対の特性がわかる。

いま， B_1 と B_2 を4つの塩基のうちの1つとすると， $B_1 \rightarrow B_2$ の発生する割合の期待値は，

$$\text{期待値 (割合)} = \frac{B_1 \text{の総数}}{\text{総塩基数}} \times \frac{B_2 \text{の総数}}{\text{総塩基数}} \dots\dots\dots (1)$$

となる。ここで B_1 と B_2 が同じ塩基でもよい。

図1は，ヒトのT-cell receptorの解析結果をグラフ表示したものである。バーは式(1)から求めた期待値を示す。縦棒は期待値まで黒で，それを越えた部分は白で示してある。また，グラフの右には当該塩基を構成している4つの塩基の割合を示した。このヒストグラムでは，対角線上にある4種の隣接塩基対はそれぞれ独立である。一方，頂点

図1 Human T-cell receptor の隣接塩基対数ヒストグラム。
Fig.1. A histogram showed the numbers of the neighboring base pairs of Human T-cell receptor.

にある同塩基対はそれぞれ対角にあるものですが、また辺上にあるものはそれぞれ相対する対辺上のものと縮退している。ここで、縮退とは DNA の順方向 (5' → 3' 方向) と逆方向 (3' → 5' 方向) が相補的な配列をもち、両方の配列が同じ統計的性質をもつならば、 $B_1 \rightarrow B_2$ と $B_2 \rightarrow B_1$ が同じ確立で出現することを意味する。一見してわかるように、これら縮退しているものどうしの塩基対の出現頻度が非常に似ている。

〈3・2〉 同一塩基の不出現間隔解析 塩基配列の 5' 側から 3' 側に当該塩基が連続して何塩基分現れないか、またその箇所が何箇所あるかを調べた。この期待値は、

$$\begin{aligned} \text{期待値} &= \frac{S}{W} \times \frac{S-1}{W-N-1} \\ &\times \prod_{i=0}^{N-1} \frac{(W-S)-i}{W-(i+1)} \dots\dots\dots (2) \end{aligned}$$

である。ここで、 N は当該塩基の不出現間隔の塩基数、 S は当該塩基数および W は総塩基数である。図 2 は A と T について、○ は実測値を実線は式 (2) から求めた理論値を図示したもので、当該塩基の現れない間隔が小さい場合には、理論値とほぼ一致していることがわかる。図で右端の○が不出現間隔の 1 番大きかったところである。すなわち、A、T、G、C の順に、46、64、75、79 間隔であり、それぞれの理論的出現確率は、式 (2) より 1.40×10^{-8} 、 1.27×10^{-10} 、 1.97×10^{-10} 、 2.36×10^{-10} であり、不出現間隔は理論的には存在の可能性の非常に小さいところまで生じている。ここで使用した塩基配列で最大であったのは、Haemophilus influenzae Rd transferrin binding protein l gene(3039bp) であり、156 塩基間で T が全く出現しなかった。ただ、その生物学的意味は不明である。

4. 考 察

ヒトの T-cell receptor の塩基配列解析において、作成したヒストグラムでは、

- (1) 同塩基の隣接塩基対の数が期待値より大きい。
- (2) C → G 対が、期待値より非常に小さい。

図2 Human T-cell receptor に対する当該塩基不出現間隔解析。
Fig.2. The interval in that a given base does not appear(Human T-cell receptor).

(3) 塩基の相補的性質から生ずる特性にほぼ従っている (例えば、A → G と C → T はほぼ等しい確率で現れた)。

(2) の現象は特に興味深い。すなわち、GC 塩基対が 3 つの水素結合で結ばれているのに対して、AT 塩基対は 2 つの水素結合で結ばれているので、前者の結合が強い。更に、スタッキング相互作用も GC が 1 番大きい⁽⁵⁾。従って、局所的に求めたエネルギー的に安定な配列の出現確立が大きいのとは言えない。

今日のゲノムに対する研究は、塩基配列の同定と、ホモロジー解析などが主である。これはゲノムがあまりにも巨大なデータを含んでおり、それを同定することに追われているからであろう。今後は、ポストゲノムの研究として、GenBank の膨大なデータを駆使した情報論的アプローチで塩基配列を解析し、統計的性質や新たな生物学的意義を抽出することも重要な課題である。

(平成 12 年 10 月 16 日受付、同 12 年 12 月 12 日再受付)

文 献

- (1) 美宅成樹, 金久賀共編:「ヒトゲノム計画と知識情報処理」, 培風館, 1995.
- (2) J. D. Parsons:「Miropeats: graphical DNA sequence comparisons」, CABIOS, Vol. 11, No. 6, pp. 615-619, 1995.
- (3) C. J. Schlueter, B. A. Schodin, S. Y. Tetin and D. M. Kranz:「Specificity and binding properties of a single-chain T cell receptor」, J Mol Biol, Vol. 256, No.5, pp. 859-869, 1996.
- (4) B. Lewin:「Genes V」, Oxford University Press and Cell Press, U.S.A., 1994.
- (5) L. V.Yakusheich:「NONLINEAR PHYSICS OF DNA」, WILEY SERIES IN NONLINEAR SCIENCE, 1998.