DNAの長距離相関

3U−8

五味壮平 遠藤教昭 白倉孝行 進藤浩一
岩手大学人文社会科学部

DNAの塩基配列は、アミノ酸、あるいはそれを構成要素とする蛋白質をコードしている。しかし、一般に高等生物のDNAで、このアミノ酸のコードに直接関係している塩基は、全体の一部に限られている。塩基配列の残りの部分（イントロンや遺伝子間配列）の役割は明らかにされていない。

ここ数年、このアミノ酸をコードしていない塩基配列（以下、非コード領域と呼ぶ。）が、極めて興味深い性質を持つことが明らかにされてきた。この性質とは、以下のようなものである。

$1/f^{\alpha}$ノイズ、相関関数のべき的減衰[1, 2, 3]

非コード領域には、長距離的な相関が存在する。すなわち、DNAの塩基配列を適当なビット列に変換したとき、このビット列のスペクトルは$1/f^{\alpha}$ ($\alpha \sim 1$)で近似的に記述される。また、同じビット列の相関関数はべき的な減衰をする。

Zipf則、冗長度[4]

DNAでの非コード領域では、自然言語や人工言語でよく見られるZipf側が近似的に成り立つ。Zipf則とは文章中の単語についての統計的法則（経験則）である。DNAの場合には、まず"単語"をビット列中の連続したnビットとして定義する。各単語の出現頻度を測定したうえ、横軸に頻度の高い順（ランク）に単語をならべ、縦軸にそれぞれの頻度をプロットした両対数グラフは直線的となる（すなわち出現頻度がランクの逆べきの関数で記述される)、というのがZipf則である。

また、非コード領域では、"冗長度"（エントロピーから定義される）が、コード領域よりも大きくなっている。

これらの性質は、DNAの非コード領域中の各塩基が、なんらかのルールに基いて配置されていることを意味する。さらに、いずれの性質も言語（文章）と共通するものであり、非コード領域が言語的な情報を持つとする考え方も存在する。

こうした性質を生み出すための基礎的なメカニズムについて、以下のようなモデルが提案されている。

1. **伸長-修正モデル（Expansion-Modification System）**[5]

最初はランダムなビット列からスタートし、そのビット列が以下のルールに従って再起的に発展していくとする。(1): あるビットが"0"ならば、確率pで"1"に、確率$1-p$で"00" に置き換える。(2): あるビットが"1"ならば、確率pで"0"に、確率$1-p$で"11" に置き換える。(3): (1)と(2)を配列中のすべてのビットに同時に適用する。このモデルでビットの増加は"重複"、またビットの変化は塩基の"置換" （重複、置換のいずれも突然変位の一種）に対応するプロセスである。

2. **一般化されたレヴィーウォークモデル（Generalized LévyWalk Model）**[6]

Long-Range Correlations in DNA Sequences
Sohei Gomi, Noriaki Endo, Takayuki Shirakura and Koichi Shindo
Iwate University, 3-18-34, Ueda, Morioka, Iwate 020, Japan

DNAの非コード領域では、4種の塩基が一様に分布しているわけではない。むしろ、ある一つの塩基の濃度が高い部分などが継ぎ合わされて、パッチ構造をとっているように見える。ここで、各パッチ内での塩基の配置がランダムに（ただし、それぞれの濃度に比例した出現確率で）決まっていると仮定する。このとき、パッチの大きさに特徴的なスケールがなく、極めて大きいものからほんの小さいものまで存在するとしたら、それだけで上記のような性質が得られると主張するのが一般化されたレヴィーウォークモデルである。具体的には、まずべき的な確率分布によって各パッチの大きさを決定する。次に、それぞれの大きさに比例した長さのビット列を1または0、どちらかの濃度を少し高くして作り出す。この際、1と0のどちらの濃度を高くするかはパッチごとにランダムに決める。最後にこうしてできたパッチを張り合わせて一つのビット列とする。尚、このモデルでは、パッチの生成過程のメカニズムは問わない。

3. 単語に対してのマルコフ過程（Markov Process for Words）[7]

非コード領域が言語的な情報を持っているとする最も極端なモデルである。そもそも"単語"が存在することを前提としている。（単語の長さはやはりnビットと仮定する。）

配列中で、ある単語が出現する確率は、その隣の単語が何かのみによって決まっていると考える。各単語の隣に来ることのできるのが、それぞれ極めて限定された単語だけだとするとき、Zipf則等、上記の性質を持ったビット列を作れることが示せる。ここで、隣に出現し得る単語を限定することは、単語間の意味的なつながりを考慮することに相当する。

本講演の目的は、これらのモデルから得られるビット列とDNA配列との比較、あるいはモデル相互での比較を行うことである。この際、当然ある特定の配列ではなく、各々のモデルから生み出されるビット列集団の統計的性質を問題とする。このための一つの方法は、それぞれのモデルの特徴的な性質が、他のモデルのビット列にもあてはまるかどうか調べることである。例えば"スケール不変なパッチ構造"や"単語についてのマルコフ性"が存在するか、それぞれのモデルで検討する。さらに、ビット列の集団間の"距離"を定義し、集団としての類似度についても議論する予定である。

参考文献

[1] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).

[2] C.-K. Peng et al., Nature **356**, 168 (1992).

[3] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[4] R. N. Mantegna et al., Phys. Rev. Lett. **73**, 3169 (1994). R. N. Mantegna et al., Phys. Rev. E **52**, 2939 (1995).

[5] W. Li, Phys. Rev. A **43**, 5240 (1991).

[6] S. V. Buldyrev et al., Phys. Rev. E **47**, 4514 (1993).

[7] I. Kantar and D. A. Kessler, Phys. Rev. Lett **74** , 4560 (1995).