タンパク質中におけるホモポリマーの遺伝暗号の解析*

2B−6

田中 剛範†　大内 東†　松嶋 範男‡　大柳 俊夫‡

†北海道大学工学部　‡札幌医科大学保健医療学部

1．はじめに

分子生物学で取り扱う、遺伝子の塩基配列およびタンパク質のアミノ酸配列は、ある決まった数種類の文字を並べた有限個の一次元配列として表される。近年、多くの生物のタンパク質のアミノ酸配列中に、ある1種類のアミノ酸だけが繰り返す特殊なパターン（ホモポリマー）が存在することがわかり、この部分に何らかの遺伝的意味があることが予想されている。

本研究では、このホモポリマーの遺伝暗号解析について、主に情報工学的見地から述べると同時に、データ処理についての生物学的な知識の関わりとも合わせて説明する。

2．解析の方法について

表1に示すように、遺伝子上ではDNA 3つが一組の暗号(コドン)となって一つのアミノ酸を決定するが、ほとんどの場合一種類のアミノ酸に対して2つから6つの異なる遺伝暗号が存在する。ホモポリマーは単一アミノ酸の繰り返しであるが、その遺伝暗号まで単一のものの繰り返しとは限らない。それらの使用頻度等を全ての生物種について調べることで、新たな事実を発見することが目的である。なお今回は、ホモポリマーを"単一アミノ酸の10回以上の繰り返し"と定義し、20種類のアミノ酸全てについてのホモポリマーを検索した。

解析は以下の手順で行なった。

1. アミノ酸配列データベースPIRおよびSwiss-protでホモロジー(相同)検索を行ない、ホモポリマーを含むタンパク質を全て調べ挙げ、それらのデータを得る。

   各データベースには、ホモロジー検索を行なうためのプログラムが用意されているので、それを利用して目的のタンパク質を調べ、メールサーバによってデータを入手する。
2. 得られたタンパク質データには、そのタンパク質をコードする遺伝子の配列がわかっている場合、その遺伝子の登録番号が書かれているので、それを用いて核酸塩基配列データベースGenBankおよびEMBLから遺伝子データを得る。

   登録番号を用いて、タンパク質データの時と同様にメールサーバによってデータを入手する。
3. 得られた遺伝子データとタンパク質データを対

1st	2nd				3rd
	T	C	A	G	
T	Phe	Ser	Tyr	Cys	T
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	終結	終結	A
	Leu	Ser	終結	Trp	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	T
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	T
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

表1: 遺伝暗号とアミノ酸の対応表

*Genetic Code Analysis of Homopolymers in Proteins

†TAKANORI TANAKA and AZUMA OHUCHI
Faculty of Engineering, Hokkaido University
‡NORIO MATSUSHIMA and TOSHIO OHYANAGI
School of Health Sciences, Sapporo Medical University

応させ、ホモポリマー部分がどのような遺伝暗号(コドン)によって表されているかを調べ、その使用頻度を数え上げる。
生の塩基配列データには、タンパク質に翻訳されないデータ領域(イントロン等)が混じっているので、その部分を取り除き(範囲はデータに書かれている)翻訳して、対応するタンパク質データと照合する。さらにその中からホモポリマー部分を抽出し、遺伝暗号の使用頻度を数え上げる。

4. 結果をまとめ、適切な分析及び考察を行なう。

大量に集められた生物情報データについて、上記の処理を迅速かつ正確に行ない、またその結果得られた二次データを保存する事は、手作業では非常に困難である。このため、これら一連の作業を自動化するプログラムを作成した。これによって、全てのデータの管理がスムーズにおこなわれ、また得られた二次データは、別のプログラムによって容易に再利用ができるようになった。

3. 結果の分析

まず最初に個々のタンパク質における、n 種類の遺伝暗号の使用頻度分布 $X=\{c_1,\cdots,c_n\}$ の偏り具合を求める。これには X の分散あるいは標準偏差を使うのが一般的であるが、今回サンプルとなるホモポリマーの長さはまちまちであり、これらを互いに比較する際に不便である。そこで、サンプルと同じ長さのホモポリマーがとりうる標準偏差の最大値に対する、実際の X の標準偏差の比を取った。この値を V として、数式で示すと以下のようになる。

$$V=\frac{\sigma}{\sigma_{max}}=\sqrt{\frac{E(X^2)-(E(X))^2}{c^2/n-(E(X))^2}}$$

$$\left(c=\sum_{k=1}^{n}c_k,\quad E(X)=c/n\right)$$

これによって V は、使用頻度の偏りが大きいほど 1 に近づき、小さいほど 0 に近い値を取る。また n はアミノ酸の種類によって 1 ～ 6 の値を取るが、特に $n=2$ のとき、上記の式は、

$$v=\frac{|c_1-c_2|}{c}$$

という単純な形で表すことが出来る。

この指標 V を、$n\geq 2$ の全てのホモポリマーについて求めたところ、半数以上のサンプルについて、$V\geq 0.5$ という結果を得た。さらにこの分布の中身を詳細に調べると、同じアミノ酸のホモポリマーが一つの生物種の異なるタンパク質に同時に存在する場合は、最頻値を示した遺伝暗号がほぼ共通していた。このことから、原始の生物の遺伝子は、局所的にみればたった 1 種類の遺伝暗号の連なりであったという推測をすることができる。

一方、$n\geq 4$ のホモポリマーの場合について、遺伝暗号をさらに 2 つ又は 3 つのグループに分け、そのグループごとの使用頻度分布について同様な分析を試みた。しかしこの場合には、グループ分けのパターンと使用頻度(の和)の偏りの相関について、何らかの生物的意味を見いだすことができなかった。したがって、このような観点から遺伝子を見ることはあまり意味を持たないいうことがわかった。

この過程で重要なことは、情報科学と分子生物学のそれぞれの理論に基づいて分析が行なわれるべきであって、一方が主で他方が従という関係になってはならないという事である。互いに補い合うことで、それぞれが単独ではなしえなかった成果をあげる事が可能になるだろう。

4. おわりに

最も原始的なタンパク質の形態であるホモポリマーおよびそれに関する遺伝子を研究することは、タンパク質や遺伝子の根本にある原理を追及することにほかならない。本稿では、研究の生物学的な意味には詳しく触れなかったが、解析の結果、DNA の構造や遺伝子レベルでの進化に関して幾つかの予想を得ることが出来た。今後はさらに DNA の構造に関する法則を探すとともに、タンパク質の立体構造予測についても研究を行なう予定である。

参考文献

[1] David Freifelder : 分子生物学の基礎. 東京化学同人, 1989.