

DNAデータベースにおけるtRNA遺伝子に関するアノテーションの問題点と改善策

国沢 隆

(kunisawa@rs.noda.tus.ac.jp)

Abstract

Transfer RNA gene sequences of plastid genomes are extracted from the GenBank/EMBL/DDBJ databases (ISDN) and are aligned to identify tRNA elements. During the course of this effort, mis-annotations of the coordinate and tRNA-assignment (Ile-tRNA, initiator Met-tRNA, or elongator Met-tRNA) are found at high frequencies. Further supports for high quality database-annotation are necessary to avoid these mis-annotations.

1. はじめに

DDBJ, GenBank, EMBLの3極で構成される国際塩基配列データベース (INSD)は世界中の科学者に公開され、公開データは実験計画の立案や理論的解析のために広く利用されている。著者もこのデータベースを用いて、ゲノム上の遺伝子の配列順序からゲノムの系統関係を探る研究をしている。この過程で、INSDデータベースの遺伝子、特にtRNA遺伝子に関するアノテーション (以下注釈という) には、利用する上でいくつかの問題があることが分かった。このことは、INSDに限らず他の多くの学術データベースにも共通して存在し、放置できない問題と考えると、この場を借りて報告する。

著者は現在、色素体ゲノムの進化を研究対象にしているので、全配列が決定された色素体ゲノム塩基配列を例にして、話を進める。また、蛋白質をコードする遺伝子の注釈についてはここでは触れず、tRNA遺伝子に関するアノテーションについてのみ言及する。まず、注釈の訂すべき点や改善すべき点を列挙し、なぜこのようなことが起こったのかを分析し、その上で改善に向けた方向づけをする。

2. tRNA遺伝子に関する注釈 (アノテーション) の品質

INSDでは塩基配列そのものが核となるデータであり、その上に何番目の塩基から何番目までの塩基が、例えばロイシンのtRNAをコードしているというような注釈が加えられている。この2次的データである注釈があることによって、塩基配列データの意味が理解でき、また他の配列とも比較でき、直ちに利用可能となりその利用価値が何倍にも高まっている。ところが、色素体ゲノム塩基配列にコードされたtRNA遺伝子の場合には、以下に述べるように、何番目の塩基から何番目までの塩基までという数の指定に誤りがあることがある。しかも、かなりの高頻度で。この数の誤りを見ると、10塩基だけずれていることが多いことに気づく。おそらく、データベースへの登録者が塩基数を数えるときの間違いがそのまま残ったものと想像される。図1には、これらの単純な間違いを取り除いたときの、Pro-tRNAの1次・2次構造を比較した。各ゲノムのPro-tRNAの配列を1行毎に表示した。*Mesostigma viride* (Mviと略されている) という色素体では71518-71591であるべきものが、10塩基余分に含む71508-71591という注釈が付いている。こうした誤りを除くと、このtRNAは長さやその配列が良く保存していることが判明する。

図1 単純ミス - Pro[TGG]の例

	Accep	Darm	Darm Anti	Anti	Tstem	Tstem Accep
	*****	++++	++++	-----	=====	=====
Cpa(+ 22308)	CGGGATGTAGCGCAGTTTGGTAGCGCGCTGCTTTGGGAGCAGGATGTCGCAGGTTCAAATCCTGTCATCCCGA					
Cca(-119549)	CGGGATATAGCGTAGCTTGGTAGCGCGCTGCTTTGGGAGCAGGATGTCGCAGGTTCAAATCCTGCTGTCCCGA					
Ppu(- 4889)	CGGGATATAGCGCAGCTTGGTAGCGCGCTGCTTTGGGAGCAGGATGTCGCAGGTTCAAATCCTGCTATCCCGA					
Osi(+ 267)	CGGGATATAGCTCAGTTTGGTAGAGTACCTGCTTTGGGAGCAGGGTGTCTAGGTTCAAATCCTACTATCCCGA					


```
Sol  orf22--ycf3 --ycf3 --psaA --psaB --rps14--fM --Ggcc *-psbZ *-Stga --psbC *-
Osa  ycf3 --ycf3 --ycf3 --psaA --psaB --rps14--fM --Rtct --atpA *-atpF *-atpF *-
Tae  ycf3 --ycf3 --ycf3 --psaA --psaB --rps14--fM --Rtct --atpA *-atpF *-atpF *-
Zma  ycf3 --ycf3 --ycf3 --psaA --psaB --rps14--fM --Rtct --atpA *-atpF *-atpF *-
```

やや込み入った話になったが、要するに、注釈はその時点での知識に基づいて付けられるが、データが増加し、知識が深まったときには、過去になされた注釈を改善しなければならない。が、この改善がなされていないということを例示している。

3. 原因と改善方策

2において、(i)数値のタイプミスという単純な誤りと、(ii)既に存在する注釈のアップデート、という2つの問題点を指摘した。(i)に対しては、塩基配列登録者の注釈をなんらかのプログラムによって確認すると共に、指摘があった場合には直ちに修正可能となる体制をデータベース側では確立する必要があるのではなかろうか。例えば、ここで指摘した数値ミスをデータベースに送り、データベース側は登録者の了解を得た上で修正するというルートが現時点ではない。(ii)に対しては、データベース構築に際して、新しいデータを何件入れたかということばかりでなく、注釈の品質を高めるという点も評価基準に入れることが必要であろう。また、注釈を付けるアノテーターが学問的情熱から、例えば、ここで問題にしたIle-tRNA, fMet-tRNA, eMet-tRNAの区別をしたとしても、そのこと自体は学術論文としてはまず認めてもらえず、発表の場も限られる。アノテーターに対するインセンティブを考慮した何らかの制度の導入が望まれる。このようなデータベースの支援によって、高品質データベースの提供が可能となる。

表1 注釈の修正点

```
----- Astasia longa (ALO)
60750 60822 - t Hgtg GB: 60749 60822 - t Hgtg
----- Arabidopsis thaliana (ATH)
35312 35404 + t Stga GB: 35312 35403 + t Stga
----- Cyanidium caldarium (CCA)
109401 109484 - t Icat GB: Mcat
156623 156696 + t fM GB: Mcat
----- Chaetosphaeridium globosum (CGL)
なし
----- Cyanophora paradoxa (CPA)
41064 41137 - t Hgtg GB: 41063 41137 - t Hgtg
85048 85132 + t Icat GB: Mcat
----- Chlorella vulgaris (CVU)
129350 129426 + t fM GB: Met
130470 130556 - t Sgga GB: 130470 130556 + t Sgga
----- Euglena gracilis (EGR)
30968 31041 + t fM GB: Mcat
60996 61067 + t Cgca GB: 60996 61056 + t Cgca
100773 100845 - t Hgtg GB: 100772 100845 - t Hgtg
----- Guillardia theta (GTH)
9223 9296 - t Hgtg GB: 9223 9295 - t Hgtg
50077 50162 - t Icat GB: Mcat
89030 89103 + t fM GB: Mcat
----- Lotus japonicus (LJA)
なし
----- Marchantia polymorpha (MPO)
42156 42229 - t fM GB: Met
----- Mesostigma viride (MVI)
71518 71591 + t Ptgg GB: 71508 71591 + t Ptgg
----- Nephroselmis olivacea (NOL)
26458 26530 - t Hgtg GB: 26458 26529 - t Hgtg
85192 85273 - t Ygta GB: 85192 85274 - t Ygta
----- Nicotiana tabacum (NTA)
47119 47205 + t Sgga GB: 47119 47197 + t Sgga
```

----- *Oenothera elata* (OEL)
28414 28487 + t fM GB: Met
28505 28578 + t fM GB: Met
38059 38129 - t Cgca GB: 38049 38129 - t Cgca
72024 72097 - t Wcca GB: 72025 72097 - t Wcca
138601 138672 + t Nggt GB: 138601 138682 + t Nggt
----- *Oryza sativa* (OSA)
1373 1407 - t Kttt_1 GB: 1363 1397 - t Kttt_1
3895 3931 - t Kttt_2 GB: 3902 3938 - t Kttt_2
6616 6687 - t Qttg GB: 6615 6687 - t Qttg
13003 13050 - t Gtcc_1 GB: 13010 13050 - t Gtcc_1
64229 64302 - t Ptgg GB: 64229 64303 - t Ptgg
133994 134068 - t Hgtg GB: 133991 134068 - t Hgtg
----- *Odontella sinensis* (OSI)
9852 9923 + t Nggt GB: 9852 9922 + t Nggt
34019 34090 + t Qttg GB: 34020 34089 + t Qttg
35294 35367 + t Hgtg GB: 35294 35368 + t Hgtg
96190 96274 + t lcat GB: Mcat
----- *Psilotum nudum* (PNU)
138107 138187 + t Lcaa GB: 138107 138188 + t Lcaa
----- *Porphyra purpurea* (PPU)
8465 8546 - t Lgag GB: 8465 8536 - t Lgag
26578 26648 - t Gtcc GB: 26585 26648 - t Gtcc
32547 32617 + t Cgca GB: 32547 32618 + t Cgca
34285 34371 - t lcat GB: Mcat
116603 116674 - t Aggc GB: 116604 116674 - t Aggc
138880 138953 - t Racg GB: 138890 138953 - t Racg
----- *Pinus thunbergii* (PTH)
47156 47228 - t Mcat GB: 47156..47288
70440 70526 - t Sgga GB: 70440 70526 + t Sgga
----- *Spinacia oleracea* (SOL)
8887 8934 + t Gtcc_2 GB: 8887 8944 + t Gtcc_2
9042 9113 + t Rtct GB: 9052 9113 + t Rtct
29731 29803 - t Ettc GB: 29741 29803 - t Ettc
30262 30333 + t Tggt GB: 30262 30343 + t Tggt
35159 35229 + t Ggcc GB: 35159 35228 + t Ggcc
44024 44110 + t Sgga GB: 44023 44110 + t Sgga
----- *Triticum aestivum* (TAE)
1385 1419 - t Kttt_1 GB: 1385 1409 - t Kttt_1
3907 3943 - t Kttt_2 GB: 3814 3943 - t Kttt_2
6687 6758 - t Qttg GB: 6686 6803 - t Qttg
64063 64136 - t Ptgg GB: 64063 64137 - t Ptgg
116344 116417 - t Racg GB: 116344 116419 - t Racg
134008 134082 - t Hgtg GB: 134005 134082 - t Hgtg
----- *Toxoplasma gondii* (TGO)
2380 2453 + t fM GB: Mcat
32544 32617 - t fM GB: Mcat
----- *Zea mays* (ZMA)

なし