解 説

ゲノム情報

9. 意味不明なDNA配列パターンのコンピュータ解析†

冨 田 勝††

1. 高等生物のジャンクDNA

「高等」な生物になればなるほど，そのゲノムサイズは大きくなり，蛋白質をコードする「遺伝子」の数も増える．しかしヒトのゲノムサイズは大腸菌のそれと比べて約700倍であるが，遺伝子の数は高々30倍である（表-1）．1つの遺伝子がコードする情報量は，すべての生物種においておおむね同じであるから，30倍の情報をコードするために700倍のスペースを使用していることになる．

このように，真核生物のゲノムには蛋白質コード領域（遺伝子）以外の「ジャンク」と呼ばれる非コード領域が多く存在する．非コード領域は遺伝子間の「スペーサー領域」と遺伝子内の「イントロン」に分類されるが，“高等”な生物になればなるほど，このジャンクDNAを数多くもつ傾向があり，ヒトのゲノムは95％が非コード領域であるといわれている．非コード領域のごく一部は「プロモータ」や「エンハンサ」といった転写制御部位であることが知られているが，そのほかの大半の部分の生物学的意味についてはほとんどわかっていない．

これに対し，「下等」な生物のゲノムには「遺伝子」以外の領域はほとんど存在しない．たとえば大腸菌のゲノムには，イントロンをもたない約3000の遺伝子が，ほとんんどすき間なく並んでいる．場合によっては2つの遺伝子領域がわずかではあるがオーバラップしていることもあり，少しでもメモリを節約しようという必死の努力が伺える．

高等生物のジャンクDNAはなんのために存在するのか．どのように進化してきたのか．長いDNAをもつと細胞分裂の際のDNA複製やDNA破損の際の修復や保守に余計なエネルギーを必要とする．よって，まったく意味のない配列が長い進化の過程で淘汰されずに存続してきたとは考えにくい．筆者らは，分子生物学者が注目することの少ない「非コード領域」のコンピュータ解析を数年前から行ってきた．

2. 意味不明な繰り返し配列

非コード領域を詳しく分析してみると，決してランダムな塩基配列というわけではない．数多くの繰り返しパターンが存在する．1～6塩基の繰り返しからなる「マイクロサテライト」から同一の数千塩基がゲノム全体に散在する「LINE」まで，さまざまな種類の繰返し配列があるが，中でも最も顕著なのは，300塩基弱からなる「ALU」と呼ばれる配列である（図-1）．ALUはヒトの全ゲノム中に約50～80万コピーあるといわれ，これは平均して5000塩基に1コピーあることになり，全ゲノムの約5％を占めている．なんのために，ALUやそのほかの繰返し配列が存在するのか．どのように進化してきたのだろうか．

ALUは「レトロトランスポジション」というプロセスによってRNAがDNAに逆転写されゲノ

表-1 ゲノムの長さと遺伝子の数

種	ゲノム長	遺伝子数
λファージ（ウイルス）	48,502	40
マイコプラズマ（原核生物）	580,073	500
大腸菌（原核生物）	4,700,000	3,000
酵母菌（真核生物）	18,060,000	6,000
ショウジョウバエ	140,000,000	8,000
ヒト	3,300,000,000	80,000

† Computer Analyses of Mysterious DNA Sequence Patterns by Masaru TOMITA (Department of Environmental Information, Keio University).

†† 慶應義塾大学環境情報学部

```
GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGAGGATTGCTTGAGC
CCAGGAGTTCGAGACCAGCCTGGGCAACATAGCGAGACCCCGTCTCTACAAAAAATACAAAAATTAGCCG
GGCGTGGTGGCGCGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGGATCGCTTGAGCCCAGG
AGTTCGAGGCTGCAGTGAGCTATGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACCCTGTC
TCAAAAAAAAAAAAAAAA
```

図-1　ALUのコンセンサス配列

ムに挿入されたと考えられている．ALUの下流（右側）付近にはpoly-A tailとよばれるアデニン（A）に富んだ領域があり，このあたりにはマイクロサテライトが数多くみられる．なぜこの領域にマイクロサテライトが多いかは不明であり，そもそもマイクロサテライトがどのようなプロセスでできるのかもよくわかっていない．

そこで筆者らは，ヒトのALU配列を12のサブファミリーに分類することによって，分子進化学的にそれぞれの配列がゲノムに飛び込んだ年代を推定し，それぞれのALUの右側のアデニンに富んだ領域を年代別に調べることによって，マイクロサテライトがどのような時間的推移で出現したかを分析している[4]．

3. イントロンの謎

真核生物の遺伝子の中にはイントロンという蛋白質をコードしないジャンク領域が点在する．バクテリアなどの原核生物には原則としてイントロンは存在しない．真核生物の細胞には遺伝子中のイントロンを“コメントアウト”する（読み飛ばす）ための「スプライシング」という驚異的なメカニズムが備わっている．この機構がないとイントロンも蛋白質をコードする領域の一部とみなされ，ナンセンスな蛋白質が合成されてしまう．スプライシングを行うSpliceosomeという酵素の複合体は，DNAから転写されたばかりのメッセンジャーRNAを眺め，イントロン部位を正確に認識し，切り捨てる．原核生物にはイントロンはなく，よってSpliceosomeをもたない（もつ必要がない）．

原則として，イントロンはGTではじまりAGで終わるほか，エキソンの終わりにはGが多いなど，イントロン／エキソンの境界領域にはあるパターンが存在する．**表-2**は筆者が作成した霊長類のイントロン／エキソン境界領域の塩基出現頻度表である．

なぜイントロンが存在するのか？　これには2つの説がある．

Gilbertら[1]が唱えたIntron-Early説によると，イントロンは個体の生存には寄与しないが，進化を促進する役割があったとする．つまり，イントロンで区切られた遺伝子のサブユニット（エキソン）は交差によって効率よくほかの染色体とやりとりができる．まったくランダムにやりとりするよりも，ユニット単位でやりとりした方が速く進化できるであろう．よって，イントロンは原始生命の時代から存在し，現在のバクテリアのイントロンは「退化」して消滅した，と考える説である．

一方，Intron-Late説によると，原始生命のゲノムにはイントロンはなく，それらは進化の過程

表-2　霊長類のイントロン／エキソン境界領域

TOTAL=3116;　1000=100%

	exon---------->						intron----------						
A	261	253	346	589	82	0	0	473	715	53	145	248	189
C	241	298	367	127	32	0	0	27	78	50	176	225	303
G	254	281	182	148	804	1000	0	473	129	847	206	352	254
T	243	167	103	134	80	0	1000	25	76	48	471	173	252
	----------intron---------->						<----------exon						
A	81	78	63	82	244	32	1000	0	232	214	221	219	244
C	412	431	465	408	320	754	0	0	155	208	273	314	296
G	120	85	66	63	218	1	0	1000	522	245	249	255	205
T	385	403	405	445	215	211	0	0	89	330	255	210	253

表-3　エキソンの長さ（3で割ったあまり）

	3N	3N+1	3N+2
霊長類	876(47.1%)	496(26.7%)	487(26.2%)
マウス	594(43.5%)	387(28.3%)	385(28.2%)
その他の哺乳類	146(55.3%)	67(25.4%)	51(19.3%)
その他の脊椎動物	290(46.0%)	173(27.4%)	168(26.6%)
無脊椎動物	976(40.2%)	775(32.0%)	674(27.8%)
植物	1264(43.6%)	759(26.2%)	873(30.1%)

でウィルスのごとく真核生物のゲノムに飛び込んできた，とする[2]．イントロンは増殖する術をもち，長い時間をかけてゲノム中に蔓延した．宿主の個体と種には何ら寄与せず，いわば寄生者であり，スプライシングはこれら寄生者に対抗するための苦肉の策である，と考える．

筆者ら[3]は，データベースに登録されている約8000個のイントロンをシステマティックに分析し，いくつかの興味深い傾向を発見した．たとえば，イントロンが割り込む際，エキソンの長さが3の倍数になるような場所を好んで入り込んでいる，ということを発見した（**表-3**）．このことはIntron-Late説では説明がつかずIntron-Early説を支持しているが，Intron-Late説を完全に否定することもできない．筆者は，イントロンは最初は寄生者のごとく飛び込んできたが，後に進化を促進する役割も果たしたのではないか，と考えている．

4. おわりに

このほかに，我々は非コード領域中のdinucleotide（2塩基対）とtrinucleotide（3塩基組）の分布のコンピュータ解析を行った．その結果，イントロン／エキソンの境界領域の上流（左側）と転写開始部位の前後のdinucleotide分布に[5]，そして翻訳開始部位の上流のtrinucleotide分布に[6]興味深いパターンがそれぞれみられた．

これらの知見は大量のDNA配列データを総合的に解析／分析することによってはじめて得ることができる．遺伝子1つ1つの塩基配列や機能を調べただけでは決してわからない生物の仕組みを，情報科学的にゲノム全体から眺めて明らかにする学問領域「情報生物学」は今後ますます重要になるであろう．20世紀の生物学は分子生物学に代表されるように「ハードウェアの生物学」であったが，結局生命の本質はソフトウェアであるから，21世紀は情報生物学・生命情報科学などの「ソフトウェアの生物学」が主流になるのではないだろうか．

謝辞 本研究課題はDoug Brutlag氏，Jurzy Jurka氏，清水信義氏，戸田好美君，清水友益君，浅川泰宏君，斎藤輪太郎君，および慶應義塾大学冨田研究室の学生諸君との共同で行われてきた．

参考文献

1) Gilbert, W. : Why Genes in Pieces?, Nature, 271:501(1978).
2) Cavalier-Smith, T. : Selfish DNA and the Origin of Introns, Nature, 315:283-284 (1985).
3) Tomita, M., Shimizu, N. and Brutlag, D.L. : Introns and Reading Frames : Correlation Between Splicing Sites and Their Codon Positions, Molecular Biology and Evolution, 13:9 (1996).
4) Toda, Y., Tomita, M. and Jurka, J. : Computer Analysis of Primate ALU Sequences and their Poly-A Tails （投稿準備中）.
5) Shimizu, T., Asakawa, Y. and Tomita, M. : Comparisons of Dinucleotide Distribution around Start Codons and Splicing Junctures Among Several Taxonomical Groups（投稿準備中）.
6) Saito, R. and Tomita, M. : Computer Analysis of ATG Trinucleotides near Start Codons（投稿準備中）.

（平成8年7月5日受付）

冨田 勝（正会員）

1981年慶應義塾大学工学部数理工学科卒業後，ペンシルバニア州カーネギーメロン大学コンピュータ科学部大学院留学．1983年修士課程（M.S）修了．1985年博士課程（Ph.D）修了．その後カーネギーメロン大学助手，助教授，準教授歴任．同大学自動翻訳研究所副所長併任．1988年米国立科学財団大統領奨励賞受賞．1990年より慶應義塾大学環境情報学部助教授およびカーネギーメロン大学コンピュータ科学部非常勤準教授．1996年より慶應義塾大学医学部助教授兼任．第三回私立大学情報教育協会賞受賞．Ph.D（カーネギーメロン大学），工学博士（京都大学）．専門分野：生命情報科学，遺伝子情報解析，分子生物学，言語処理，人工知能．