様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 １０ 日現在

機関番号：１７１０４
研究種目：挑戦的萌芽研究
研究期間：2013～2015
課題番号：２５６４０１１２

研究課題名（和文）遺伝子のde novo誕生の機序に迫るバイオインフォマティクス研究

研究課題名（英文）Bioinformatics analysis reveals a putative scenario for de novo origination of genes

研究代表者
矢田　哲士（Yada, Tetsushi）

九州工業大学・大学院情報工学研究院・教授

研究者番号：１０３２２７２８

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：ゲノム研究の進展にともない、多くの遺伝子がde novoに誕生していることが明らかになった。しかし、その誕生のシナリオはほとんど分かっていない。ここでは、遺伝子がde novoに誕生した前後の祖先配列を推定し、それらの祖先配列の間の変化を観察することで、遺伝子のde novo誕生のシナリオを描くことを試みた。そして、出芽酵母に至る系統における観察では、次のような遺伝子のde novo誕生のシナリオを描くことに成功した。(1)はじめに高GC含量のゲノム領域ありき、(2)そのゲノム領域に中立な突然変異が蓄積する。そして、(3)ORF長の伸長が起こり、続いて、(4)翻訳開始シグナルが生成される。

研究成果の概要（英文）：Recent advances in genome research clearly show that many genes originate de novo from non-genic DNA sequences. However, little is known about scenarios of the origination. We have sketched here a putative scenario of how genes arose from non-genic sequences by applying bioinformatics analysis to Saccharomyces cerevisiae genome. That is, we have reconstructed the homologous ancestral DNA sequences before and after de novo gene origination and have observed changes between the two sequences. A putative scenario which we have successfully sketched is as follows. (1)In the beginning was GC-rich genome region. (2)Neutral mutations were accumulated in the region. (3)ORFs were extended/combined, and then (4)translation signature (Kozak sequence) was recruited.

研究分野：ゲノム情報生物学

キーワード：遺伝子進化　de novo誕生　分子進化　比較ゲノム　バイオインフォマティクス

１．研究開始当初の背景

これまで、新しい遺伝子は、既にある遺伝子の重複や混成によって生み出されると考えられ、*de novo* に生み出されること (突然変異の蓄積によって遺伝子間領域に新しい遺伝子が生み出されること) はほとんどないと考えられてきた (Kaessmann, H., *Genome Res.* 2010)。ところが、次世代シークエンサーの登場により、RNA-seq やリボゾームプロファイリングのデータが蓄積されて遺伝子の転写や翻訳の実態が詳らかになると、これまで考えられてきたよりずっと多くの遺伝子が *de novo* に生み出されていることが明らかになった (Carvunis, A.R. *et al.*, *Nature* 2012)。Carvunis らの報告によると、*Saccharomyces cerevisiae* (*S.cer*) と *Saccharomyces paradoxus* (*S.par*) の分岐後に生み出された *de novo* 遺伝子の数は、重複や混成によって生み出された遺伝子の数の 5 倍にも達する。

２．研究の目的

どのようにして遺伝子は生まれるのだろうか？これは、生物学の根幹に関わる問いのひとつである。これまで、遺伝子の重複や混成による遺伝子の誕生と多様化の過程については多くのことが明らかになったが、遺伝子の *de novo* 誕生の過程についてはほとんど分かっていない。ここでは、*de novo* 遺伝子がどのような過程によって生み出されるのかを明らかにする。

３．研究の方法

遺伝子の *de novo* 誕生の過程を明らかにするために、ここでは、遺伝子が *de novo* に誕生した前後の祖先配列を推定し、それらの祖先配列の間に生じた変化を観察することを着想した (図 1)。すなわち、*de novo* 遺伝子の相同領域のゲノム配列を近縁種の間で比較し、遺伝子が *de novo* に誕生した前後の共通祖先のゲノム配列を推定する。そして、それらのゲノム配列について、遺伝子の *de novo* 誕生の前後に蓄積された突然変異や塩基配列に観察される統計的な特徴を調べ、それらの特徴から遺伝子の *de novo* 誕生のシナリオを描く。統計的な特徴として、(a) 蓄積された突然変異の種類と頻度、(b) ゲノム配列の GC 含量の変化、(c)ORF の伸長と縮退、(d) 翻訳開始のシグナル配列の生成と消失を調べた。

ここでは、モデル生物として *S.cer* を採用した。*S.cer* では、体系的で網羅的な解析により、$\sim 1{,}900$ の *de novo* 遺伝子が、近縁種における保存度とともに同定されている (Carvunis, A.R. *et al.*, *Nature* 2012)。加えて、*S.cer* では、多くの近縁種のゲノム配列が決定されており、また、それらの間の系統関係が明らかになっている (`http://www.yeastgenome.org/`)。

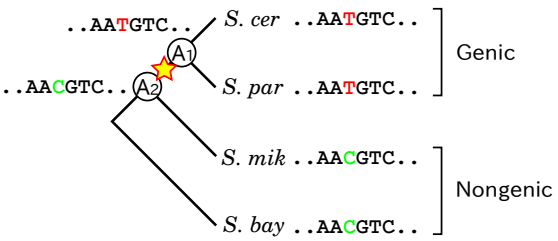

図 1: 遺伝子の *de novo* 誕生を観察するためのアイデア。ここで、*S.cer* ゲノムと *S.par* ゲノムの相同領域には *de novo* 遺伝子が存在し、*S.mik* (*Saccharomyces mikatae*) ゲノムと *S.bay* (*Saccharomyces bayanus*) ゲノムの相同領域には遺伝子が存在しないとする。また、これら 4 種のゲノム配列と系統樹が与えられているとする。今、最も少ない事象で *de novo* 遺伝子の存在を説明しようとすると、この遺伝子は、系統樹上の星印の枝で生まれたと考えるのが妥当である。そこで、これらの相同領域のゲノム配列をアラインメントすることで、遺伝子の *de novo* 誕生の前後に当たる共通祖先の最も尤もらしい相同領域のゲノム配列を推定する。この例では、*S.cer* と *S.par* の共通祖先 A_1 と *S.cer* と *S.par* と *S.mik* の共通祖先 A_2 について、相同領域のゲノム配列を推定する。以上の操作を全ての *de novo* 遺伝子について繰り返し、A_2 に当たるゲノム配列と A_1 に当たるゲノム配列の間に生じた変化を観察する。

４．研究成果

S.cer に同定されている *de novo* 遺伝子の各々について、その誕生の前後の祖先配列を推定し、それらの祖先配列の間に生じた変化から次のような統計的な特徴が観察された。(a) 遺伝子の *de novo* 誕生の前後には、中立な突然変異が蓄積される (図 2)。(b) 遺伝子の *de novo* 誕生の前後のゲノム配列は、いずれも高い GC 含量を示す (図 3)。(c) 遺伝子の *de novo* 誕生の前後では、ORF 長の伸長が数多く観察される (表 1、図 4)。(d) 遺伝子の *de novo* 誕生の前後では、翻訳開始のシグナル配列の偏った生成は観察されない (表 2)。

これらの観察に、Carvunis *et al.*(*Nature* 2012) の報告 (*de novo* 遺伝子では、その年齢が増すにつれて、翻訳開始のシグナル配列を持つ割合いが増す) を加味すると、遺伝子の *de novo* 誕生に関する次のようなシナリオを描くことができる。(1) はじめに高 GC 含量のゲノム領域ありき、(2) そ

のゲノム領域に中立な突然変異が蓄積する。そして、(3)ORF 長の伸長が起こり、続いて、(4) 翻訳開始のシグナル配列が生成される。

今回、遺伝子が *de novo* に誕生した前後の祖先配列を推定し、それらの祖先配列の間に生じた変化を観察することで、遺伝子の *de novo* 誕生のシナリオを描くことができた。一方で、新しい疑問も生まれた。遺伝子が生まれる前からゲノム領域の GC 含量が高くなり、それが保持される仕組みはどのようなものだろうか。また、 遺伝子が生まれる前に、上流のゲノム配列は転写活性を持つのだろうか。さらに、今回明らかになったシナリオは、他の生物種でも成り立つのだろうか。

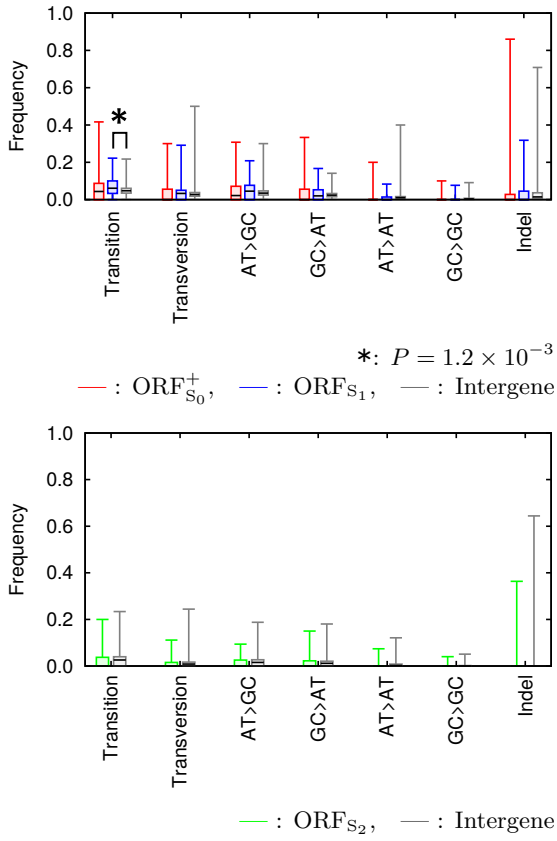

図 2: 遺伝子の *de novo* 誕生の前後には、中立な突然変異が蓄積される。上図は、共通祖先 A_1 から *S.cer* に至る間に生まれた *de novo* 遺伝子 ($\mathrm{ORF}^{+}_{\mathrm{S}_0}$ と $\mathrm{ORF}_{\mathrm{S}_1}$) について、その期間に該当するゲノム領域に蓄積された突然変異の頻度を表わす。下図は、共通祖先 A_2 から共通祖先 A_1 に至る間に生まれた *de novo* 遺伝子 ($\mathrm{ORF}_{\mathrm{S}_2}$) について、その期間に該当するゲノム領域に蓄積された突然変異の頻度を表わす。コントロールとして、同じ期間に遺伝子間領域 (intergene) に蓄積された突然変異の頻度を記す。機能的制約のない遺伝子間領域には、中立な突然変異が蓄積されていると考えられる。*de novo* 遺伝子が生まれる期間にそのゲノム領域に蓄積された突然変異の頻度は、同じ期間に遺伝子間領域に蓄積された突然変異の頻度とほぼ同じであった。

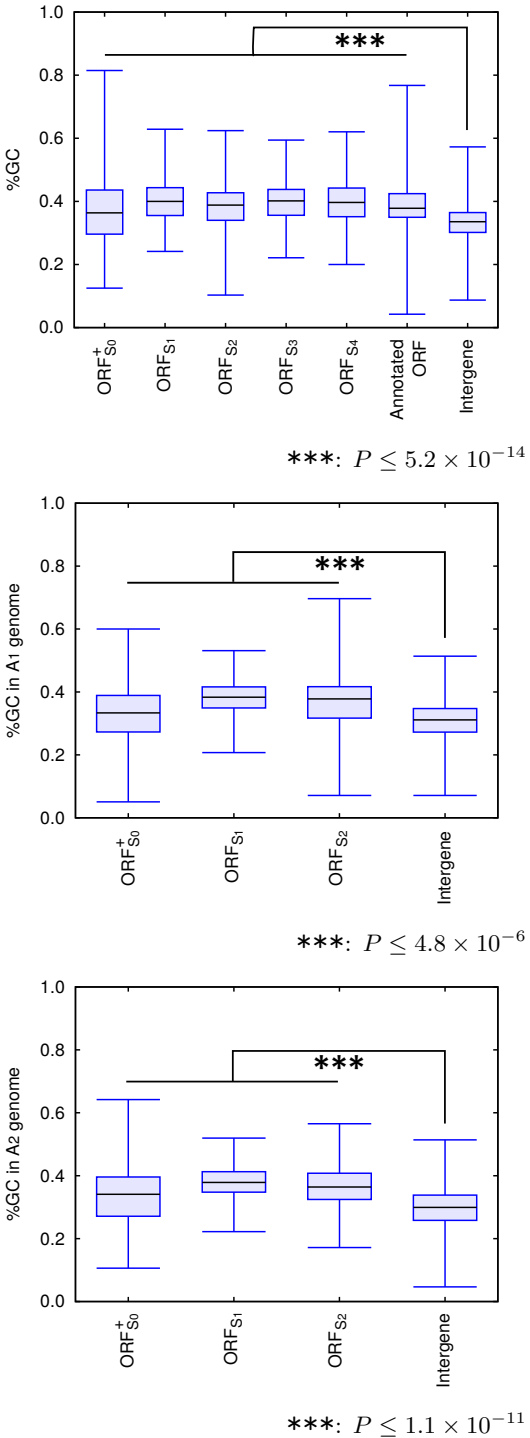

図 3: 遺伝子の *de novo* 誕生の前後のゲノム配列は、いずれも高い GC 含量を示す。上図は、*S.cer* ゲノムにおける *de novo* 遺伝子 ($\mathrm{ORF}^{+}_{\mathrm{S}_0}$ と $\mathrm{ORF}_{\mathrm{S}_{1\sim4}}$) の GC 含量、データベースにアノテーションされている遺伝子 (annotated ORF) の GC 含量、遺伝子間領域 (intergene) の GC 含量を表わす。各遺伝子の GC 含量は、遺伝子間領域に比べて高い値を示す。中図は、共通祖先 A_1 ゲノムにおける *de novo* 遺伝子 ($\mathrm{ORF}^{+}_{\mathrm{S}_0}$ と $\mathrm{ORF}_{\mathrm{S}_{1\sim2}}$) 相同領域の GC 含量、遺伝子間領域 (intergene) 相同領域の GC 含量を表わす。$\mathrm{ORF}^{+}_{\mathrm{S}_0}$ と $\mathrm{ORF}_{\mathrm{S}_1}$ は、共通祖先 A_1 では生まれていないにも関わらず、その領域の GC 含量は、遺伝子間領域に比べて高い値を示す。下図は、共通祖先 A_2 ゲノムにおけ

る *de novo* 遺伝子 ($ORF^{+}_{S_0}$ と $ORF_{S_{1\sim2}}$) 相同領域の GC 含量、遺伝子間領域 (intergene) 相同領域の GC 含量を表わす。これらの *de novo* 遺伝子は、共通祖先 A_2 では生まれていないにも関わらず、その領域の GC 含量は、遺伝子間領域に比べて高い値を示す。

表 1: 遺伝子の *de novo* 誕生の前後では、ORF 長の伸長が数多く観察される。共通祖先 A_1 から *S.cer* に至る間に生まれた *de novo* 遺伝子 $ORF^{+}_{S_0}$ と ORF_{S_1} について、A_1 の相同領域に存在する最長の ORF 長と $ORF^{+}_{S_0}$、ORF_{S_1} の ORF 長をそれぞれ比べた (上表と下表)。$ORF^{+}_{S_0}$ は、データベースにアノテーションはされていないが、転写と翻訳が実験的に確認された *de novo* 遺伝子、ORF_{S_1} は、データベースにアノテーションはされている *de novo* 遺伝子である。'<' は ORF 長が長くなった場合、'>' は ORF 長が短くなった場合、'=' は ORF 長が変わらなかった場合を表わす。ORF 長は、3 つの基準 (最上流 ATG から終止コドンまで、最上流 Kozak 配列 `A..ATG` から終止コドンまで、最上流 Kozak 配列 `A..ATG.C` から終止コドンまで) を用いて測った。$ORF^{+}_{S_0}$ では ORF 長の変化に偏りは観察されなかったが、ORF_{S_1} では多くの場合で ORF 長が長くなった。

Length changes in A_1 > *S.cer*	# of $ORF^{+}_{S_0}$		
	The most upstream `ATG`	The most upstream Kozak seq.	
		`A..ATG`	`A..ATG.C`
<	264	55	15
>	285	88	16
=	264	80	14

Length changes in A_1 > *S.cer*	# of ORF_{S_1}		
	The most upstream `ATG`^{***1}	The most upstream Kozak seq.	
		`A..ATG`^{**2}	`A..ATG.C`
<	83	24	8
>	21	3	2
=	10	4	0

^{***1}: $P = 1.3 \times 10^{-9}$, ^{**2}: $P = 9.8 \times 10^{-5}$

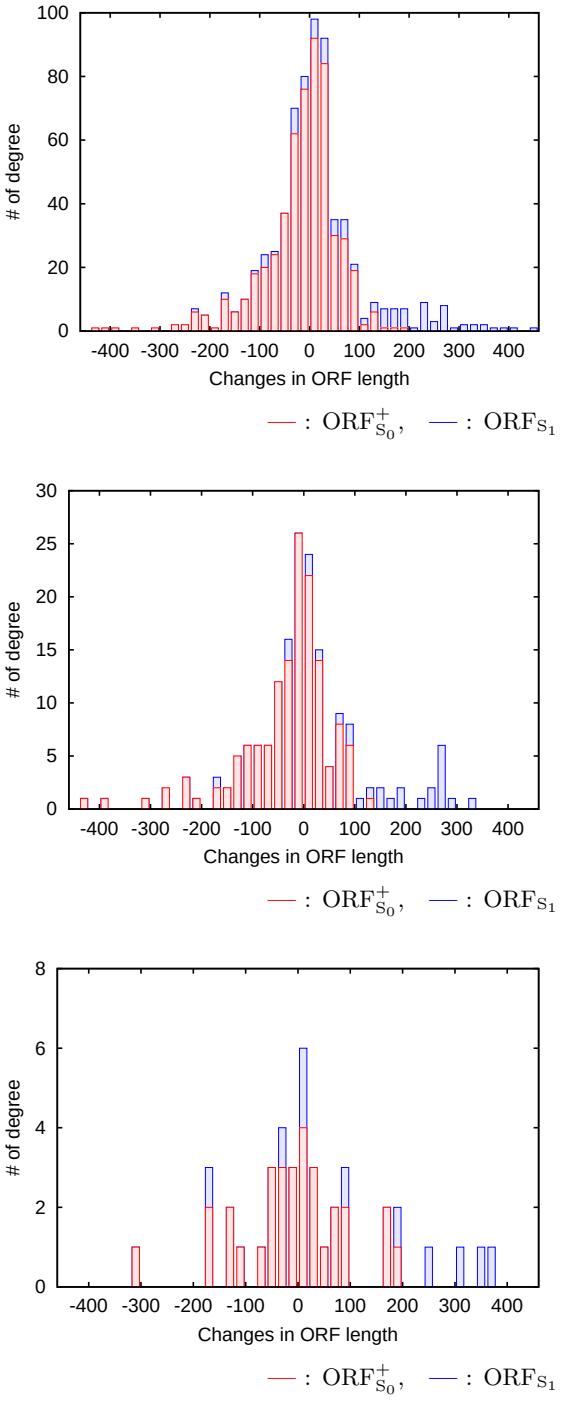

図 4: 遺伝子の *de novo* 誕生の前後における ORF 長の伸長は、多くは段階的だが、大きな変化を伴なう場合も観察される。共通祖先 A_1 から *S.cer* に至る間に生まれた *de novo* 遺伝子 ($ORF^{+}_{S_0}$ と ORF_{S_1}) について、A_1 の相同領域に存在する最長の ORF 長と $ORF^{+}_{S_0}$、ORF_{S_1} の ORF 長の差を調べた。ORF 長は、3 つの基準、最上流 ATG から終止コドンまで (上図)、最上流 Kozak 配列 `A..ATG` から終止コドンまで (中図)、最上流 Kozak 配列 `A..ATG.C` から終止コドンまで (下図) を用いて測った。

表 2: 遺伝子の *de novo* 誕生の前後では、翻訳開始のシグナル配列 (Kozak 配列) の偏った生成は観察されない。共通祖先 A_1 から *S.cer* に至る間に生まれた *de novo* 遺伝子の $ORF^{+}_{S_0}$ と ORF_{S_1} について、Kozak 配列の生成と消失の場合の数を調べた。ここでは、Kozak 配列のコンセンサス配列として、`A..ATG`(上表) と `A..ATG.C`(下表) を用いた。None は、A_1 の *de novo* 遺伝子相同領域に存在するどの ORF にも Kozak 配列が観察できず、また、*S.cer* の *de novo* 遺伝子にも Kozak 配列が観察できなかった場合を表わす。Conserved は、A_1 の *de novo* 遺伝子相同領域に存在する ORF の少なくともひとつに Kozak 配列が観察でき、また、*S.cer* の *de novo* 遺伝子にも Kozak 配列が観察できた場合を表わす。Disappear は、A_1 の *de novo* 遺伝子相同領域に存在する ORF の少なくともひとつに Kozak 配列が観察できたが、*S.cer* の *de novo* 遺伝子には Kozak 配列が観察できなかった場合を表わす。Appear は、A_1 の *de novo* 遺伝子相同領域に存在するどの ORF にも Kozak 配列が観察できなかったが、*S.cer* の *de novo* 遺伝子には Kozak 配列が観察できた場合を表わす。いずれの場合においても、Kozak 配列の偏った生成は観察されなかった。

Kozak seq. (`A..ATG`) in A_1 > *S.cer*	# of $ORF^{+}_{S_0}$	# of ORF_{S_1}	Total
None	407	50	457
Conserved	223	31	254
Disappear	176	51	227
Appear	58	20	78

Kozak seq. (`A..ATG.C`) in A_1 > *S.cer*	# of $ORF^{+}_{S_0}$	# of ORF_{S_1}	Total
None	725	99	824
Conserved	45	10	55
Disappear	32	8	40
Appear	27	9	36

５．主な発表論文等

(研究代表者、研究分担者及び連携研究者には下線)

[学会発表] (計 3 件)

① Yada T., Taniguchi T. A putative scenario for *de novo* gene origination in *Saccharomyces cerevisiae* genome, 日本進化学会 2016 年年会, 2016 年 8 月 25 日〜28 日, 東京工業大学大岡山キャンパス.

② Yada T. A putative scenario for *de novo* gene birth in *Saccharomyces cerevisiae* genome, BIT2016, 2016 年 3 月 3 日〜4 日, National Yang-Ming University, Taiwan.

③ Yada T., Taniguchi T. Observing *de novo* gene birth through reconstruction of ancestral DNA sequences, 日本バイオインフォマティクス学会 2015 年年会, 2015 年 10 月 29 日〜31 日, 京都大学宇治キャンパス.

６．研究組織

(1) 研究代表者

矢田　哲士 (YADA, Tetsushi)
九州工業大学・大学院情報工学研究院・教授
研究者番号：10322728