Title	同義語コドン使用の偏りと5'非翻訳領域の保存性の関係のコンピュータ解析
Sub Title	Correlation between sequence conservation of 5' untranslated region and codon usage bias
Author	坂井, 寛章(Sakai, Hiroaki) 冨田, 勝(Tomita, Masaru)
Publisher	慶應義塾大学湘南藤沢学会
Publication year	2001-03
Jtitle	優秀修士論文
JaLC DOI	
Abstract	本書は、9種のバクテリアのDNA配列を用いて、タンパク質の生産効率の良し悪しを左右すると言われる同義語コドンの使用の偏りと、タンパク質の生産開始ポイントにおいて重要視されるShine-Dalgarno(SD)配列の保存性の相関関係についてコンピュータ解析を行ったものである。
Notes	冨田勝研究室 2000年 修士論文2000年度(平成12年度)
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=0302-0000-0433

SFC－MT2000－002

修士論文 同 Keio University Shonan Fujisawa Academic Society

義語コドン使用の偏りと5'非翻訳領域の保存性の関係のコンピュータ解析

2000年

坂井　寛章
慶應義塾大学 大学院 政策・メディア研究科 修士課程
冨田　勝　研究室

慶應義塾大学湘南藤沢学会

修士論文　2000年度　(平成12年度)

同義語コドン使用の偏りと5'非翻訳領域の保存性の関係のコンピュータ解析

慶応義塾大学大学院　政策・メディア研究科

坂井　寛章

修士論文要旨 2000年度 (平成12年度)

同義語コドン使用の偏りと5'非翻訳領域の保存性のコンピュータ解析

論文要旨

本研究では、9種のバクテリアのDNA配列を用いて、タンパク質の生産効率の良し悪しを左右すると言われる同義語コドン使用の偏りと、タンパク質の生産開始ポイントにおいて重要だとされるShine-Dalgarno (SD) 配列の保存性の相関関係についてコンピュータ解析を行った。これまで同義語コドン使用の偏りとタンパク質生産効率の関係については詳細な研究が行われてきたが、SD配列の保存性との関係については本研究が初めての研究であり、その関係が明らかになることは、生物のタンパク質生産メカニズムを理解する上で重要な知識となることが期待される。解析結果により、大腸菌をはじめとする6種のバクテリアで、同義語コドン使用の偏りとSD配列の保存性の間に正の相関が見られた。このことから、同義語コドン使用の偏りが強い遺伝子は、偏りの弱い遺伝子よりもより強く保存されたSD配列を持ち、そのことで高いタンパク質生産効率を維持しているのではないかということが示唆された。

また、マウスについても同様の解析を行った。マウスのような真核生物ではバクテリアに見られるようなSD配列は存在しないが、Kozakのコンセンサス配列の存在が知られており、同じようにタンパク質生産の開始段階で重要な役割を担っていると言われている。そこで、同義語コドン使用の偏りと、Kozakのコンセンサス配列の保存性の相関関係について解析を行った。その結果、同義語コドン使用の偏りの強い遺伝子は、偏りの弱い遺伝子よりもより保存されたKozakのコンセンサス配列を持っていることが明らかになった。

本研究結果の重要なポイントは、バクテリアやマウスでは、同義語コドン使用パターンと共に、SD配列やKozakのコンセンサス配列にも、進化的な選択圧が働いたということである。進化の過程で、同義語コドン使用パターンには自然選択が起こり、その遺伝子の翻訳効率の決定に大きく影響していることは'80年代から研究され詳細に理解されている。今回の解析により、タンパク質を大量に生産する必要のある遺伝子では、タンパク質の生産開始の段階でも、同義語コドン使用パターンと連動した形でDNA配列に進化的な圧力がかかっていたということが明らかになった。

キーワード

1. 翻訳　2. 同義語コドン　3.CAI値　4.Shine-Dargalno配列　5.Kozakのコンセンサス配列

慶応義塾大学大学院政策・メディア研究科
坂井寛章

Abstract of Master's Thesis

Academic Year 2000

Correlation between sequence conservation of 5' untranslated region and codon usage bias.

Summary

In this study, I have analyzed the correlation between synonymous codon usage bias and Shine-Dalgarno (SD) sequence conservation, using complete genome sequences of 9 procaryotes. Synonymous codon usage bias is said to be related with the efficiency of the protein synthesis and SD sequence is known to be important to start protein synthesis. Previously it has well investigated that synonymous codon usage bias is associated with the efficiency of protein synthesis. On the other hand, the correlation between synonymous codon usage bias and SD sequence conservation has not studied well. Thus this study is expected to help to understand the mechanism of protein synthesis more precise. As a result, I found that there exists a clear correlation between synonymous codon usage bias and SD sequence conservation in 6 bacterial species such as *Escherichia coli*. I could suggest that genes with highly biased codon usage have well-conserved SD sequence to maintain the high efficiency of protein synthesis.

I have also analyzed the correlation between synonymous codon usage bias and the sequence conservation of Kozak's consensus sequence in mouse genes. Though there is no sequence like SD sequence found in bacterial species, Kozak's consensus sequence is well known in vertebrate species and is said to be important to start protein synthesis. It has revealed that a clear correlation between synonymous codon usage bias and the extent of conservation of Kozak's consensus sequence, suggesting that genes with highly biased synonymous codon usage have well-conserved Kozak's consensus sequence.

Finally we have concluded that the extent of conservation of SD sequence in bacteria and Kozak's consensus sequence in mouse genes is important factor in modulation of the efficiency of protein synthesis as well as synonymous codon usage bias.

Key Words

1.translation 2.synonymous codon 3.Codon Adaptation index value
4.Shine-Dargalno sequence 5.Kozak's consensus sequence

Keio University Graduate School of Media and Governance

Hiroaki Sakai

目 次

第1章　研究背景

1.1　遺伝情報とは

遺伝物質が核酸であるという考えは、1928年、Griffithらの形質転換の発見に端を発し、1944年、Averyらの研究により、デオキシボ核酸(DNA)こそが形質転換因子であることが化学的に示された。

DNAは化学結合したサブユニットの配列で形成されており、窒素を含む塩基とペントース(環状構造の五炭糖)、およびリン酸基から成り立っている。窒素を含む塩基はピリミジンとプリンという二つの型に分かれる。DNAは4種類の塩基から構成されており、2種類のプリン塩基にはアデニン(adenine)とグアニン(guanine)、ピリミジン塩基にはシトシン(cytosine)とチミン(thymine)が含まれる。塩基は通常頭文字で表現される。ペントースには五つの炭素原子が含まれ、それぞれに1'(1ダッシュ)〜5'という数字が割り振られている(図1.1参照)。

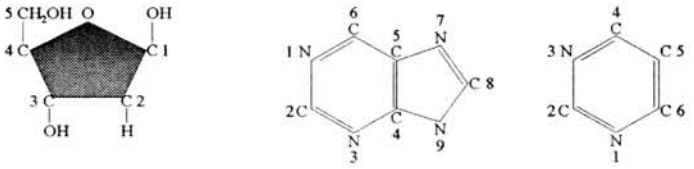


図1.1：ペントース(左)　と塩基　塩基はプリン塩基(中央)とピリミジン塩基(右)の2種類存在する。

図1.2にDNAの骨格の略図を示す。ペントース環の5'の位置にはリン酸基が結合しており、次のペントース環の3'の位置とホスホジエステル結合を形成してつながっている。つまり、このように、DNAの一方の端には、リン酸基が結合した5'が遊離しており、もう一方の端には、3'が遊離した状態で存在していることになる。DNAの塩基配列は、5'-3'の方向、すなわち左側に5'の末端を書き、右側に3'の末端がくるように表現するのが慣例となっている。

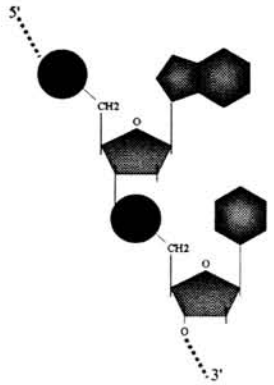


図 1.2：DNA の構造モデル　塩基に糖が結合するとヌクレオシドと呼ばれる物質になる。これにリン酸基が一つ加わって塩基-糖-リン酸の構造をした物質はヌクレオチドと呼ばれる。ヌクレオチドの糖の 3'-ヒドロキシル基は、次のヌクレオチドの 5'-リン酸基とホスホジエステル結合を形成し、ヌクレオチドがつながる。

ここまで、遺伝情報を司る DNA の構造について述べてきたが、DNA の塩基配列はつまり ATGC という 4 文字が連なった長大な文字列であると考えることができる。個体が、受精卵から無数の細胞分裂を繰り返して成体となるのに必要な情報は、全てこの DNA に書き込まれていると言っても過言ではない。そうした情報は無数の遺伝子として DNA 中にちりばめられており、それぞれの遺伝子には、タンパク質を合成するためのいわば設計図が記されている。タンパク質は、20 種類のアミノ酸が複数個連なって構成されており、このアミノ酸の配列によって、タンパク質の構造は一義的に決定される。タンパク質中のアミノ酸の数は数十から数百にもなり、長いものだと千を超えるものもあるから、20 種類のアミノ酸の組合わせは何通りにもなるわけで、その数だけタンパク質の種類も多様になるわけである。遺伝子には、アミノ酸の配列に関する情報が記されている。

1.2　遺伝子からタンパク質が出来るまで ～遺伝子発現～

タンパク質というのは全てもとは生物由来の物質であり、DNA にコードされている遺伝子が発現することによって生産される。遺伝子発現とは、DNA からタンパク質合成までの一連の流れを指すが、それは大きく二つのステップに分けることが出来る。

1.2.1　転写

転写は、RNA ポリメラーゼという酵素によって行われる、遺伝子発現の最初のステップである。RNA ポリメラーゼは、DNA 中でタンパク質となる領域 (遺伝子) を認識し、メッセンジャー RNA(mRNA) と呼ばれる物質を合成する。mRNA は、DNA 中の遺伝子のアミノ酸配列を正確にコピーしたもので、この後に説明する、翻訳という転写の次のステップで必要になる。mRNA ではチミン (T) の代わりにウラシル (U) が使われる。

1.2.2 翻訳

DNA は ATGC という 4 文字から成る文字列であると述べたが、そうすると、タンパク質をコードするアミノ酸配列もまた、ATGC4 文字の文字列であるといえる。4 文字でどうやって 20 種類のアミノ酸を区別するのかは、非常にシンプルかつ巧妙な仕組みになっている。

それぞれのアミノ酸は、ATGC4 文字のうちの 3 文字で表現されており。例えば、CTG(mRNA では CUG) という 3 文字はロイシンをコードし、GCA はアラニンをコードしている。この、アミノ酸を表現する 3 文字の文字列を「コドン」と呼ぶ。ATGC を使った 3 文字の組合わせというのは、$4 \times 4 \times 4$ で 64 通り考えられるが、アミノ酸の種類は 20 種類である。64 通りのコドンは、3 通りはタンパク質コード領域の終りを意味する「終止コドン」をコードしており、アミノ酸はコードしていない。残りの 61 通りのコドンでアミノ酸をコードしているわけだが、一つのアミノ酸に対して複数のコドンが対応することによって、全てのコドンが何かしらのアミノ酸をコードするようになっている。例えば、ロイシンというアミノ酸には、CUG の他に、UUA、UUG、CUU、CUC、CUA の 5 つのコドンが対応しており、合計 6 つのコドンによってロイシンが表現される。このように、一つのアミノ酸をコードする複数のコドンを同義語コドンと呼ぶ。

翻訳は、mRNA 上のコドンに対応するアミノ酸を一つずつつなげていく作業で、文字配列からアミノ酸配列への翻訳作業とも解釈できることから、このように呼ばれている。実際にアミノ酸をつなげてタンパク質を合成するのは、リボソームと呼ばれる、複数の分子が会合した複合体の仕事となる。

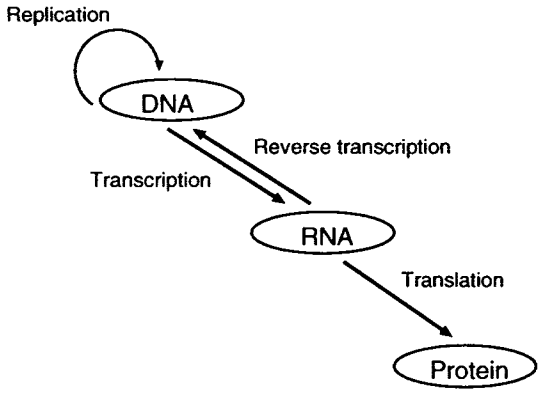


図 1.3：セントラルドグマ　あらゆる細胞で、遺伝情報の流れは DNA から RNA(転写 [transcription])、RNA からタンパク質 (翻訳 [translation]) へ向かう。また自己複製のために DNA から DNA(複製 [replication]) という矢印も存在する。逆転写酵素の発見により、RNA から DNA へ向かう矢印も追加された。

1.3 コドン使用の偏り

遺伝子の中での同義語コドンの使用頻度はランダムではなく、生物種固有の偏りというものを持ってるということが、'80 年に Grantham らによって示された [1]。表 1.1 は大腸菌の *tuf* 遺伝子のコドン頻度の一部であるが、ロイシン (Leu) では 6 つある同義語コドンのうち CUG を多用し、アルギニン (Arg) では CGU を多用しているといったように、各アミノ酸で特定のコドンを多用する傾向が見られる。大腸菌と酵母菌において、同義語コドン使用の偏りは細胞内の tRNA 分子量と相関関係にあるということが国立遺伝学研究所の池村淑道氏によって示された [2, 3]。'85 年には、複数の遺伝子のタンパク質生産量を測定

し、生産量の多い遺伝子ほど、codon usage bias は激しいということを明らかにした。このことから、同義語コドン使用の偏りはタンパク質の生産効率に関わっているということが示唆された [4]。

コドンからアミノ酸へ翻訳するためには、リボソームと、tRNA という分子の存在が不可欠である。tRNA 分子にはアミノ酸が結合し、コドンとアミノ酸を結び付ける役割を担っている。例えば CUG というコドンには、必ず CUG に対応する tRNA がロイシンを運んでくるわけである。このようにそれぞれのコドンに対して、対応する tRNA は異なる。

しかし、リボソームは CUG に対応する tRNA を自分で見つけて取ってくることは出来ない。そのため、CUG に対応する tRNA がリボソームにやってくるまで、何回も tRNA の選別を行う。ここで、tRNA 分子が仮に細胞中に 10 分子存在し、その中で CUG に対応する tRNA が 1 分子しか含まれていなかった場合、リボソームが CUG に対応する tRNA 分子に出会う確率は 1/10 だが、これが 5 分子含まれているときは 1/2 になるわけである。つまり、あるコドンに対応する tRNA 分子の量が多ければ多いほど、翻訳はスムーズに行えるようになるということである。また、ゲノム DNA には、進化の過程で重複して存在している遺伝子が多く、tRNA 遺伝子もまたゲノム中で重複して存在しているケースが多い。'99 年には、18 種のバクテリアゲノムで、ゲノム中の各 tRNA 遺伝子の重複遺伝子数と同義語コドン使用の偏りの間に正の相関関係があることが明らかになった [5]。

同義語コドン使用の偏りは、発現量の高い遺伝子において特に顕著であることが分かり、これは効率よく翻訳を進めるための選択であるという説が一般に広く受け入れられるようになった [6, 7, 8, 9, 10]。

'91 年には M. Bulmer によって *selection-mutation-theory* が提唱され、同義語コドン使用の偏りは、進化の過程で、翻訳効率を高めるコドンを選択する力と、そうでないコドンへの変異とのバランスで決定されると定義された [11]。発現量の高い遺伝子では選択の影響が変異のそれよりも強いために偏ったコドン頻度が形成され、発現量の低い遺伝子では変異の影響方が強いため、コドン頻度はランダムに近くなっていくというものである。

アミノ酸のコード表を見ると、同義語コドンのうち、2 文字目までは変わらず、3 文字目だけ変化しているものが多いのに気が付く。つまり、アミノ酸コードはコドンの 3 文字目に対して寛容であると言うことができるわけで、ゲノム全体の変異圧を最も反映しやすいのがコドンの第 3 文字目なのである。一般的に、ゲノム全体の G+C 含有量がそのゲノムの変異圧を把握するのに用いられる。

選択圧と変異圧のバランスは種により大きく異なるもので、*Mycoplasma capricolum* はゲノムの G+C 含量が 25%と低く、93%のコドンが A あるいは U で終わっている。一方 *Micrococcus luteus* のゲノム G+C 含量は 74%であり、約 95% のコドンが G あるいは C で終わっている。これらの種では、ほとんどの遺伝子が極めて似たコドン頻度を持っており、選択圧は変異圧に圧倒されている [12, 13]。

一方、*E.coli* と近縁種である *Serratia marcescens* はゲノムの G+C 含量は 51%で (*E.coli* は 59%)、発現量の低い遺伝子はゲノム G+C 含量を反映しているが、発現量の高い遺伝子ではコドン頻度は *E.coli* に非常に近くなることが知られている。こういった傾向は原核生物に限らず、*Dictyostelium discoideum* という粘菌 (G+C 含量は 25%) ゲノムでは、発現量の低い遺伝子のコドン 3 文字目の G+C 含量は 10%程度であるが、発現量の高い遺伝子では約 30% に増加し、翻訳を効率化するコドンの頻度を反映している [10]。これらの種は、変異圧よりも選択圧が上回った結果、遺伝子間でコドン使用に変化が見られるわけである [14]。

アミノ酸	コドン	頻度	アミノ酸	コドン	頻度
Leu	UUA	0		CGG	0
	UUG	0		AGA	0
	CUU	1		AGG	0
	CUC	0	Ile	AUU	3
	CUA	0		AUC	26
	CUG	27		AUA	0
Arg	CGU	21	Glu	GAA	30
	CGC	2		GAG	7
	CGA	0			

表 1.1：大腸菌 *tuf* 遺伝子のコドン頻度の一部

第2章　研究目的

先にも述べたように、遺伝子からタンパク質が合成されるまでには様々なステップを踏む過程がある。その中の翻訳の過程で、タンパク質の生産量を制御出来るポイントとしてコドン使用の偏りが考られるわけであるが、発現量の高い遺伝子が進化の過程で高い翻訳効率を維持するためにそうした手段を取ったならば、他にも何か翻訳効率に影響を与えるような要素があってしかるべきである。我々はその一つとしてmRNAの5'非翻訳領域に注目した。mRNAの5'側の非翻訳領域には、リボソームのサブユニットがmRNAを認識するためのコンセンサス配列が存在する。この後にも説明するように、そうしたコンセンサス配列に変異が入ると、リボソームのmRNA認識能が低下し、翻訳量が低下すると言われている。それならば、効率良く発現する必要のある遺伝子にとっては、コンセンサス配列に変異が入ることは不利な条件になるので、より保存されているような選択圧が進化の過程でかかってきたのではないかという仮説をたてることができる。

我々はこの仮説を検証するために、原核生物ゲノムと真核生物ゲノムにおいて、同義語コドン使用の偏りと5'非翻訳領域の保存性の間の相関関係についてコンピュータ解析を行った。

第3章　原核生物におけるコドン使用の偏りと 5' 非翻訳領域の保存性の関係

3.1　原核生物の翻訳開始機構

DNA から RNA へ転写されたあと、RNA からタンパク質が合成される翻訳機構は、翻訳開始、翻訳伸長、翻訳終止と大きく 3 つのプロセスに分けることができるが、ここでは本研究に直接関係のある翻訳開始のメカニズムについて説明していく。

3.1.1　翻訳開始

翻訳開始のプロセスでまず最初に起こるのは、先に述べたリボソームの mRNA への結合である。原核生物のリボソームは、30S と 50S という二つのサブユニットから構成されており、翻訳開始は、本来のリボソームでは起こらず、解離している 30S サブユニットによって起こる。しかし、30S サブユニットはそれ単体では mRNA に結合することは出来ず、それにはさらに開始因子 (IF(Initiation Factor)) 3 が必要である。原核生物には三つの開始因子があり、それぞれ IF-1、IF-2、IF-3 と名付けられている。それぞれ違った役割を持っており、IF-3 は 30S サブユニットと結合し、mRNA と 30S サブユニットを結合させるのを補助する役割を担っている。

IF-3 が結合した 30S サブユニットは、mRNA 上のリボソーム結合部位を認識して結合する。その後、mRNA の中でタンパク質をコードする領域の先頭を示す特別な開始コドンを認識する。開始コドンはたいてい「AUG」であるが、原核生物では GUG や UUG も使われる。開始コドンが認識された後に、開始コドンに対応する特別な tRNA 分子が結合し、30S サブユニットと 50S サブユニットが会合することで 70S のリボソームが形成され、翻訳が開始される。IF-3 は 30S と 50S サブユニットが会合する前に遊離し、他の 30S サブユニットを探して再びこの一連の反応を開始する。

原核生物において、mRNA 上のリボソーム結合部位の認識配列は、多くの種で、タンパク質コード領域の手前に Shine-Dalgarno(SD) 配列と呼ばれるコンセンサス配列が存在することが知られている。

3.1.2　Shine-Dalgarno 配列

Shine-Dalgarno 配列 (以降 SD 配列) を認識するのは、16S rRNA と呼ばれる、30S サブユニットを構成する RNA 分子の一つである。図 3.1 に SD 配列と 16S rRNA の結合の様子を示す。16S rRNA の 3' 末端近くには、mRNA 上の SD 配列に相補的な配列が存在し、両者が塩基対合を形成することで 30S サブユニットが mRNA へ結合することができる。図 3.1 に示している SD 配列は大腸菌で知られるものであ

り、通常 6 文字から 7 文字のコアとなる配列が存在し、開始コドンの 5' 末端側十数 bp の位置に保存されている。

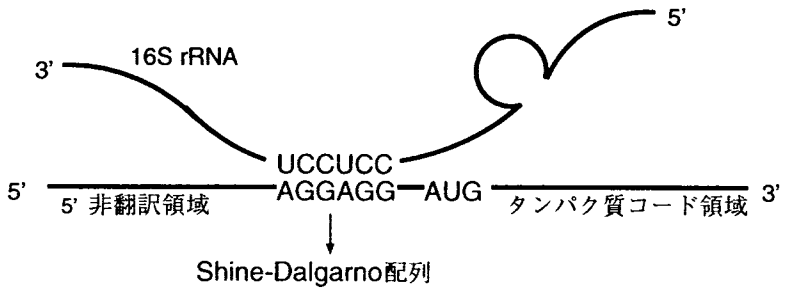


図 3.1：SD 配列と 16S rRNA の対合 (大腸菌の例)

3.2 解析に使用したデータ

今回の解析の対象とした生物種を以下に示す。

species	class
Archaeoglobus fulgidus[16]	Archaebacteria
Bucillus subtilis[17]	Eubacteria
Escherichia coli K-12[18]	Eubacteria
Haemophilus influenzae Rd[19]	Eubacteria
Methanobacterium thermoautotrophicum[20]	Archaebacteria
Methanococcus jannaschii[21]	Archaebacteria
Mycoplasma genitalium[22]	Eubacteria
Mycoplasma pneumoniae[23]	Eubacteria
Synechocystis PCC6803[24]	Eubacteria

表 3.1：解析の対象とした原核生物 9 種

3.3 方法

3.3.1 CAI 値

CAI 値は、Sharp らによって提案された、コドン頻度の偏りを定量化する指標の一つであり、ribosomal protein 遺伝子のような、一般的に発現量が高いとされる遺伝子のコドン頻度を参照して、全ての遺伝子についてコドン頻度の偏りを数値化するものである [25]。

まず、参照する遺伝子群のコドン頻度から、RSCU(relative synonymous codon usage) 値を算出し、レファレンステーブルと呼ばれるものを作成する。RSCU 値は、各アミノ酸における同義語コドンの使用頻度の平均に対する、各同義語コドンの割合を意味する。

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}}$$

X_{ij} は i 番目のアミノ酸に対する j 番目の同義語コドンの使用回数である。

$$w_{ij} = \frac{RSCU_{ij}}{RSCU_{imax}} = \frac{X_{ij}}{X_{imax}}$$

各アミノ酸の中で最も使用頻度が高いコドンを 1 とした、その他の同義語コドンの使用頻度を w 値として定義する。以上のようにして算出した w 値をもとにして、遺伝子ごとに CAI 値を計算する。

$$CAI = (\prod_{k=1}^{L} w_k)^{\frac{1}{L}}$$

ここで w_k は、遺伝子の中の k 番目のコドンの w 値、L はその遺伝子の中に含まれるコドンの総数を表す。

今回の解析ではレファレンステーブルを作成するための遺伝子として、ribosomal protein、elongation factor、heatshock protein、outer membrane protein、RNA polymerase subunit、をコードする遺伝子群を利用した。

表 3.2 に大腸菌の W 値の一覧を示す。ロイシン (Leu) を始め多くのアミノ酸について、特定のコドン (W value = 1.0000) 以外のコドンの W 値が低いことが分かる。このことは、そのアミノ酸については、その特定のコドンを偏って多用しているということを意味している。

Amino acid	Codon	W value	Amino acid	Codon	W value	Amino acid	Codon	W value
Leu	UUA	0.1156	Ala	GCU	0.9645	Thr	ACU	0.5252
	UUG	0.1180		GCC	0.5950		ACC	1.0000
	CUU	0.1095		GCA	0.8349		ACA	0.1605
	CUC	0.1255		GCG	1.0000		ACG	0.3742
	CUA	0.0263	Val	GUU	1.0000	Ile	AUU	0.6069
	CUG	1.0000		GUC	0.4139		AUC	1.0000
Arg	CGU	1.0000		GUA	0.5343		AUA	0.0503
	CGC	0.6190		GUG	0.6713	Asn	AAU	0.4131
	CGA	0.0520	Gly	GGU	1.0000		AAC	1.0000
	CGG	0.0693		GGC	0.8198	Phe	UUU	0.6556
	AGA	0.0324		GGA	0.1245		UUC	1.0000
	AGG	0.0128		GGG	0.1671	Tyr	UAU	0.8963
Pro	CCU	0.2400	Ser	UCU	0.8463		UAC	1.0000
	CCC	0.1033		UCC	0.7547	Glu	GAA	1.0000
	CCA	0.2721		UCA	0.2625		GAG	0.3433
	CCG	1.0000		UCG	0.3264	Cys	UGU	0.7578
Gln	CAA	0.3259		AGU	0.3731		UGC	1.0000
	CAG	1.0000		AGC	1.0000	His	CAU	0.7513
Lys	AAA	1.0000					CAC	1.0000
	AAG	0.3122				Asp	GAU	1.0000
							GAC	0.9791

表 3.2：大腸菌 (*E.coli*) の W 値一覧

3.3.2 SD配列の保存性

各遺伝子のSD配列の保存度は、SDモチーフ配列の相補配列と、開始コドンの上流配列(5'UTR)との間の自由エネルギーを計算することによって数値化した。

自由エネルギーとは、ある反応で要求される、あるいは放出されるエネルギー量を表す一つの熱力学定数である。自由エネルギーは、パラメーターΔGで表され、kcal/molで測定される。エネルギーを要求する反応は正の値をとり、エネルギーを放出する反応は負の値をとる。塩基が対合した構造を取るときはエネルギーが放出されるので、全体として負の値をとることになる。構造の安定性は放出されたエネルギーの量で決定されるので、負の値が大きいほど、その二本鎖は安定した対合を形成するということになる。

SDモチーフ配列はTompa, M.氏が開発したアルゴリズムによって予測された7文字配列を用い、自由エネルギーの計算は、慶應大学の長田氏のアルゴリズムを参考に行った[26, 27]。

今回の解析で用いたそれぞれの種のSDモチーフ配列を以下に示す。

species	SD motif sequence(5' → 3')
A.fulgidus	GGAGGTG
B.subtilis	AAGGAGG
E.coli	TCAGGAG
H.influenzae	TAAGGAG
M.thermoautotrophicum	CGGTGAT
M.jannaschii	GGTGATA
M.genitalium	CGGTTGT
M.pneumoniae	GGAGGTG
Synechocystis	CGATCGC

表3.3：原核生物各種のSDモチーフ配列(Tompa氏のアルゴリズムで予測された)

まず、各遺伝子の5'UTRから40bpを抽出し、SDモチーフ配列の相補配列とアラインメントする。塩基対合の自由エネルギーはダブレットの組合わせによって決定される(表3.4参照)ので、7文字配列をアラインメントさせたときにはダブレットが6通り存在することになる。40bpの5'UTRの5'末端側からSDモチーフ配列を1bpずつスライドさせ、各ポジションでの6通りのダブレットの自由エネルギーの合計を算出した。このとき、ギャップ、誤対合の許容はしなかった。図3.2に示すように、mRNA上でSDモチーフ配列がよく保存されている領域では、モチーフ配列の相補配列とmRNA間で安定した塩基対合を形成することが出来るため、自由エネルギーは低くなる。

以上のようにして、全遺伝子について自由エネルギーを計算した後、CAI値別に、上位200、平均値周辺の200、下位200遺伝子について、5'UTR各ポジションにおける自由エネルギーの平均値を求め、グラフにプロットした。2種の*Mycoplasma*菌はゲノムサイズが小さいため、上位、平均、下位それぞれ50遺伝子の平均で計算した。

doublet		Δ G(kcal/mol)	doublet		Δ G(kcal/mol)
AU doublet	AA			GA	
		-0.9			-2.3
	UU			CU	
	AU			GU	
		-0.9			-2.1
	UA			CA	
	UA		GC doublet	CG	
		-1.1			-2.0
	AU			GC	
mixed doublet	CA			GC	
		-1.8			-3.4
	GU			CG	
	CU			GG	
		-1.7			-2.9
	GA			CC	

表 3.4：ダブレット (2 塩基配列) の組合せによる塩基対合のエネルギー　上段のダブレットは 5' から 3' 方向に伸びる鎖を左から右方向へ表し、下段のダブレットは、それと相補的な配列を左から右へ 3' から 5' の方向へ表している。[28] より。

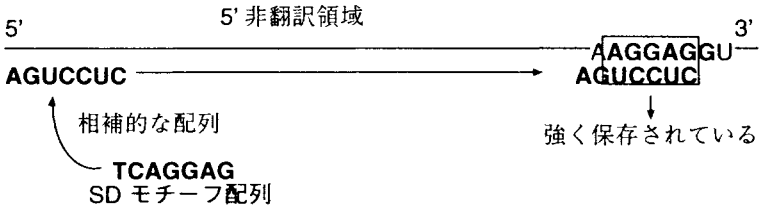


図 3.2：自由エネルギーの計算　mRNA 配列の 5' 末端に SD モチーフ配列の相補配列をアラインメントし、自由エネルギーを測定する。その後 3' 側に 1bp ずらしながら各ポジションでの自由エネルギーを算出していく

3.4 結果

3.4.1 CAI 値と自由エネルギーの関係

図 3.3 に、各種における、開始コドン上流領域の自由エネルギーの変化を示す。*B.subtilis*、*E.coli*、*H.influenzae* の 3 種については、-12〜- 10 のポイントで、自由エネルギーが大きく下がる領域が現れているのが分かる。このことは、この領域で SD 配列のモチーフが強く保存されていることを意味している。さらに、CAI 値が低い 200 の遺伝子から、平均値周辺の 200、高い 200 の遺伝子になるにつれて、自由エネルギーの低下が大きくなっていくことも分かる。このことは、古細菌 3 種についても同じような結果になった。このことから、*B.subtilis*、*E.coli*、*H.influenzae*、*A.fulgidus*、*M.thermoautotrophicum*、*M.jannaschii* の 6 種では、同義語コドン使用の偏りが強い遺伝子は、低い遺伝子に比べてより保存された SD 配列を持っているということが示唆される。

一方、SD 配列に明確なコンセンサスを持たないことで知られる *Mycoplasma* 菌では、開始コドン上流に自由エネルギーが低下する領域は現れなかった。

Synechocystis は、Tompa, M. の解析において、開始コドン上流 20bp 以内に SD モチーフ配列を持たない唯一の生物であり、今回の解析では、上流 40bp 以内に現れる高頻度の 7 文字配列を用いたが、自由エネルギーが大きく低下する領域は現れなかった。

A.fulgidus と *M.pneumoniae* の 2 種については、-1 のポジションで自由エネルギーが大きく低下しているが、これは、SD モチーフ配列の最後の 2 文字が、「TG」であり、今回の解析で抽出してきた上流配列に開始コドンを含んでいたため、開始コドンと対合して自由エネルギーが低下したものである。

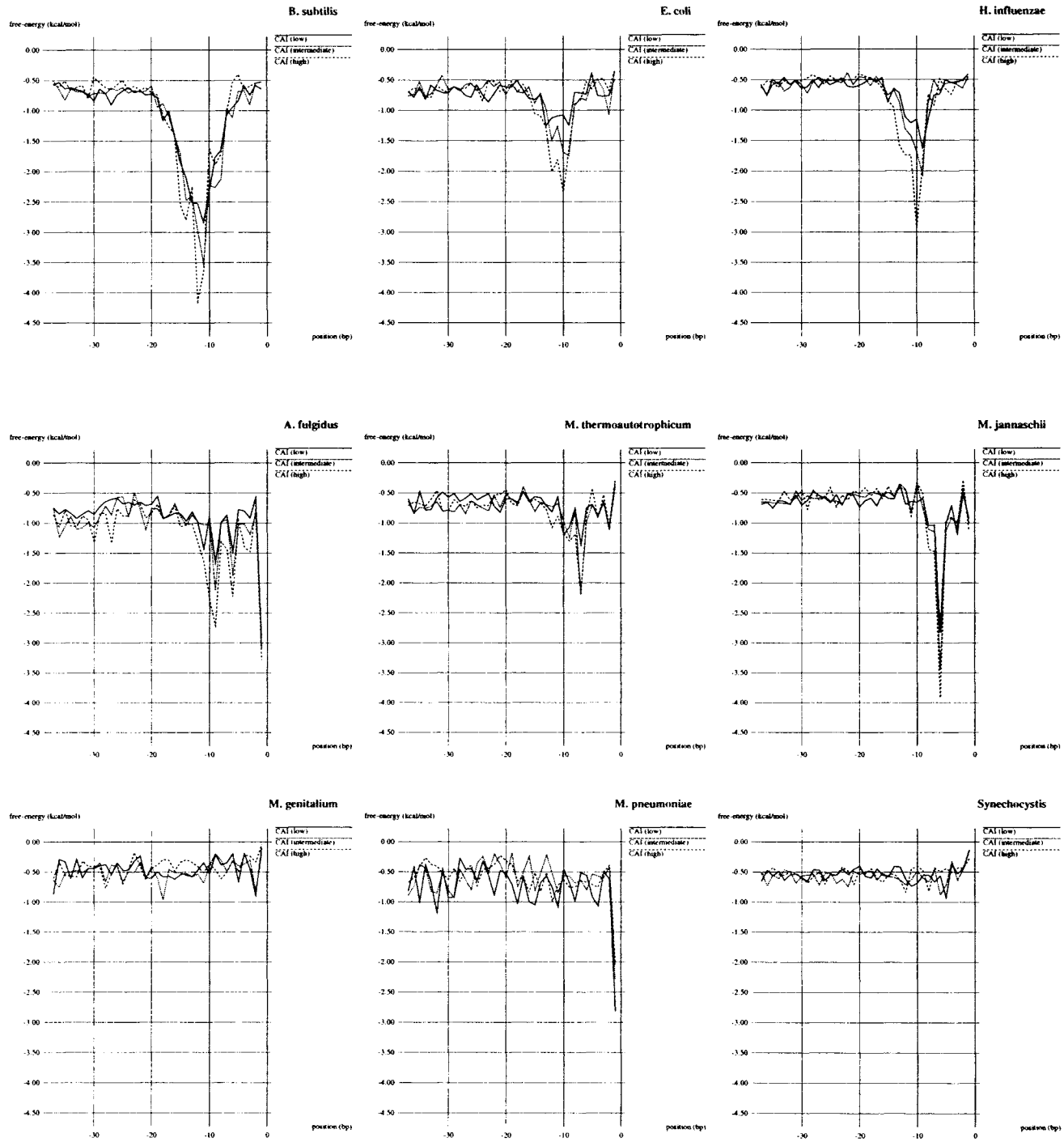


図 3.3:CAI 値別に見た自由エネルギーの平均値の変化　横軸は開始コドンからの距離を表す。

3.4.2 CAI値と自由エネルギーの最低値の関係

前の解析では上位、平均、下位の遺伝子について部分的に自由エネルギーの値を見たが、次に、全遺伝子について見たときにCAI値と自由エネルギーの間にどのような関係があるのかを調べた。まず開始コドンから-7bp～-17bpの間での自由エネルギーの最低値を各遺伝子について求め、CAI値に並べた後に任意の数ごとの平均を求めグラフにプロットした(*B.subtilis*、*E.coli*、*Synechocystis*については100遺伝子、*A.fulgidus*、*H.influenzae*、*M.thermoautotrophicum*、*M.jannaschii*は50遺伝子、*M.genitalium*、*M.pneumoniae*は15遺伝子)。また、データの信頼性を示すために、Spearman's rank correlation coefficient (r_s) を求めた。

図3.4にその結果を示す。*E.coli*、*B.subtilis*、*H.influenzae*の3種については自由エネルギーの最低値が徐々に低下していることが分かる(*E.coli* r_s=-0.88、*B.subtilis* r_s=-0.76、*H.influenzae* r_s=-0.75)。つまりCAI値が高くなるにつれてSD配列もより保存されているということであり、全遺伝子について見たときにも、コドン使用の偏りとSD配列の保存性との間に相関があることが確かめられた。3種の古細菌については、CAI値別に見た自由エネルギーでは違いが見られたものの、全体での相関はそれほど強くない(*A.fulgidus* r_s=-0.67、*M.thermoautotrophicum* r_s=-0.25、*M.jannaschii* r_s=-0.33)。2種の*Mycoplasma*菌と*Synechocystis*については前の解析と同様に相関は見られない。

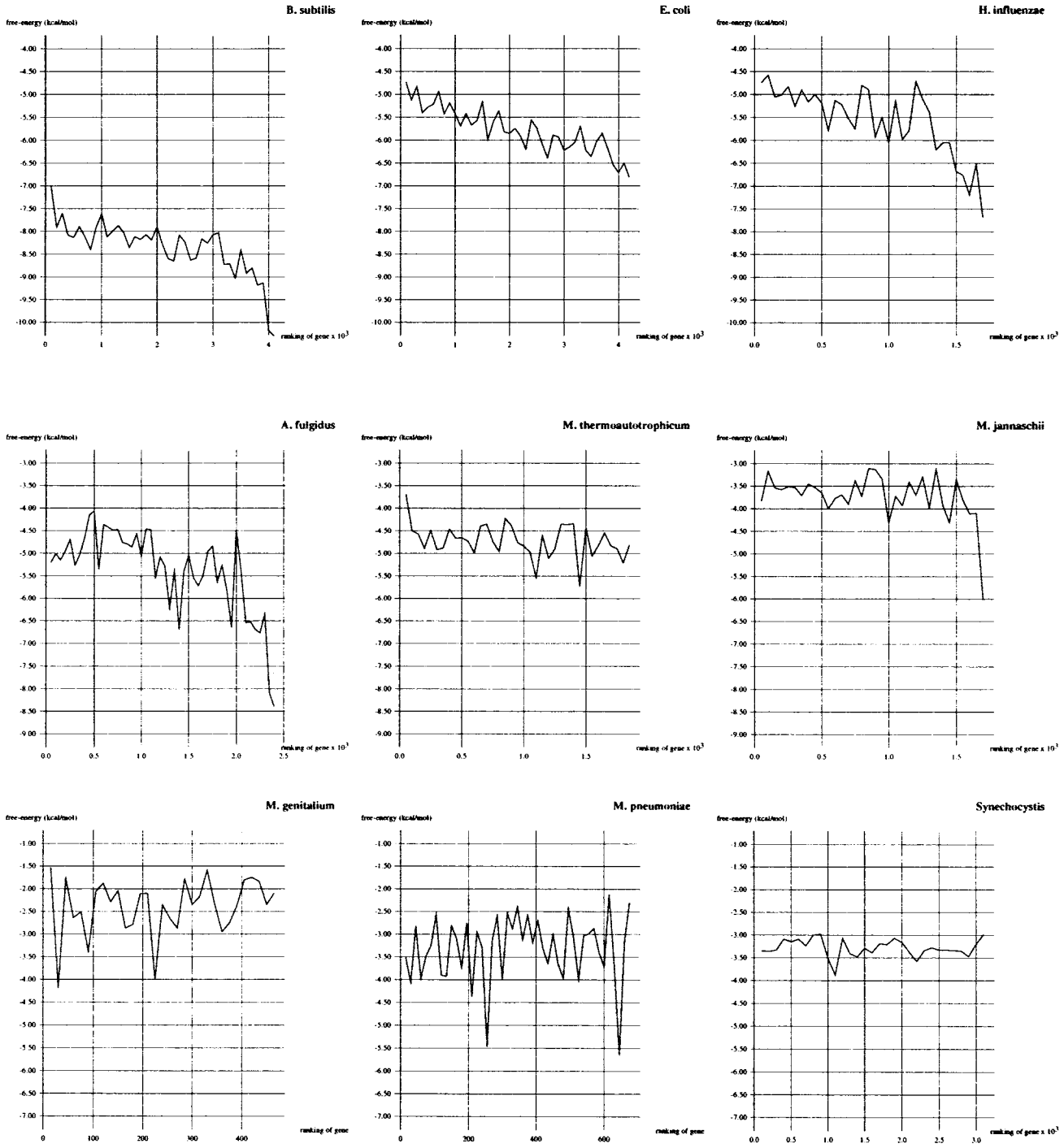


図 3.4:CAI 値別に見た自由エネルギーの最低値の変化　横軸は CAI 値のランキングを表し、左から右へ CAI 値が高くなっていく

3.5 考察

今回の解析結果から、多くの原核生物において、CAI値とSD配列の間に顕著な相関関係を見ることができた。CAI値は発現量と相関があるということが示されており[10]、SD配列の保存性というものが翻訳効率に関係しているという考えを支持するものである。頻繁に発現される遺伝子においては、発現量がそれほど高くない遺伝子に比べて、リボソームがSD配列をより効率的に認識する必要があるのではないかと考えられる。SD配列に変異が入ると、翻訳効率が著しく低下するという実験結果があることから[29]、今回の解析により、このことが多くの原核生物に当てはまるモデルであるということが示唆される。おそらく進化の過程で、高発現遺伝子は、高い翻訳効率を維持するために、同義語コドンの使用に偏りを持たせると共に、SD配列も効率のよい形に保存されてきたのだろう。

ただ、古細菌ではそれほど強い相関が見られなかったため、原核生物に普遍的な特性であると断言することはできない。古細菌の遺伝子発現系は真正細菌とは異なり、真核生物に見られる特徴も持っていることから、翻訳プロセスがSD配列の保存性にそれほど強く依存していないのではないかということが考えられる。

我々は、各ゲノムにおいて、各開始コドンがどのくらいの割合で使用されているのかということを解析した。先にも述べたように、原核生物は開始コドンとして大抵AUGコドンを使用するが、その他にUUGやCUGも開始コドンとして認識されることが知られている。しかし、AUG以外のコドンが開始コドンとして存在するとき、これを認識するtRNAの認識効率が低下するということが考えられる。

図3.5に、今回解析を行った9種の原核生物における、CAI値とAUGの使用頻度との関係を示す。*E.coli*、*B.subtilis*、*H.influenzae*と3種の古細菌については、CAI値が高い遺伝子群ほど、開始コドンとしてAUGを使用する割合が高くなっているのが分かる。つまり、発現量の高い遺伝子には、進化の過程でAUG以外の開始コドンへの変異が保存されにくかったということが考えられる。

*Synechocystis*と2種の*Mycoplasma*菌ではCAI値とAUGコドンの使用頻度にも相関は見られない。この3種の原核生物については、今回のどの解析においても傾向を見つけることが出来なかった。SD配列のような保存された配列が存在しないことから、リボソームがどのようにしてmRNAを認識しているのかという問題も含め、これらの原核生物の翻訳プロセスについては未知の領域が多く存在する。

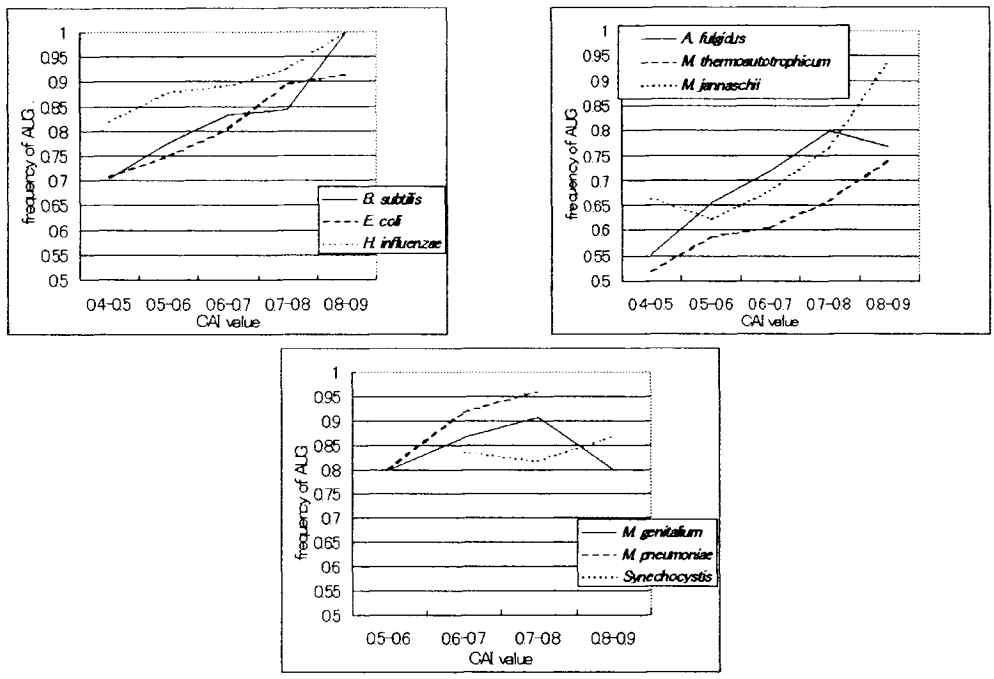


図 3.5：CAI 値と AUG の使用頻度の関係　横軸は、0.1 ずつ区切った CAI 値、縦軸はそれぞれの CAI 値のグループに含まれる遺伝子の AUG 開始コドンの使用頻度

第4章　脊椎動物におけるコドン使用の偏りと5' 非翻訳領域の保存性の関係

4.1　真核生物の翻訳開始機構

原核生物と真核生物の翻訳開始プロセスの違いは、まずリボソーム結合部位の認識の方法である。原核生物の場合は、開始コドン上流にあるSD配列で開始複合体が直接形成される。一方、真核生物では、40Sサブユニット (原核生物の30Sサブユニットに相当) がまずmRNAの5' 末端を認識して結合し、それからmRNA上を滑るように移動して (スキャニング) 開始部位に到達し、60Sサブユニット (原核生物の50Sサブユニットに相当) と会合する。真核生物の翻訳開始にも、原核生物のように開始因子というものが必要であり、原核生物よりもさらに多くの開始因子がそれぞれ適当な段階で機能している。

5' 末端に結合した40SサブユニットがmRNA上を移動しながらどのように開始部位を認識するのかについては、多くのmRNAで、5' 末端から最初のAUGコドンが開始コドンとして認識されていることが確かめられ [34]、さらに詳細な解析で、開始コドンの周辺にGCC(A/G)CCaugGというコンセンサス配列 (Kozakのコンセンサス配列) が存在することが明らかになった [34]。特に開始コドンから-3の位置にあるプリン塩基 (AorG) と開始コドン直後 (ポジション+4) のG残基が効率的な翻訳に重要で他のポジションに比べ強く保存される傾向があり、これらのポジションへの変異は、翻訳量を低下させ、leaky scanning(リボソームが開始コドンを認識できずに通りすぎてしまうこと) を起こしやすくするということも様々な実験で確かめられている [33, 35, 37, 38]。Kozakのコンセンサス配列は、脊椎動物以外の生物には当てはまらないという研究もあり、脊髄動物のmRNAに特有のものではないかということが言われている [39]。

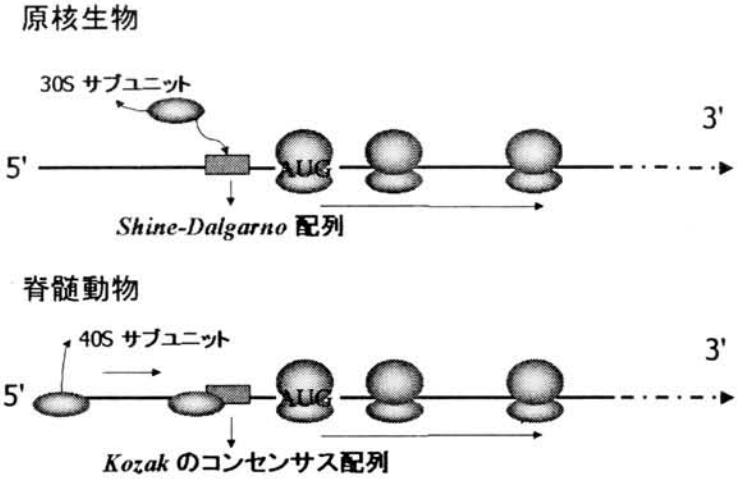


図 4.1：原核生物と脊髄動物の翻訳開始メカニズムの違い

4.2 解析に使用したデータ

マウスについては、原核生物のような全ゲノム配列データは存在しないが、理化学研究所によって完全長 cDNA ライブラリーが開発されており、今回の解析ではこのデータを使用した。

4.2.1 cDNA

cDNA とは、細胞から抽出された mRNA が逆転写酵素によって DNA コピーにつくり替えられたものである。この技術を適用するメリットは、理論上タンパク質コード領域を正確に得られるというところにある。

真核生物の遺伝子には、タンパク質をコードする領域 (エキソン) の他に、イントロンと呼ばれる、遺伝子領域内には存在しているにも関わらずアミノ酸配列をコードしていない領域が存在する。こうした非コード領域 DNA も、核内でいったんは一本の mRNA として転写されるが、その後核膜を通過して細胞質へ移動する前にスプライシングという特殊なプロセッシングを受けて mRNA から除去される。cDNA はスプライシングされた後の成熟した mRNA 配列から逆転写されるため、イントロンを含まず、タンパク質コード領域のみの状態で得ることができる。

cDNA は、基本的に細胞中の全 mRNA を対象にしていることから、細胞中で頻繁に転写されるような遺伝子の場合は同じ cDNA がいくつも合成されるということがある。本研究では、21076 本の cDNA からそうした重複配列を除去し、最終的に 8962 本の cDNA を対象に解析を行った。

4.3 方法

4.3.1 CAI 値

CAI 値については原核生物と同じ方法で算出した。その際、W 値を求めるために参照する発現量の高い遺伝子群については、GenBank から ribosomal protein、elongation factor、heatshock protein、outer membrane protein、RNA polymerase subunit、をコードする 52 の遺伝子を抽出した。表 4.1 に、マウスの W 値の一覧表を示した。

Amino acid	Codon	W value	Amino acid	Codon	W value	Amino acid	Codon	W value
Leu	UUA	0.0985	Ala	GCU	0.7864	Thr	ACU	0.5341
	UUG	0.3222		GCC	1.0000		ACC	1.0000
	CUU	0.3172		GCA	0.4707		ACA	0.5295
	CUC	0.4992		GCG	0.2231		ACG	0.3205
	CUA	0.1536	Val	GUU	0.3519	Ile	AUU	0.5939
	CUG	1.0000		GUC	0.4722		AUC	1.0000
Arg	CGU	0.4396		GUA	0.1744		AUA	0.1331
	CGC	1.0000		GUG	1.0000	Asn	AAU	0.6865
	CGA	0.6667	Gly	GGU	0.5203		AAC	1.0000
	CGG	0.9304		GGC	1.0000	Phe	UUU	0.6228
	AGA	0.6850		GGA	0.5319		UUC	1.0000
	AGG	0.6850		GGG	0.4584	Tyr	UAU	0.5968
Pro	CCU	0.8275	Ser	UCU	0.8651		UAC	1.0000
	CCC	1.0000		UCC	1.0000	Glu	GAA	0.5678
	CCA	0.8626		UCA	0.4706		GAG	1.0000
	CCG	0.2690		UCG	0.2457	Cys	UGU	0.8881
Gln	CAA	0.2335		AGU	0.4360		UGC	1.0000
	CAG	1.0000		AGC	0.8927	His	CAU	0.6883
Lys	AAA	0.4688					CAC	1.0000
	AAG	1.0000				Asp	GAU	0.8660
							GAC	1.0000

表 4.1：マウスの W 値一覧 (参照した遺伝子は GenBank から抽出した)

4.3.2 増加情報量

増加情報量は、複数の配列を参照して、各ポジションでのコンセンサスの強さを数値化する指標の一つであり、下記の式によって算出される。増加情報量が高いということは、その塩基ポジションで何らかの塩基の強いパターンが存在するということを意味している。

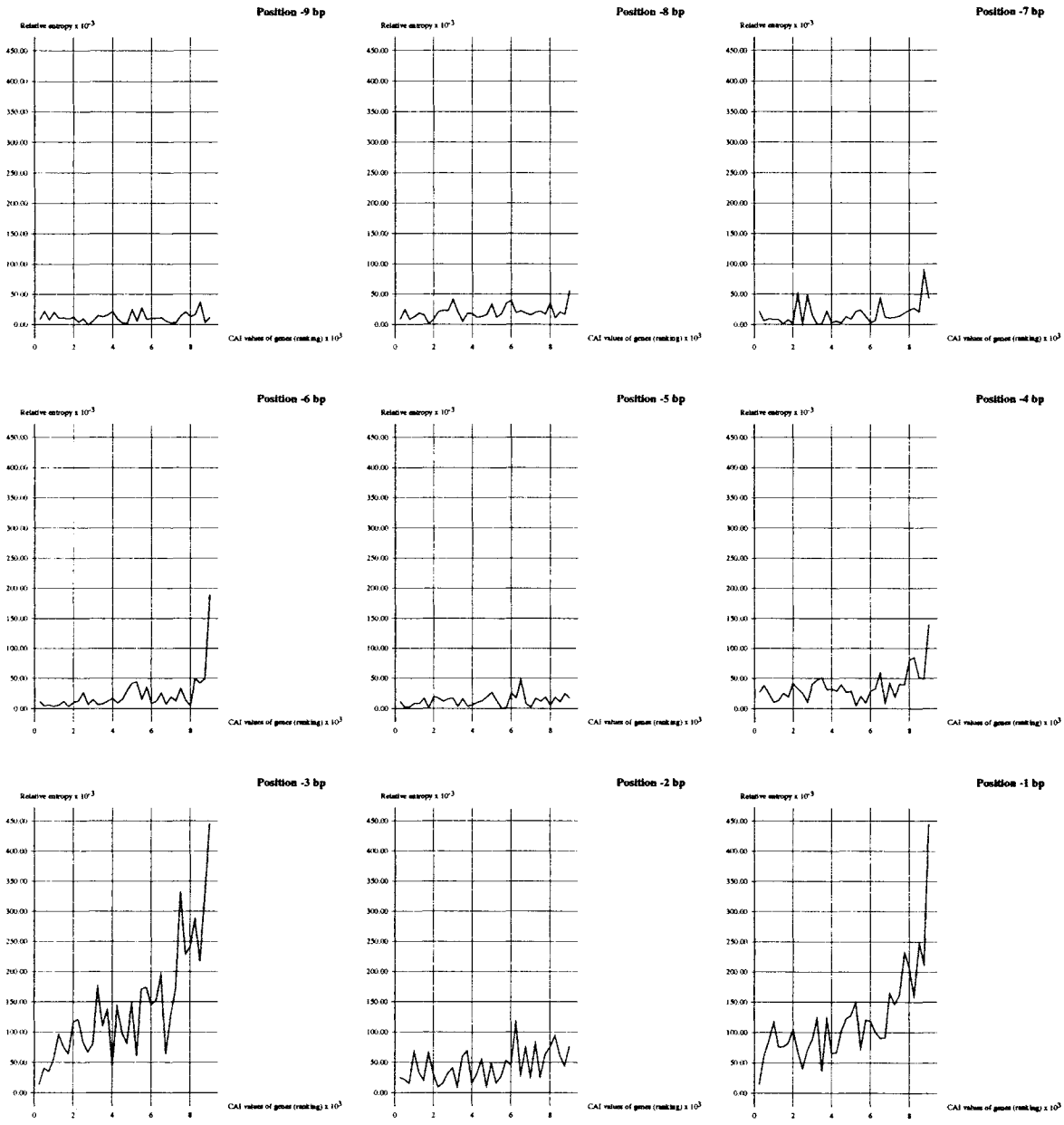
$$I(L) = \sum_{b=A,T,G,C} f(b,L) log_2 \frac{f(b,L)}{f(b)}$$

今回の解析では、まず対象とした約 9,000 本の cDNA について CAI 値を算出し、開始コドン周辺 (上流 20bp、下流 15bp) を抽出した。次に、周辺配列を CAI 値順に並べ、下から 250 遺伝子ずつの配列セットを作成し、各配列セットについて、各塩基ポジションの増加情報量を求め、ポジション別にグラフにプロットした。

4.4 結果

4.4.1 CAI 値と増加情報量の関係

マウスゲノムにおける、開始コドン周辺のコンセンサスの強さと CAI 値の相関を見た結果を以下に示す。


Position -9 bp
Position -8 bp
Position -7 bp
Position -6 bp
Position -5 bp
Position -4 bp
Position -3 bp
Position -2 bp
Position -1 bp
Relative entropy x 10-3
CAI values of genes (ranking) x 10³

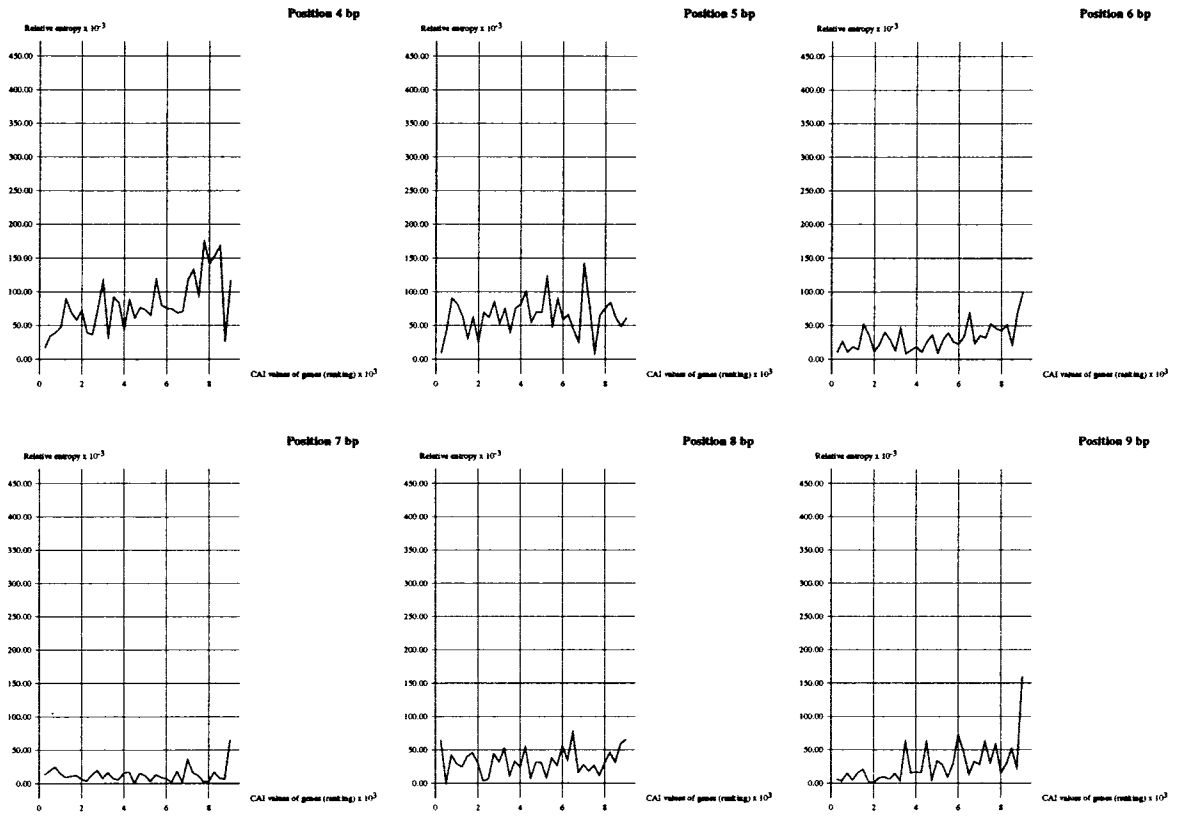


図 4.2:CAI 値と各ポジションの増加情報量の相関関係　横軸は CAI 値の低い順に 250 遺伝子ごとのセット、縦軸が増加情報量を表している。

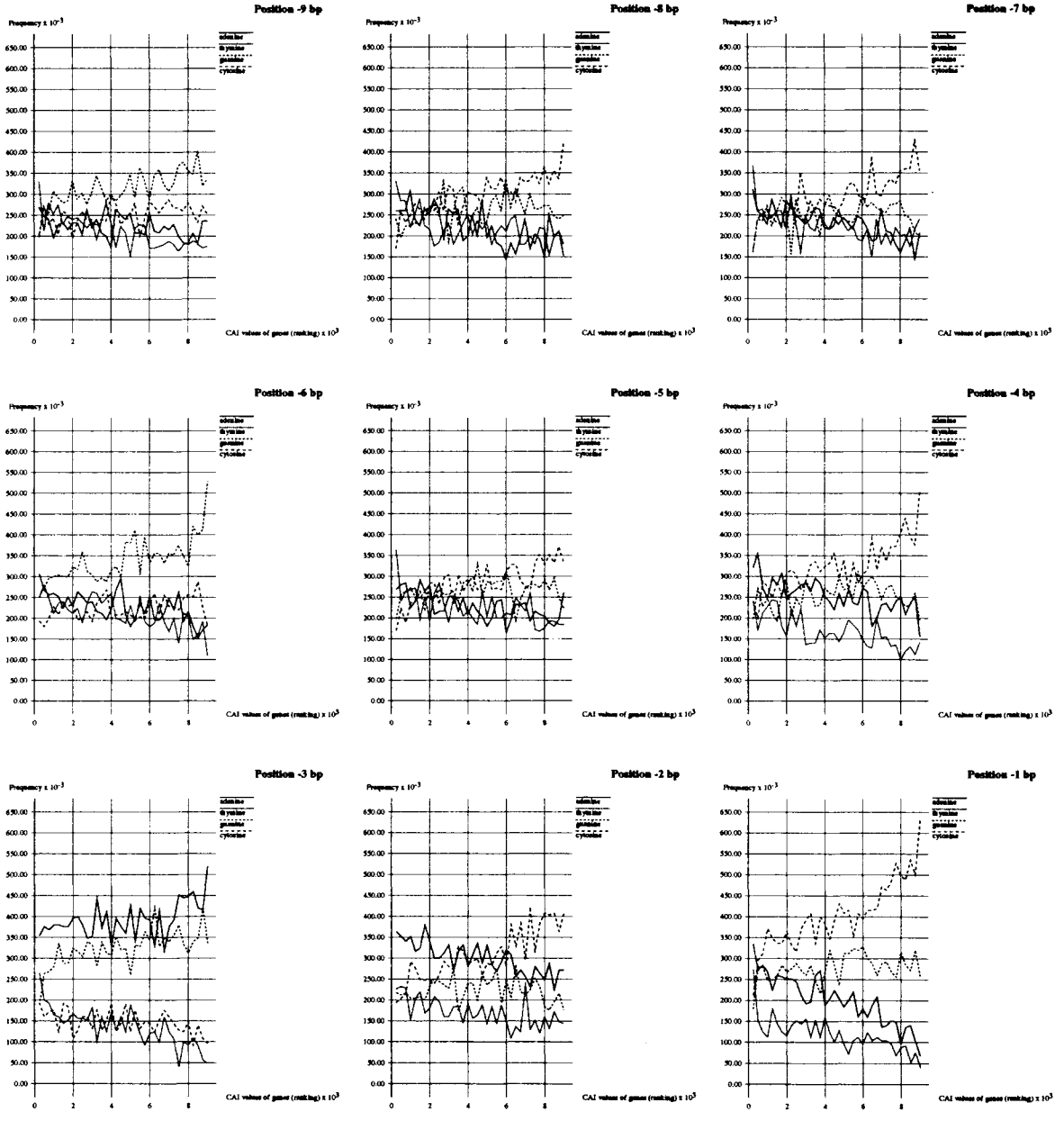
Kozak のコンセンサス配列 (GCCA/GCCatgG) のうち、翻訳開始に特に重要であると言われているポジション-3 と+4 の内、ポジション-3 では、CAI 値と増加情報量の間に顕著な相関関係が見られるのが分かる。さらに、ポジション-1 でも相関が見られ、ポジション -4、-6、-8 では相関関係は見られないものの、CAI 値の高い遺伝子群ではコンセンサスが強くなっていることが明らかになった。ポジション+4 では -3bp 程強い相関関係は見られなかった。

4.4.2 CAI 値と塩基含有量の関係

前回の解析で、開始コドン周辺のどの辺りにコンセンサスの強い領域が存在するのかということが明らかになった。次に、コンセンサスの強いポジションにおいて、実際にどういった塩基が好まれているのかを調べるために、各ポジションにおける A、T、G、C 各塩基の含有量の解析を行った。

図 4.3 にその結果を示す。CAI 値と増加情報量の間に顕著な相関関係が見られたポジション-1bp と-3bp では、-1bp でシトシン、-3bp でプリン塩基 (アデニン or グアニン) の含量が、CAI 値が高くなるのに伴って増加していることが分かる。さらに、- 1bp におけるシトシンと-3bp におけるプリン塩基は Kozak のコンセンサス配列の塩基組成と一致している。これら 2 つのポジションほど強い相関が見られなかったポジ

ション+4bp では、同様に Kozak のコンセンサス配列に対応するグアニンの含量が CAI 値と相関を示している。こういった傾向はポジション-2、-4、- 5、-6 でも見られる (ポジション-2bp、-4bp、-5bp はシトシン、-6bp はグアニン)。



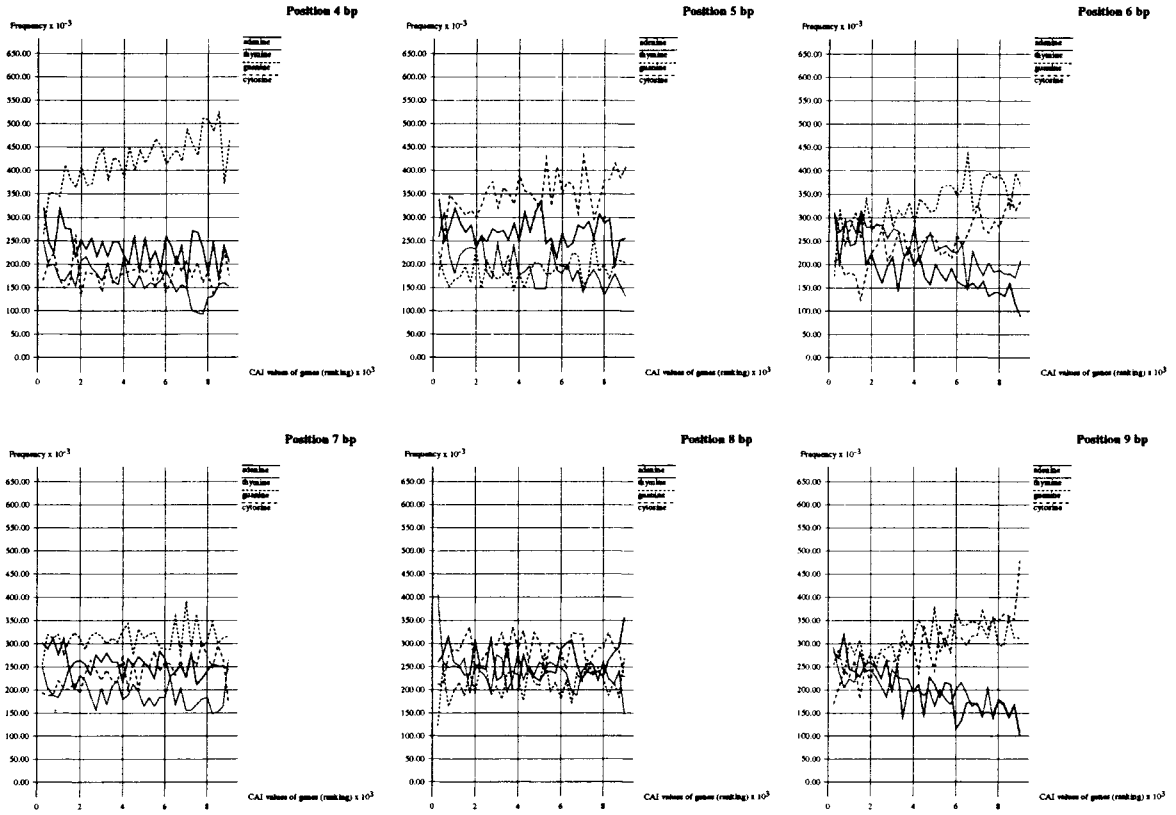


図 4.3:CAI 値と各ポジションの塩基含有量の相関関係　横軸は CAI 値の低い順に 250 遺伝子ごとのセット、縦軸が頻度を表している。

4.5 考察

真核生物においては SD 配列は存在しないが、脊髄動物で翻訳開始制御に重要だと考えられる Kozak のコンセンサス配列が見つかっていることから、この配列が 18S rRNA と相互作用するということが考えられる。実際 mRNA 上に 18SrRNA と対合し得る配列が保存されているという研究結果や [30]、実際に mRNA と rRNA が対合することがあるという実験結果が存在する [31]。しかし、18S rRNA 配列と 5'UTR の相補性が高いと、逆に発現量を低下させてしまうという結果が出ていることから、Kozak コンセンサス配列が 18S rRNA と実際に相互作用しているということはどうやら言えない。我々の解析結果は、Kozak のコンセンサス配列の保存性が高い程、翻訳には有利であるということを示しており、もし、18S rRNA の配列中に Kozak のコンセンサス配列と対合するような領域が含まれている場合は、Tranque 氏らの実験結果と相反するものになってしまうからである。

それでは Kozak のコンセンサス配列はリボソームの開始コドン認識にどのような形で寄与しているのだろうか。

CAI 値と増加情報量の解析では、ポジション-1、-3、+4 以外では目立った相関が見られなかったが、

Kozakコンセンサスに含まれる上流6ポジションでは、Kozakコンセンサスの塩基組成に相当する塩基の含量がCAI値が高くなるのに伴って増加していることが確認できた。さらに上流、-7、-8、-9ポジションでも、-7と-8ポジションではシトシンの含量、ポジション-9ではグアニンの含量に注目するとCAI値と正の相関を示していることが分かる。これらのポジションについては、Kozak氏の研究の中では、高等な真核生物において翻訳開始のコンセンサス配列となっているという風に述べられている[34]。今回は-10bpより上流については解析を行っていないが、さらに上流の塩基について調べて、今回の解析で見られたような傾向がどこまで波及しているのかということを明確にする必要がある。

今回の解析結果は、Kozakのコンセンサス配列がrRNAと相互作用するかどうかという問いに対する答えを提示するものではないが、少なくとも発現量と相関関係がありそうだということが明らかになった。このことは、真核生物においても、リボソームの開始コドン認識による翻訳開始の効率の良し悪しが、たんぱく質の生産量にも影響してくるということを示しており、コドン選択のパターンと同様に、翻訳効率を最適化するような圧力がかかった可能性というものも提示している。

第5章　最後に

本研究において明らかになったことをポイントをしぼってまとめる。

- 原核生物において、SD配列は、CAI値が高くなるにつれてより強く保存される傾向がある
- CAI値の高い遺伝子は、開始コドンとしてAUGを使用する傾向が強い
- 脊髄動物において、CAI値の高い遺伝子程、開始コドン周辺にKozakのコンセンサス配列がより強く保存される傾向がある

図5.1に、原核生物、脊髄動物における発現量の高い遺伝子の翻訳プロセスのモデルを示す。まず翻訳開始のステップでは、原核生物ではSD配列、脊髄動物ではKozakのコンセンサス配列が存在し、発現量の高い遺伝子においてはこれらの配列は特に強く保存されているために、リボソームによるmRNA、開始コドンの認識がスムーズに行われる。ここで、原核生物では、開始コドンとしてAUG以外のコドンが使われないようになっている。ほとんどの遺伝子において開始コドンはAUGであるので、AUGを使うことが翻訳の効率を上げる積極的な要因になっているというよりは、スムーズな翻訳開始を妨げるようなAUG以外のコドンの使用が進化の過程で避けられてきたと考える方が自然である。

タンパク質コード領域内においては、対応するtRNA分子量の多いコドンが多用されることにより、リボソームによるアミノ酸伸長が効率良く行われていく。一本のmRNAはリボソームが複数結合したポリリボソーム(ポリソーム)と呼ばれる単位になっている。mRNA中にマイナーコドン(対応するtRNAの分子量の少ないコドン)が存在すると、そこでリボソームがアミノ酸伸長に手間取り(ribosomal pausing)、後ろからやってくるリボソームがつまり、翻訳がスムーズに流れなくなるということが知られている[40, 41, 42]。このことから、翻訳開始からの一連の流れがスムーズにいくことは、リボソームが次々とmRNAに結合して翻訳を行っていくメカニズムにも適していると言える。

タンパク質のコード領域の最後には必ず終止コドン(UAA、UAG、UGA)が存在する。このコドンに対応するtRNAは存在せず、タンパク因子によって直接認識される。この因子は遊離因子(RF;release factor)と呼ばれ、リボソームが翻訳を終了するのに関与しており、最近の研究で、終止コドン周辺の塩基パターンにも、同義語コドン使用の偏りと相関が見られることが分かり、発現量の高い遺伝子には、翻訳をスムーズに終了させるようなコンセンサス配列が存在するのではないかと言われている[43]。

同義語コドン使用の偏りについては、翻訳機構との関連のみならず、様々な視点から研究が行われており、翻訳効率との関係だけでこれを説明することは出来ない。Romeoらは、*Chlamydia trachomatis*のコドン頻度の解析を行った結果、この細菌のコドン使用のパターン形成が最低4つの要因による結果であるということを発表した[44]。まずコドン使用のパターン形成に最も影響を与えると言われるのがDNA鎖特異的な変異圧である。DNAの二重鎖は、それぞれ複製メカニズムの違いからリーディング鎖(leading strand)とラギング鎖(lagging strand)と呼ばれる。二重鎖の間では、鎖に入る変異圧に違いがあり、リー

ディング鎖にはグアニンとチミン、ラギング鎖にはシトシンとアデニンが偏るということが知られている。多くのアミノ酸で、コドンの三文字目は自由であるケースが多いことから、ゲノム全体でそれぞれのDNA鎖のコドン頻度を調べてみると、リーディング鎖では三文字目がGあるいはT、ラギング鎖ではCあるいはAであるコドンが統計的に顕著に偏っていることが明らかにされた。次に、複製、翻訳レベルでの自然選択である。リーディング鎖で同義語コドン使用の偏りについて解析した結果、発現量の高い遺伝子では、三文字目がCあるいはAであるコドンが多用されるケースも見られることが示された。第三にコドン使用のパターン形成に関わっているとされるのが遺伝子のヒドロパシー(疎水性親水性指標)である。親水性の遺伝子と疎水性の遺伝子を調べたところ、各グループで顕著に偏るコドンが多く存在することが分かり、さらにその内、親水性遺伝子で好まれるコドンの三文字目はAあるいはT、疎水性遺伝子で好まれるコドンの三文字目はCあるいはGである傾向が見られた。最後にアミノ酸の保存度である。*C. trachomatis*と近縁種にある*Chlamydia pneumoniae*との間で、相同な遺伝子の間でコドン頻度を調べたところ、アミノ酸が保存されているとコドンの三文字目も保存される傾向が強いということが明らかにされた。このように、コドン使用のパターン形成は複合的な要因によって形作られているということが分かる。同義語コドン使用の偏りと発現量の関係というポイントに焦点を当てたときに、大腸菌ゲノムに他生物由来の遺伝子を組み込んだときに、そこにマイナーコドンが含まれていた場合、発現量が低下するかというと、必ずしもそうではないという説もある。現実的には、発現量を積極的に制御するのであれば、翻訳段階よりも転写段階の方が効果的であると考えられるが、転写量は細胞内の状態や成長率などの外的要因により大きく増減するものであるため、転写量だけで発現量との相関を示すのも困難である。実際、*Saccharomyces cerevisiae*において、mRNA量とタンパク質量との間には十分な相関関係が見いだせないという報告もある[45]。

発現量の制御メカニズムを明らかにしようとするならば、同義語コドン使用の偏りの他に、転写量から、tRNA、翻訳開始因子、リボソームの分子量までさまざまな要素を複合的に捉える必要があるが、そういった要素は細胞の状態によっても多用に変化していくので、一義的に相互関係を理解するのは極めて困難である。

今回の一連の解析により、同義語コドン使用の偏りが翻訳機構の中でどのような位置付けにあるかということについては、翻訳効率をオペレートするものであり、発現量の高い遺伝子において、進化の過程で翻訳に有利な変異が蓄積した結果であるとの一つの解釈を提示することが出来た。

同義語コドン使用の偏りの生成メカニズムについて考えたとき、*E.coli*のように、tRNAの遺伝子数、分子数と密接に関係のある種については、コドン使用とtRNAの共進化[32]ということが言われており、分子数の多いtRNAに読まれるコドンが多用されるということは既に述べた通りである[2, 3]。それではなぜ種によって異なるtRNA遺伝子が重複してそれに合った同義語コドン使用の偏りが形成されたのかということについては明らかになっていない。今後原核生物や真核生物のゲノム配列が明らかになっていく中で、ゲノム同士で比較を行うことで、重複しているtRNA遺伝子になんらかの共通点が見い出せる可能性があると考えている。また、ゲノム中での遺伝子の位置であるとか、ゲノムのG+C含有量を網羅的に調べることにより、同義語コドン使用の偏りを体系的に理解するための手がかりを得ることができるのではないかと考えている。

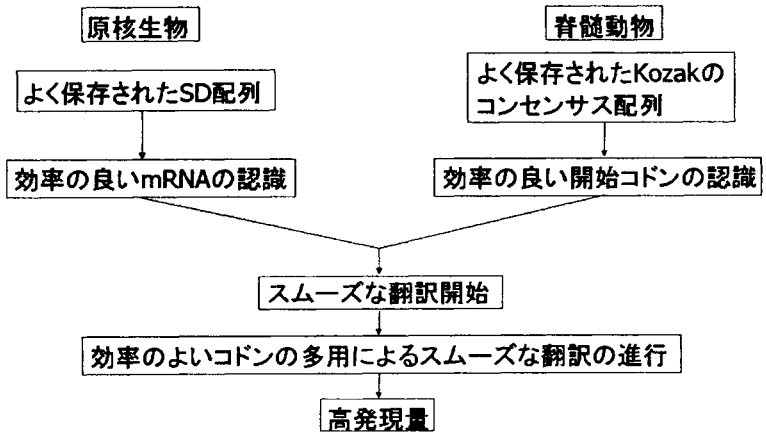


図 5.1：高発現遺伝子の翻訳プロセスのモデル

謝辞

本研究を進める上で、慶應義塾大学環境情報学部の冨田勝教授、奈良先端科学技術大学院大学の森浩禎教授、慶応義塾大学大学院政策・メディア研究科の鷲尾尊規氏、理化学研究所ゲノム科学総合研究センターの斎藤輪太郎氏には、貴重な御意見、アドバイスをいただきました。そして、理化学研究所ゲノム科学総合研究センター遺伝子構造・機能研究グループプロジェクトディレクターの林崎良英氏には、マウスゲノムの解析を行う上でcDNAデータを提供していただいた他有益な助言をいただきました。また、原核生物におけるSD配列の自由エネルギーによる解析では、同じく慶應義塾大学大学院政策・メディア研究科の長田木綿子氏に方法論について御意見をいただきました。その他、プログラミングやデータの整理では、慶應義塾大学環境情報学部の大熊祐介氏、大野浩氏、小澤陽介氏、花岡悟史氏、同大学総合政策学部の今村千秋氏に助力を賜わりました。本研究をサポートしてくださいました全ての方々に、この場をお借りして厚くお礼申し上げます。

関連図書

[1] Grantham R et al. (1980) Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8:r49-r62

[2] Ikemura T (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. 146:1-21

[3] Ikemura T (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. J. Mol. Biol. 158:573-597

[4] Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2:13-34

[5] Kanaya S et al. (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238:143-155

[6] Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10:7055-7074

[7] Grantham R et al. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9:r43-74

[8] Grosjean H, Fiers W (1982) Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene 18:199-209

[9] Sharp PM et al. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14:5125-5143

[10] Sharp PM, Devine KM (1989) Codon usage and gene expression level in Dictyostelium discoideum: highly expressed genes do 'prefer' optimal codons. Nucleic Acids Res. 17:5029-5039

[11] Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897-907

[12] Wright F and Bibb MJ (1992) Codon usage inthe Grich Streptomyces genome. Gene 113:55-65

[13] Ohama T, Muto A, and Osawa S (1990) Role of GC-biased mutation pressure on synonymous codon choice in Micrococcus luteus, a bacterium with a high genomic GC-content. Nucleic Acids Res. 18:1565-1569

[14] Sharp PM et al. (1993) Codon usage: mutational bias, translational selection, or both: Biochem. Soc. Trans. 21:835-841

[15] Sorensen MA, Kurland CG, Pedersen S (1989) Codon usage determines translation rate in Escherichia coli. J. Mol. Biol. 207:365-377

[16] Klenk HP et al. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. Nature 390:364-370

[17] Kunst F et al. (1997) The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature 390:249-256

[18] Blattner FR et al. (1997) The complete genome sequence of Escherichia coli K-12. Science 277:1453-1474

[19] Fleischmann RD et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496-512

[20] Smith DR et al. (1997) Complete genome sequence of Methanobacterium thermoautotrophicum delta H: functional analysis and comparative genomics. J. Bacteriol. 179:7135-7155

[21] Bult CJ et al. (1996) Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. Science 273:1058-1073

[22] Fraser CM et al. (1995) The minimal gene complement of Mycoplasma genitalium. Science 270:397-403

[23] Himmelreich R et al. (1996) Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic Acids Res. 24:4420-4449

[24] Kaneko T et al. (1996) Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3:109-136

[25] Sharp PM, Li WH (1987) The codon adaptation index –a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281-1295

[26] Tompa M (1999) An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. Proc. Seventh International Conference on Intelligent System for Molecular Biology, 262-271

[27] Osada Y et al. (1999) Analysis of base-pairing potentials between 16S rRNA and 5'UTR for translation initiation in various prokaryotes. Bioinformatics 15:578-581

[28] Levin B (1997) GENES VI, Oxford Univ. Press and Cell Press, pp. 104

[29] Barrick D et al. (1994) Quantitative analysis of ribosome binding sites in E.coli. Nucleic Acids Res. 22:1287-1295

[30] Tranque P et al. (1998) rRNA complementarity within mRNAs: A possible basis for mRNA-ribosome interactions and translational control. Proc. Natl. Acad. Sci. USA 95:12238-12243

[31] Hu MC et al. (1998) rRNA-complementarity in the 5' untranslated region of mRNA specifying the Gtx homeodomain protein: evidence that base- pairing to 18S rRNA affects translational efficiency. Proc. Natl. Acad. Sci. USA 96:1339-1344

[32] Bulmer M (1987) Coevolution of codon usage and transfer RNA abundance. Nature 325:728-730

[33] Kozak M (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44:283-292

[34] Kozak M (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNA. Nucleic Acids Res. 15:8125-8148

[35] Kozak M (1989) Context effects and inefficient initiation at non-AUG codons in eukaryotic cell-free translation systems. Mol. Cell. Biol. 9:5073-5080

[36] Kozak M (1997) Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. EMBO J. 16:2482-2492

[37] Handley-Gearhart PM et al. (1994) Human ubiquitin-activating enzyme, E1. Indication of potential nuclear and cytoplasmic subpopulations using epitope-tagged cDNA constructs. J. Biol. Chem. 269:33171-33178

[38] Shire D et al. (1995) An amino-terminal variant of the central cannabinoid receptor resulting from alternative splicing. J. Biol. Chem. 270:3726-3731

[39] Cavener DR, Ray SC (1991) Eukaryotic start and stop translation sites. Nucleic Acids Res. 19:3185-3192

[40] Chen GT and Inoue M (1990) Suppression of the negative effect of minor arginine codons on gene expression;preferential usage of minor codons within the first 25codons of the Escherichia coli genes.Nucleic Acids Res. 18:1465-1473

[41] Chen GT and Inoue M (1994) Role of the AGA/AGG codons,the rarest codons in global gene expression in Escherichia coli. Genes Dev. 8:2641-2652

[42] Vervoort ET et al.(2000) Optimizing heterologous expression in Dictyostelium :importance of 5' codon adaptation. Nucleic Acids Res. 28:2069-2074

[43] Ozawa Y et al. Comprehensive sequence analysis of translation termination sites in various eucaryotes. in submission

[44] Romero H et al. (2000) Codon usage in Chlamydia trachomatis is the result of strand-specific mutational biases and a complex pattern of selective forces. Nucleic Acids Res. 28:2084-2090

[45] Gygi SP et al. (1999) Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. 19:1720-1730

発表

学会発表

- "Fop values, Start codons, SD sequence conservation, and their correlation to gene expression level." Sakai H, Imamura C, Ohno H, Washio T, and Tomita M; Third Annual Conference On Computational Genomics, 1999 年 11 月, Baltimore
- "Fop 値、開始コドン、SD 配列の保存性と遺伝子発現との関係" 坂井寛章、今村千秋、大野浩、鷲尾尊規、冨田勝; 第 22 回日本分子生物学会年会, 1999 年 12 月, 福岡
- "Fopvalues, Start codons, SD sequence conservation, and their correlation to gene expression level." Sakai H, Imamura C, Ohno H, Washio T, and Tomita M; GIW'99, 1999 年 12 月, 東京
- "Correlation between codon usage bias and conservation of 5'untranslated region." Sakai H, Ohkuma Y, Imamura C, Shinagawa A, Itoh M, Shibata K, Carninci P, Konno H, Kawai J, Fukunishi Y, Hayashizaki Y, and Tomita M; The 14th International Mouse Genome Conference, 2000 年 11 月, 成田
- "5' 非翻訳領域の配列保存性とコドン使用の偏りとの相関について" 坂井寛章、大熊裕介、今村千秋、品川朗、伊藤昌可、柴田一浩、Carninci Piero、今野英明、河合純、福西快文、林崎良英、冨田勝; 第 23 回日本分子生物学会年会, 2000 年 12 月, 神戸
- "Correlation between sequence conservation of 5'UTR and codon usage bias." Sakai H, Ohkuma Y, Imamura C, Shinagawa A, Itoh M, Shibata K, Carninci P, Konno H, Kawai J, Fukunishi Y, Hayashizaki Y, and Tomita M; GIW2000, 2000 年 12 月, 東京

論文発表

- Sakai H, Imamura C, Osada Y, Saito R, Washio T, and Tomita M. (2001) Correlation between Shine-Dalgarno sequence conservation and codon usage of bacterial genes. Journal of Molecular Evolution. in press.
- Sakai H, Washio T, Saito R, Shinagawa A, Itoh M, Shibata K, Carninci P, Konno H, Kawai J, Hayashizaki Y, and Tomita M. Correlation between sequence conservation of the 5' untranslated region and codon usge bias in Mus muscuus genes. in submission.

同義語コドン使用の偏りと5'非翻訳領域の保存性の関係のコンピュータ解析

2001年3月30日　　初版発行

著者　坂井寛章

監修　冨田勝

発行　　湘南藤沢学会

〒252-0816　神奈川県藤沢市遠藤5322

TEL:0466-49-3437

Printed in Japan　　印刷・製本　　ワキプリントピア

SFC-MT2000-002

■本論文は修士論文として優秀と認められ、出版されたものです。