東京大学大学院新領域創成科学研究科
情報生命科学専攻

平成 17 年度

修士論文

コムギ胚芽無細胞タンパク質合成系における翻訳効率決定要因の配列解析

2006 年 3 月提出
指導教員　高木　利久　教授

46918　藤田　直也

目次

要旨

ポストゲノム時代と言われる今，タンパク質の立体構造解析や創薬のスクリーニングなど大規模研究において，活性を持ったタンパク質を十分量用意することが必要不可欠になっている．そのための方法として無細胞合成系と呼ばれる *in vitro* のタンパク質合成系が注目されている．中でも，コムギ胚芽無細胞タンパク質合成系では，ハイスループットに転写，翻訳，精製するシステムが構築され，数百種類のタンパク質の大量合成が可能であり，プロテオミクスの基盤技術になることが期待されている．このコムギの系における技術改良の 1 つに，翻訳に重要である 5'-UTR を翻訳効率の高い配列に統一した点が挙げられる．よって，自動化で反応条件がほぼ等しいことと 5'-UTR が共通であることから，どのようなタンパク質でも合成量が同じぐらいになることが予想される．しかしながら，報告されている合成量には 20 倍以上のばらつきがある．このばらつき程度は実験誤差だけでは説明できず，合成量は各タンパク質に依存して決定されていると考えられる．つまり，合成量に影響を与える因子は，アミノ酸配列とそれをコードする核酸配列の特徴に由来する可能性が高い．そこで本研究では，合成量のばらつきを配列の特徴から説明することができるか試みた．

具体的には，コムギ胚芽無細胞合成系で合成した，シロイヌナズナのタンパク質リン酸化酵素 423 個を対象に，配列に基づく様々な特徴を抽出し，合成量と相関があるかどうかを解析した．その結果，翻訳効率に影響を与える因子が 4 つ同定された．(i) 核酸

配列に基づく特徴からは，開始コドン付近の塩基対形成が合成量低下に影響を与えることが示唆された．開始コドンの上流 25nt.～下流 25nt.内での二次構造を予測し，自由エネルギーを求めると，合成量と正の相関が観察された．したがって，その領域で塩基対が多い mRNA は翻訳しにくいことが示唆された．(ii) アミノ酸配列に基づく特徴に関して，アミノ酸インデックスデータベース(AAindex)を使用し，様々な特徴量を調べた．AAindex とは，20 個のアミノ酸を 20 個の数値に変換する指標のデータベースで，現在 516 もの指標が登録されている．解析の結果，油谷らの点変異実験に基づくタンパク質の熱力学的安定性の指標が，合成量に影響を与えることが示唆された．N 末側領域 11aa.～200aa.での熱力学的安定性の指標の総和と，合成量との間には正の相関があった．このことから，熱力学的に安定なタンパク質は合成量が増加し，不安定なものは低下する傾向にあると考えられる．(iii) 他にも，N 末側領域内で disorder 領域の割合が合成量に影響を与えることが観察された．disorder 領域とは特定の構造をとらない領域のことである．N 末側領域 20aa.～206aa.の予測 disorder 残基の割合と合成量には負の相関があった．つまり，N 末側領域がふらふらしているタンパク質は合成量の低下を招くという示唆が得られた．(ii), (iii)よりタンパク質リン酸化酵素では，N 末側領域の特徴が特に合成量に影響を与えている可能性が考えられる．ドメイン構造からの相対的な位置ではなく，末端からの絶対的な位置であることが興味深い．(iv) さらに別のアミノ酸配列由来の特徴として，coiled-coil 構造を持つと予測されたタンパク質は合成量が低下する傾向

にあった．また，従来言われている，コドン使用頻度はこのデータセットに関しては影響がないことを確認した．

最後にこれら因子の組み合わせで，合成量の多寡をどの程度説明できるか検証した．4 因子を説明変数，合成量を目的変数として回帰木を構築した．その結果，合成量が低いタンパク質の約半数がこれらの因子で説明可能であった．加えて，合成量が高いタンパク質に関するルールも発見した．翻訳効率の低下に関係する因子群の否定集合となるタンパク質は，合成量が高くなると判別された．

最近，転写因子 661 個の合成量を計測したデータも入手した．リン酸化酵素に対して同定された，塩基対形成に関する傾向はこのデータでも観察されたが，他の 3 因子との相関は見られなかった．あらゆるタンパク質に対して，一般性のある因子の抽出は難しく，タンパク質ファミリーで個別の要因を探索する必要があるのかもしれない．今後，AAindex 由来の安定性以外の指標や，核酸配列由来の他の特徴から，さらなる因子の探索を行う予定である．また，*in silico* で同定した翻訳効率決定因子の有効性を実験的に検証し，無細胞合成系の技術改良に応用していくことを目指す．

1. 序論

1.1. 無細胞タンパク質合成系

ポストゲノム時代の今，ゲノム配列は数多くの生物種で解読された．しかし，機能未知の遺伝子は未だに多い．そのため，遺伝子の翻訳産物であるタンパク質の構造，機能同定は，ゲノム解読に続く，差し迫った課題である．しかしながら，対象となるタンパク質を十分量得ることが難しく，ボトルネックであった．よって，ハイスループットなタンパク質合成技術が，構造，機能プロテオミクスで渇望されている(Yokoyama 2003)．

人工的にタンパク質を合成する方法は，化学合成系，*in vivo* の系，そして *in vitro* の系と，大きく 3 種類に分けられる．その中で，*in vitro* の系である無細胞タンパク質合成系と呼ばれるシステムが，ハイスループットな合成系として注目されている(Spirin 2004)．以下に 3 手法の特徴を比較しながら，無細胞タンパク質合成系がハイスループットな合成に適している点を述べる．

まず，化学合成系とはペプチド結合でポリペプチドを生産する手法である(Blaschke *et al.* 2000)．生物学的に有用なタンパク質は概ね数百残基であるが，化学合成系では結合可能な長さは 70 残基程度しかない．よって，解析対象のタンパク質を合成すること自体が，そもそも困難である場合が多い．

次に，*in vivo* の系として遺伝子工学的なタンパク質発現系がある．この系は，十分な合成量を誇り，大腸菌や酵母を宿主として，現在よく用いられている．しかしながら，

宿主にとって有害なタンパク質は，合成することができないという制限があり，網羅性に限界がある(Chrunyk *et al.* 1993)．また，細胞への遺伝子の組み込み，それらの培養など合成過程が煩雑で，大規模な合成には人手と日数を要する．よって，ハイスループットに生産するためには，それら問題点を克服する必要がある．

最後に，*in vitro* の系，無細胞タンパク質合成系は，試験管内に細胞抽出液，アミノ酸やATPなどの基質，そして合成したいタンパク質のmRNAを入れ，タンパク質合成を行うものである．この方法自体は古く，遺伝暗号解読の際にはすでに用いられていた(Lengyel *et al.* 1969)．昔は翻訳システムの安定性が低く，生化学，構造生物学的な解析ができるほどの合成効率ではなかったが，その後，合成反応の持続時間を長くする改良が行われた(Spirin *et al.* 1988)．その結果，生産量は *in vivo* の系に匹敵するぐらいまで改善され，大量合成システムとしての可能性が見出された．近年は，実際に数百のタンパク質を一度に合成することができる自動機械が構築され，ハイスループットな生産が可能になっている(Sawasaki *et al.* 2004)．

以上のような特徴から，化学合成や *in vivo* の系に比べて，無細胞タンパク質は手続きの自動化，そしてハイスループットな生産が可能なところに利点があり，大量タンパク質合成技術として期待されている．

無細胞タンパク質合成系は，さらに，細胞抽出液の由来となる生物種に従い細分化される．主な提供源としては，大腸菌(Kim *et al.* 1996)，ウサギ網状赤血球(Ryabova *et al.*

1989)，コムギ胚芽(Sawasaki *et al.* 2002)の 3 つがある．その中で，コムギの系は次のような特徴，利点がある．コムギ胚芽内には，発芽の時期を前にして，翻訳に必要な全ての要素が濃縮された状態で蓄えられている．そのため，翻訳因子の供給源として使用することは大変都合が良い．コムギは食用としても身近なもので，生物汚染の心配もなく，簡便かつ多量に入手できることも利点である．また，コムギの合成系では真核，原核生物を問わずタンパク質のフォールディングが正しく行われていることが確認されており，その点でも優れている．対照的に，原核生物である大腸菌の系では，真核生物のタンパク質に対し，正しくフォールディングが行われない可能性がある．

上記のように有用なコムギ胚芽由来の系を，ハイスループットな合成システムにするために，様々な技術改良が行われてきた．反応環境の改良として，まず，強い洗浄による胚乳の除去が挙げられる．胚乳部分にはプロテアーゼ，リボヌクレアーゼ等，生合成阻害剤となるものが含まれている．そのため，できるだけ胚乳が混入しないことで，合成効率はかなり改善された(Madin *et al.* 2000)．他には細胞抽出液と，tRNA，ATP などの基質が入った溶液を，二層に分離した重層法という方法が開発された．二層に分離することで，基質を長時間ゆっくりと安定的に供給することが可能になり，合成反応時間が大幅に延長された(Sawasaki *et al.* 2002)．

また，mRNA 配列における改良として，全て遺伝子に対する共通の 5'-, 3'-の非翻訳領域(UTR)付加が挙げられる．真核生物の転写物に必須な 5'-^{7}mGpppG キャップと

poly(A)-tail の付加がなくとも，その条件に匹敵するぐらいの合成量が得られる配列が発見された．特に 5'-UTR は翻訳の開始に重要な領域であるため，タバコモザイクウィルス由来の翻訳促進配列であるΩ配列(Gallie 2002)に置き換え，統一された(Kamura *et al.* 2005)．この改良で，キャップ構造の付加等の化学修飾の工程を省略でき，かつ，5'-UTR を揃えることで，合成量をある程度一定にすることが可能になった．

以上のような改良を基に，数百種類ものタンパク質を一度に合成することができる，ハイスループットな生合成機械が開発された(Sawasaki *et al.* 2004)．これを用いると，合成したいタンパク質の cDNA クローンを用意すれば，転写，翻訳，精製の工程を全自動で行うことができるため，とても簡便である．今後は，プロテオミクスの基盤技術として用いられる機会もよりいっそう増え，コムギ胚芽無細胞タンパク質合成系の需要はますます高まると思われる．

1.2. 本研究の目的

コムギ胚芽無細胞合成系は上で述べたように，翻訳に重要な 5'-UTR が統一されていることと，エネルギーや tRNA の十分な供給，そして機械化により反応条件が均質であるということから，どのようなタンパク質でも同等量の合成が行われると期待される．しかしながら，実際の合成量は 0.1 から 2.3mg/ml とばらつきがある(Sawasaki *et al.* 2002)．

ばらつきの理由として，異なる部分である，mRNA のコード領域とタンパク質の配列

の違いに起因する可能性が高い．このコムギの系は再現性が高いため(Sawasaki *et al.* 2004)，そのような実験誤差以外の要因が十分に考えられる．そこで，合成量の違いを生み出す要因を，それらの核酸配列，アミノ酸配列から見つけ出すことを，本研究の目的とする．

そのことによって,無細胞合成系の改良に貢献することを最終目標とする．ハイスループットな合成系としては，得手不得手なく，どのようなタンパク質でも合成できることが望ましい．よって，本研究のように翻訳効率を変動させる要因を突き止めることは有意義である．そして,発見した要因への対策を基に，合成系の性能を少しでも向上させることを目指す．

1.3. 翻訳効率の先行研究

過去に，配列に基づく翻訳効率の研究では，どのような知見が得られているのだろうか．5'-UTR 領域が翻訳の制御に重要であるというのが，様々な生物を通じて共通の見解である(Kozak 2005).

開始コドン付近で，発現量の増加に関連する共通配列の存在が，幾つか報告されている．原核生物には SD 配列，downstream box (Stenstrom *et al.* 2001)，脊椎動物には Kozak 配列と呼ばれる共通配列が，翻訳効率の高い遺伝子の間で見つかっている．さらに，植物の遺伝子でも-3 が A，+4 が G の場合に翻訳効率が促進されるという報告がある

(Lukaszewicz *et al.* 2000). ただし，開始コドン AUG の A が+1 である．また，塩基対形成が翻訳効率の低下に影響するという傾向がある(Ohashi *et al.* 2005; Ringner *et al.* 2005). mRNA は，リボソームが通過する際に，一本鎖になっている必要があり，塩基対が多いと合成に不利になると考えられている．塩基対形成の場所は，特に，5'-UTR や開始コドン付近のものが重要であると言われている．

また先行研究と比べ，翻訳に特化した系を用いている点と，生成タンパク質そのものの量を計測している点が，本研究の特徴である．今回の研究のように，網羅的に合成効率を解析した研究はあるが,その解析では,翻訳途中のポリソーム形成能を翻訳効率と定義している(Ringner *et al.* 2005)．その場合，ポリソーム形成と生成タンパク質の量がどれほど比例するか明らかになっていないし，その後の折りたたみや活性についても議論できない点で，翻訳の効率と言えるのか疑問である．無細胞タンパク質合成系の性能向上を目的とした研究では，いかに機能性タンパク質を合成できるかが重要な目標であるため，生成タンパク質の量で議論することが大切だと考えられる．

1.4. 本研究の流れ

本研究では，可溶化タンパク質の量(合成量)を翻訳効率と定義した．そして,合成量の違いをその配列から説明することが可能かどうかを検証した．データセットとしてシロイヌナズナ(*Arabidopsis thaliana*)のタンパク質リン酸化酵素 423 個を用いた．

まずタンパク質の分子量やコドン使用頻度など基本的な統計量から翻訳効率に影響を与えている量があるかどうか確認した．次に，先行研究に従い，核酸配列，特に開始コドン付近の塩基対形成が合成量と関係があるのか調べた．その後，アミノ酸配列からも合成量と相関する指標が見つかるか探索した．そして最後に，見つかった翻訳効率決定因子の候補群の組み合わせで，合成量のばらつきがどの程度まで説明可能か検証した．

2. 方法

2.1. データセット AtPK (*Arabidopsis thaliana* protein kinases)

PlantsP データベース(http://plantsp.sdsc.edu)に登録されている, 1067 個のシロイヌナズナのタンパク質リン酸化酵素の中から 439 個を選び, コムギ胚芽無細胞タンパク質合成系で合成した(Sawasaki *et al.* 2004). その中で合成量が特に低かった 14 個は, プライマーの設計ミスなど翻訳過程以外での要因も考えられるためデータセットから除いた. また, 分子量が極端に少ない 2 つのタンパク質も除き, 423 個を本研究でのデータセット AtPK として用いた(付録参照).

合成量として, ロイシン同位体置換で計測した可溶化タンパク質の量を用いた. また, 不溶化タンパク質も計測し, 可溶化率も算出した. よって, データセットには 1 つのエントリーにつき, 以下の情報がある. 核酸配列, アミノ酸配列, 分子量, 合成量(μg/ml), 合成量(nM), 可溶化率. 基本的な統計量を表 1 に示す.

表 1 データセット AtPK

	合成量 (μg/ml)	合成量 (nM)	可溶化率 (%)	分子量	アミノ酸配列長
max	47.89	1615.0	100.00	140012	1263
mean	21.86	365.0	80.61	67337	604
S.D.	9.24	209.6	13.23	23466	215
min	0.93	11.0	32.10	26894	238

2.2. 配列解析の手法

mRNA の二次構造予測には Vienna RNA package (Hofacker 2003)の RNAfold を用いた. AAindex(Kawashima *et al.* 2000)の解析には，NA となっている指標は 0 に置き換え，それ以外の値はそのまま使用した. Disorder 領域の予測には DISOPRED2 プログラムを偽陽性率 2%の条件で使用した(Ward *et al.* 2004). coiled-coil 構造は COILS プログラムで予測し，推奨されているウィンドウ幅(14,21,28 残基)を使用した(Lupas *et al.* 1991). β aggregation は TANGO プログラムを用いて予測した(Fernandez-Escamilla *et al.* 2004).

配列のペアワイズアラインメントには ALIGN(Pearson *et al.* 1992)を用い，マルチプルアラインメントには ClustalW(Chenna *et al.* 2003)を用いた.

回帰木は JMP パッケージ(SAS Institute Inc.)の対話的パーティショニングを使用し構築した. その際の条件として, 分岐統計量を最大にすることを分岐の基準として選択し, 分岐の最小サイズを 20 エントリーにした. また，重回帰分析，ニューラルネットワークも同様に JMP パッケージを用いて解析した.

3. 結果と考察

3.1. データセットの基本統計量

コムギ胚芽無細胞合成系で合成した，シロイヌナズナのタンパク質リン酸化酵素 423 個を本研究のデータセット AtPK とした．合成反応時間は 17 時間で，反応液上澄み部分の可溶化タンパク質に対し，ロイシンの同位体置換を基に測定した値を合成量と定義した(付録参照)．

合成量は，ほぼ正規分布に従っている(図 1)．ただし，合成量が極端に多いあるいは少ないものの割合が若干高い．範囲は 0.93～47.89 μg/ml である(表 1)．

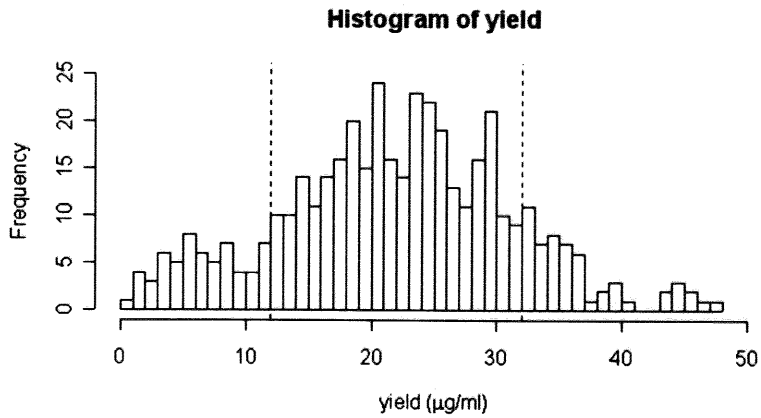

図 1　合成量(yield)の分布

横軸が合成量のヒストグラム．破線(12 μg/ml と 32 μg/ml)を境にして low, middle, high の 3 グループを定義する．

ここで合成量の単位として，グラム濃度かモル濃度のどちらを用いるべきかという問題があるが，今回はグラム濃度を採用した．合成量は nM で評価することもできるが，

転写産物を電気泳動で見た結果，転写の時点でグラム濃度でそろっていることが観察された．翻訳前にグラム濃度の単位で一定であるため，翻訳後のばらつきの評価もグラム濃度で行うことが妥当である．したがって，合成量の評価もグラム濃度，μg/ml で行うこととする．実際，先行研究もグラム濃度で評価しているものが多い(Yokoyama 2003; Sawasaki *et al.* 2004; Spirin 2004)．nM 単位ではタンパク質の長さと負の相関(corr = -0.551)があり，短いタンパク質ほどモル濃度の点で合成効率が良い．リボソームでのアミノ酸付加反応の速度がほぼ一定だとすると，長いタンパク質ほど生産するのに時間がかかるのは自明な傾向である．逆にμg/ml 単位では，タンパク質の長さ，ひいては分子量とは相関がない．

今回，合成量が 32 μg/ml 以上のタンパク質を合成量が多いグループ(high, 55 個)，12 μg/ml 以下を少ないグループ(low, 60 個)，そして残り(middle, 308 個)と大まかに 3 種類に分類し，後の検証に用いた．

3.1.1. 核酸配列の基本統計量

データセットの特性として，まずコドン使用頻度を調べた．コドン使用頻度の偏りが遺伝子の発現量と関係があることは従来から知られており，シロイヌナズナでも，発現量の多い遺伝子のコドンは偏っているという報告がある(Wright *et al.* 2004)．NCBI の Reference Sequence (Refseq)に登録されている *Arabidopsis thaliana* の配列を元にリボソー

ムタンパク質，いずれも発現量の高いタンパク質，のコドン使用頻度を調べてみると，*Arabidopsis thaliana* の全配列と比べて大いに偏ったコドンの使われ方であった(図2左)．例えば，図2左の上部にあるはずれ値はAAGコドンの使用頻度で，全遺伝子では3.2%の使用頻度なのに対し，リボソームタンパク質では8.8%と偏って使用されている．そこで，我々のデータセットでも，合成量が多いタンパク質はコドンの使用頻度に偏りがあるかどうかを確認してみた．しかし,合成量が多いグループ(high)と全データを比べてもほとんど使用頻度に偏りは見受けられず，合成量と関係するような傾向は観測されなかった(図2右)．その理由として，タンパク質ファミリーがタンパク質リン酸化酵素に限定されているため，データセットの中で使用頻度に偏りが生じにくいことが考えられる．

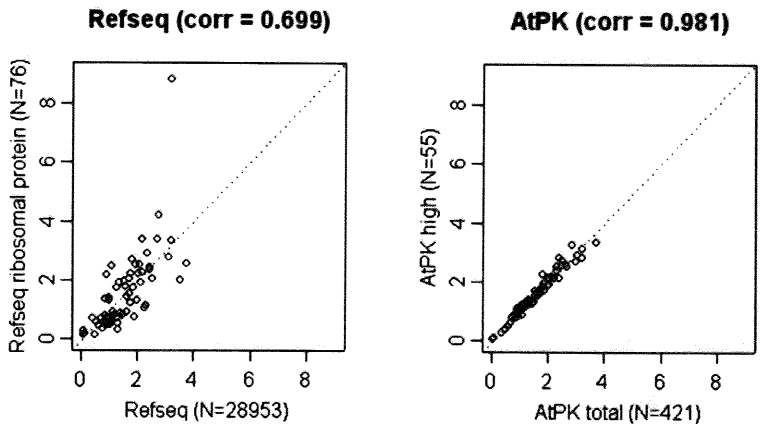

図 2　コドン使用頻度

縦軸，横軸共にコドンの頻度(%)であり，散布図中の1つの点が1つのコドンに対応する．左の図はデータセットAtPKに関し，横軸は全データセット421個の遺伝子におけるコドン使用頻度を示し，縦軸は合成量上位55個(highグループ)のコドン使用頻度である．右の図はRefseqに基づくもので，横軸はシロイヌナズナの全遺伝子，縦軸はシロイヌナズナのリボソームタンパク質76個のコドン使用頻度である．

また，他の核酸配列の特徴としてGC含量やthird codonを調べたが，合成量と関係するような傾向は観察されなかった．さらに，レアコドンの影響は，tRNAやATPなどの基質を過剰に供給しているため，生じにくいと思われる．

3.1.2. アミノ酸配列の基本統計量

アミノ酸使用頻度に関しても，AtPKの全データと上位55個の相関係数は0.981と高く，ほぼ同じ使用頻度であることが確認された．コドン使用頻度と同様に，翻訳物であるアミノ酸の使用頻度でも，翻訳効率に関係する有意な偏りは見られなかった．次にタンパク質の予測二次構造の割合は合成量と関係があるかどうかを確認した．二次構造予測プログラムにはPSIPREDを用いた(McGuffin *et al.* 2000)．各タンパク質の二次構造含量と合成量との関係を調べてみたが有意な傾向は得られず，αヘリックス含量との相関は-0.054, βストランドとは0.100，ランダムコイルは-0.022であった．

次に，PROSITE を利用してタンパク質に特徴的なモチーフまたはドメイン構造があるか調べた．しかし，タンパク質リン酸化酵素としてキナーゼドメインが検出されたのみで，ドメインの位置なども合成量と関係する傾向はなかった．

さらに，ドメイン領域だけではなく配列全体の類似性を見るために，タンパク質ファミリーによって合成効率に違いはあるか確認した．図3は，PlantsPデータベースに基づいて分類したファミリーでの合成量分布である．この中でグループEが，全体と比較

し最も偏った分布をしている．しかし，E は未分類タンパク質の集合であるため，配列の類似性から論ずることはできない．さらに，分散分析(ANOVA)で平均値の差を検定した結果，*P* value は 0.227 であり，このファミリー分類による合成量の違いには統計的有意性が得られなかった．よって，全体的な配列の類似性は，合成量とあまり関係がないことが示唆された．

また，ClustalW(Chenna *et al.* 2003)を用い分類分けを行った結果，ほぼ PlantsP と一致し，PlantsP の分類で配列の大域的な類似性を議論できることを確認した．

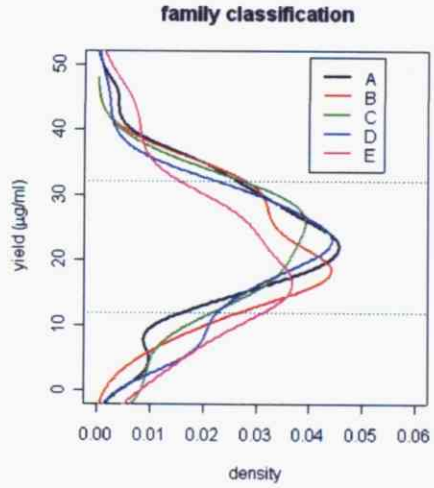

図 3　PlantsP ファミリー分類による合成量の分布

PlansP データベースの分類で 5 つに分けた合成量の密度分布．縦軸が合成量，横軸が密度である．分類は色分けして表示し，図に示すようにファミリーA～E に対応している．各ファミリーは以下のように PlantsP クラスで分類している．カッコ内はエントリー数である．

A (206): Transmembrane Receptor Kinase and Related non-Transmembrane Kinases

B (30): ATN1/CTR1/EDR1/GmPK6 like Kinase

C (11): Casein Kinase I

D (157): Non-Transmembrane Protein Kinases

E (20): Other and Unclassified Protein Kinase

ここまでの解析でアミノ酸配列の様々な基本的特徴を抽出した．しかし，有意な傾向は発見できず，合成効率はアミノ酸配列の大域的な特徴からは見つけることが難しいことが示唆された．

3.2. 合成量と mRNA の二次構造の関係

5'-UTR 上で，塩基対形成が多い遺伝子は発現量が低下するという報告がある(Kozak 2005; Ringner *et al.* 2005)．また，5'-UTR のみならず，開始コドン下流も含めた領域での塩基対形成も，発現量低下に影響するという傾向が *in vivo* の大腸菌の系で確認された (Ohashi *et al.* 2005)．

同様のことが，本研究のデータセットにも当てはまるか検証した．ただし，コムギの系は 5'-UTR が統一されている点に注意しなければならない．5'-UTR はΩ配列の上流に GAA を付加した配列(GAAΩ)に統一されている．またΩ配列は CAA リピートを特徴とした，グアニン(G)が極端に少ない，二次構造を形成しにくい配列である(図 4)．つまり，塩基対形成の観点から述べると，翻訳効率の高い 5'-UTR であると言える．よって，全データセットの 5'-UTR の安定化自由エネルギーは等しく高い．ところが，それであるがゆえに合成量のばらつきを 5'-UTR から議論することができない．

```
          ...(((....)))..........................................................................
omega     gaaguauuuuuacaacaauuaccaacaacaacaacaaacaacaacaacauuacauuuuacauucuacaacuaccacccaccaccaccaAUG
```

図 4　オメガ配列

コムギ胚芽無細胞タンパク質合成系の 5'-UTR の配列．すべての遺伝子でこれに統一している．上段は二次構造予測結果である．大文字の AUG は開始コドンを表す．

そこで，大橋らの手法(Ohashi *et al.* 2005)のように，開始コドン下流も含めた領域で調べることが有効であると考えた．合成量のばらつきが，開始コドン付近の塩基対形成に影響される可能性がある．よって，5'-UTR と CDS の間の相補対形成を予測し，自由エネルギーの観点から論ずる．用いた予測プログラムは，Vienna RNA package (Hofacker 2003)の RNAfold である．調べる領域は，大橋らの手法と同じく(Ohashi *et al.* 2005)，開始コドンの上流 25bp と下流 25bp に絞った(UTR+CDS)．また，比較対象として，CDS 領域を 5'-側をそろえて端から 50bp の部分(5'-end CDS)と，逆に 3'-側をそろえた 50bp の部分(3'-end CDS)を用いた(図 5)．

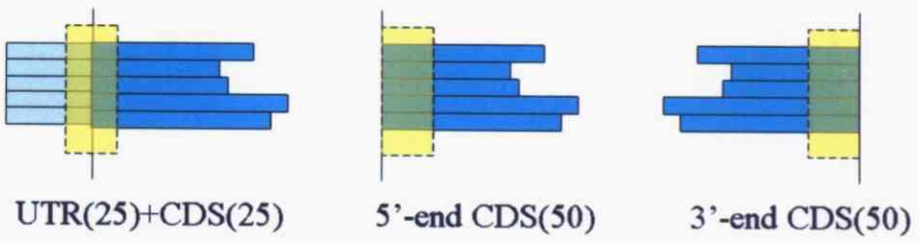

図 5　mRNA 局所的二次構造の探索領域

水色の四角は 5'-UTR を，青色の四角は CDS を表す．黄色い四角が探索領域である．

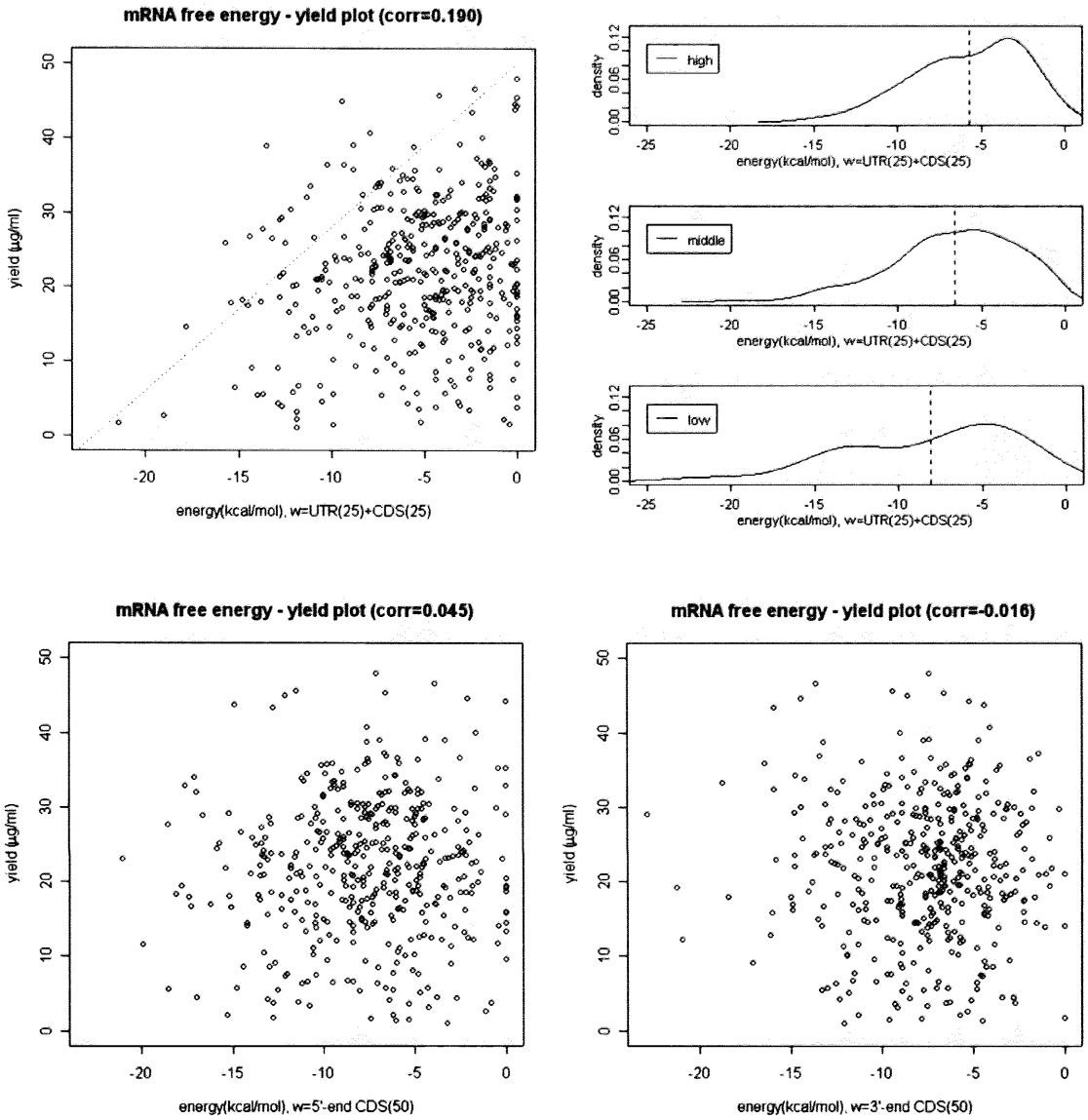

図 6　mRNA の安定化エネルギーと合成量との関係

3 つの散布図は，図 5 の 3 つの条件における，ウィンドウ内の mRNA 安定化エネルギー(横軸)と合成量(縦軸)の関係を示す．1 つの点が 1 つのタンパク質に対応する．左上の散布図が，ウィンドウ UTR(25)＋CDS(25)の場合であり，左下が 5'-end CDS(50)，右下が 5'-end CDS(50)である．

また，右上の図は UTR(25)+CDS(25)の条件で，合成量を high, middle, low(図 1 参照)に分けた時の密度分布を示す．横軸は mRNA の安定化エネルギーである．

結果は図6に示す．比較対象の2つと比べて，開始コドン付近の安定化自由エネルギーと合成量の間には，弱いながらも有意な正の相関が見られることがわかる(corr = 0.190, $p = 8.2 \times 10^{-5}$)．よって，5'-UTR 領域と相補塩基対を形成する 5'端 CDS 配列の存在により，タンパク質の翻訳効率が低下している可能性が示唆される．また，UTR+CDS の条件と，5'-end CDS の条件では，配列の半分(25bp)が重複しているにもかかわらず，合成量との相関係数が異なっている点が興味深い．5'-UTR と CDS という組み合わせがいかに重要であるかを示していると思われる．

また，mRNA の二次構造ではなく，一次構造である塩基 ATGC，またはコドンの並びからは有用な傾向は特定できなかった．先行研究で，植物の遺伝子では-3 が A，+4 が G の場合に翻訳効率が促進されるという報告がある(Lukaszewicz *et al.* 2000)．しかし，今回のデータセットにはあてはまらない．我々の統一した 5'-UTR では-3 の位置は C であり，また+4 も G で合成量が高いという傾向は特になかった．コムギの系は，真核生物の翻訳の基本であるキャップ構造と polyA が存在しない．そのため，リボソーム認識機構も例外的な振る舞いをし，-3A，+4G というルールに当てはまらないのではないかと考えられる．

大腸菌の研究では以前から，開始コドン後の数コドン(初期コドン，initial codons)が翻訳効率に影響を与えているという報告がある．しかし，我々の研究では同様の傾向を見出すことはできなかった．今回のデータでは，塩基対形成の自由エネルギーという観点

からは関連性が観察されたが，その傾向は必ずしもコドンの並びに現れるほど，明示的な傾向ではなかったと言える．

今回の解析で，5'-UTR の配列をタバコモザイクウィルス由来の配列に統一したという条件ではあるが，真核生物の mRNA でも開始コドン付近の塩基対形成が翻訳効率に影響を与えることが示唆された(図 6)．

3.3. 合成量と AAindex の関係

次に，アミノ酸配列上の特徴から合成量の違いを議論できるか解析した．そこで様々なアミノ酸の傾向を調べるために，アミノ酸インデックスデータベースである AAindex(Kawashima *et al.* 2000)を用いた．

AAindex とは，ある特徴に基づき 20 種類のアミノ酸を 20 個の数値に変換する指標で，現在 516 個もの指標が登録されている．それらは以下の 4 つに大別される：(i) α ヘリックスとターンの傾向 (ii) β ストランドの傾向 (iii) 疎水性 (iv) 物理化学的傾向．

タンパク質全長に対し AAindex で変換した数値の総和を求め，それを長さで正規化した値を各タンパク質固有の特徴量とした，そして，その AAindex 由来の特徴量と合成量との間で相関があるかを解析した．その結果，516 個全ての AAindex の場合による相関の分布は，無相関であるゼロを中心にほぼ正規分布していた．正の相関が最も高い 3 つ，負の相関が最も高い 3 つの計 6 つの AAindex を表 2 に示す．

表 2　合成量と相関の高い AAindex

AAindex ID	相関係数	Annotation
YUTK870102	0.195	Unfolding Gibbs energy in water; pH9.0
SNEP660101	0.191	Principal component I
NAKH900112	0.182	Transmembrane regions of mt-proteins
OOBM850104	-0.181	Optimized average non-bonded energy per atom
PUNT030102	-0.188	Knowledge-based membrane-propensity scale from 3D_Helix in MPtopo databases
VHEG790101	-0.194	Transfer free energy to lipophilic phase

一番正の相関の高い指標(YUTK870102)は，タンパク質の熱力学的安定性に関するものであった(Yutani *et al.* 1987)．指標の値は変性のギブスエネルギー変化(ΔG)である．この値が高いほど変性状態よりも天然状態のタンパク質の数が多い割合で平衡状態にある，つまり安定したタンパク質であることになる．この AAindex の値が正の相関を示したことから，合成量の高いタンパク質ほど安定なタンパク質である可能性が示唆される．逆に一番負の相関の高い指標(VHEG790101)は，タンパク質の膜通過転移(Trans-membrane translocation)の指標である(von Heijne *et al.* 1979)．値は，親油層への転移自由エネルギーである．

上記の解析では，タンパク質の長さで正規化しているため，大きなタンパク質ほど AAindex の値は平均的な値に近づく傾向にあることが問題である．また，今回のデータセットではアミノ酸使用頻度が似ていることため，全長に基づく AAindex の総和は似る傾向にある．それよりも，局所的な領域を探索した方が，より相関の高い特徴量が得

られると期待される．

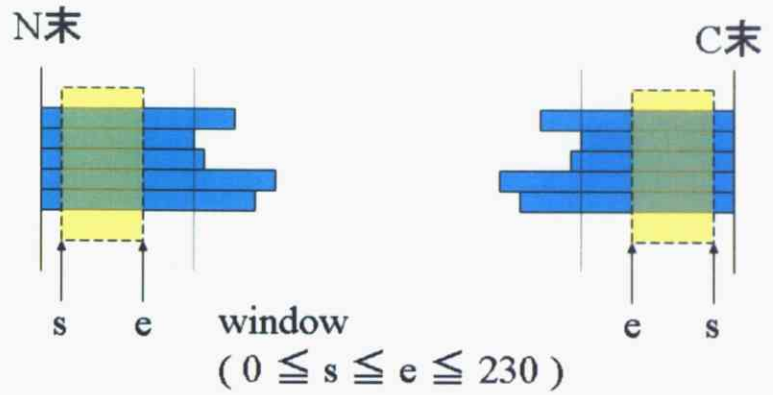

図 7　アミノ酸配列の探索領域

アミノ酸配列に対し，ウィンドウを設定して探索した方法を示す．左の図は，全ての配列を N 末端にそろえて解析した場合で，開始位置(s)と終了位置(e)をステップサイズ 10 残基で動かす．終了位置(e)の最大値が 230 なのは，データセット内の最小のアミノ酸配列長が 238 残基であるためである．同様に，C 末端でそろえた場合が右の図である．

そこで次に，一定の長さの部分配列に焦点を絞った(図 7)．アミノ酸配列に対しある長さのウィンドウを設けて，その中での AAindex の総和を求め合成量との相関係数を求めた．各 AAindex に対し，最も相関が高くなるようなウィンドウの幅と位置を決定した．ただし,ウィンドウの位置は全タンパク質で共通のものを用い，N 末端あるいは C 末端からの位置で揃えてある．個々のタンパク質ごとに可変にすることも可能である．しかし，異なる位置の間で生物学的な意義を論じるのは難しい上，作為的にウィンドウを配置してしまう恐れがあるため行っていない．最小値が 0，最大値が 230 の中で，ウィンドウを 10 残基刻みで変化させた．データセットの中で最も短いタンパク質が 233 残基であるため，その値を用いた．

その結果，C 末側領域に比べて N 末側領域特異的に高い相関が得られた．表 3 が相関が高かった AAindex のリストで，正の相関，負の相関共に上位 3 つは N 末側領域由来の指標である．C 末側領域由来のものの最大の相関係数は 0.207 であった．また，前回の配列全体での解析と比較しても，N 末側領域由来の相関係数は高くなった(表 2, 3)．

表 3　合成量と相関の高い AAindex(N 末端領域)

AAindex ID	相関係数	Annotation	s	e
YUTK870101	0.306	Unfolding Gibbs energy in water; pH7.0	10	200
ARGP820102	0.297	Signal sequence helical potential	90	180
ARGP820103	0.297	Membrane-buried preference parameters	90	180
GUYH850101	-0.281	Partition energy	10	200
PUNT030101	-0.283	Knowledge-based membrane-propensity scale from 1D_Helix in MPtopo databases	10	200
PUNT030102	-0.288	Knowledge-based membrane-propensity scale from 3D_Helix in MPtopo databases	10	200

表 3 で s はウィンドウの開始位置，e は終了位置を示す．例えば，YUTK870101 という AAindex は N 末端から数えて 11 残基目から 200 残基目の幅 190 残基の領域に対し，その指標で総和を取ると合成量との相関は 0.306 になることを示している．この領域を設けた探索でも，一番正の相関が高かった指標は熱力学的安定化エネルギーに関するもの(YUTK870101)であった．また，負の相関は PUNT030102 という指標が最も相関が高く，これは膜貫通タンパク質に関するものである(Punta *et al.* 2003)．PUNT030102 は前回の指標，VHEG790101(表 2)と似ている(AAindex 間の類似度，corr = 0.873)．よって，

全体での探索でも，領域に限定した探索でも同様の傾向を検出したことが示唆される．

最も相関のあった指標 YUTK870101 と合成量との散布図を以下に示す(図 8)．

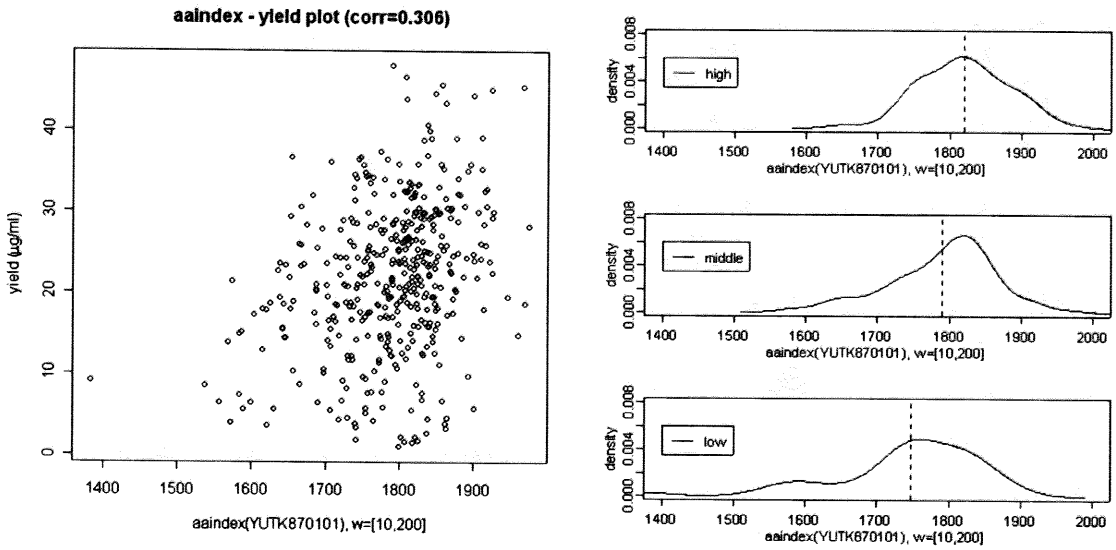

図 8　タンパク質の熱力学的安定性と合成量との関係

左図は，アミノ酸インデックス(YUTK870101)に基づく値(横軸)と合成量(縦軸)との関係を示す．右図は，合成量を high, middle, low(図 1 参照)に分けた場合の密度分布を示す．横軸は左図と同じである．

弱いながらも有意な正の相関があることが確認された(corr = 0.306, $p = 1.2 \times 10^{-10}$)．指標の大きい，つまり，熱力学的に安定なタンパク質は合成量が多く，逆のものは合成量が低い傾向にある．

以上より，様々な指標が合成量と関係のあることが示唆された．

しかしながら，各指標の間で相関が強く，完全に独立した指標ではないという傾向があり，注意が必要である．例えば表 3 の中で，YUTK870101(w=[10,200]) と

PUNT030102(w=[10,200])の間の相関係数は-0.861 であり，YUTK870101(w=[10,200])とARGP820102(w=[90,180])の間の相関係数は 0.750 である．つまり，正負を問わず，合成量との相関の高い AAindex 由来の指標は，互いに相関が高い傾向が観察された．特に，探索ウィンドウの位置が同じ値どうしではこの傾向がより顕著であった．この理由として，AAindex 間で似ているものがあるということと，window 内でアミノ酸の使用頻度が似ている傾向にあるという 2 点が考えられる．また別の注意点として，単純に総和をとることが適切かどうかということは議論の余地がある．しかし今回は当てはめるモデルの知識がないため第一近似として総和を使用した．

このように，AAindex を扱う際の注意点が残ってはいるが，それでも複数の特徴を網羅的に探索できることはこのデータベースを使用する強みであり，相関の高い指標(表3)は，実際に合成量に関係する特徴であると思われる．

よって，独立性の問題を踏まえ，最も相関の高かった指標(YUTK870101，タンパク質の熱力学的安定性)のみを AAindex 由来の因子として採用し，今後議論することにした．

3.4. 合成量と N 末端 disorder の関係

その他の翻訳効率決定因子を探索する際に，タンパク質の"凝集"のような現象を想定した．つまり生成ペプチド間で相互作用しやすいようなタンパク質は合成量が低下するのではないかと仮定したのである．そして今回，そのような相互作用部位として disorder

領域，coiled-coil 構造，β aggregation 部位という 3 つのタンパク質部分構造を候補に上げ，合成量との関係を解析した．

disorder 領域とは二次構造が主にランダムコイルから形成される，柔軟で決まった折りたたみ構造をとっていない領域のことで，真核生物のタンパク質に多い構造である(Ward *et al.* 2004)．今回，Disorder の予測に DISOPRED2 プログラムを用い，アミノ酸の各残基に対し disorder 状態を取るかどうかを判断した．その後，各アミノ酸の disorder 残基の割合を計算し，合成量との相関係数を求めた．その結果，タンパク質全長に基づく disorder 残基の割合と合成量との間には，弱い負の相関が観察された(corr = -0.122)．これに対応する無相関検定の *P* value は 0.012 で，両者に相関があるかどうか判断するのか難しい．したがって全体の傾向を見るよりも，部分領域に分けて調査すると，より相関の高い領域が発見できるのではないかと考えた．

タンパク質を N 末側領域 30%，中央領域 40%，C 末側領域 30%の三つの部分配列に分けそれぞれの領域ごとでタンパク質の合成量との相関を求めた．その結果，N 末側領域の disorder の割合と合成量の間に，特異的な負の相関(corr = -0.249)が確認できた．それに対しタンパク質の残る 2 つの領域では有意な相関はなかった(中央領域　corr = -0.018，C 末側領域　corr = -0.059)．

よって，N 末側領域に着目し AAindex と同様のウィンドウ探索(図 5)で相関が最大になる位置を求めた．その結果，N 末端から数えて 20 残基目から 206 残基目までの領域

の disorder 残基の割合が最も相関が高くなった(corr = -0.278, $p = 6.0 \times 10^{-9}$). 合成量と N 末端 disorder の割合に有意な負の相関がある. 加えて図 9 の右上部分が空白になっていることも重要である. そこから N 末側領域に disorder 領域が多いタンパク質は合成量が低下する傾向にあると言える.

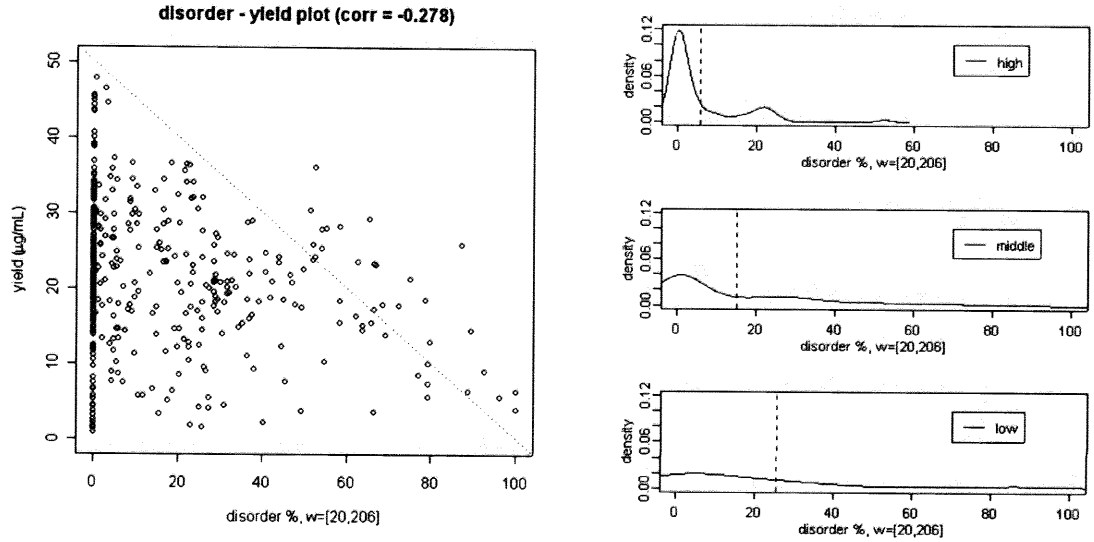

図 9　合成量と disorder との関係

左図は, N 末側領域の disorder の割合 (横軸)と合成量(縦軸)との関係を示す. 右図は, 合成量を high, middle, low(図 1 参照)に分けた場合の密度分布を示す. 横軸は左図と同じである.

しかしながら, その逆は成り立たない. disorder 領域がないタンパク質であっても合成量が増加しているものだけではないからである. disorder 残基を全く持たないタンパク質は全体の 43% (184/423)もあるが, 合成量の分布は広い. これは翻訳効率が多因子要因で決定されており, disorder だけで必ずしも全てが説明できるものではないことを示唆している. よって, 複数要因を組み合わせることが必要であり, その効果について

3.6.と 3.7.で議論する.

またN末端からの絶対的な位置ではなく，PROSITEに基づくキナーゼドメインからの相対的な位置を基準とした，disorderの割合を調べた．ドメイン内と，N末側ドメイン外，C末側ドメイン外と3分割し，それぞれで合成量との相関を調べた．その結果，N末側ドメイン外の相関が-0.131，ドメイン内が-0.086，C末側ドメイン外が0.068となった．この解析からも，N末側に基づく相関が一番高くなったが，N末端からの絶対的な位置の場合ほど有意な傾向は見られなかった.

以上より，N末側領域の構造がふらふらしているタンパク質は合成量が低いという関係が示唆された．タンパク質の二次構造ランダムコイルの含量では，なんら有意な傾向を観察することはできなかったが(corr = -0.022, 3.1.参照)，見方を変え精査することにより合成量と関係のありそうな因子 disorder (corr = -0.278, 図9)を抽出することができた.

3.5. 合成量と coiled-coil の関係

coiled-coil構造とは2〜5のタンパク質多量体内で形成される構造で，両親媒性ヘリックスがお互いにねじれ合って超螺旋構造を形成し，それによりcoiled-coilは新たな一時的凝集部位となる可能性がある．COILSプログラム(Lupas *et al.* 1991)で予測し，1部分でもcoiled-coil構造があると判定されたタンパク質をCoil+グループ，そうではないものをCoil-グループに分けた．その結果，119個(28%)のタンパク質がCoil+，304(72%)

個のタンパク質が Coil-と予測された．それらの合成量の分布は図 10 の通りで，coiled-coil 構造があると判定されたタンパク質の合成量は wilcoxon 検定 $P < 0.05$ で有意に低下する傾向にあった．

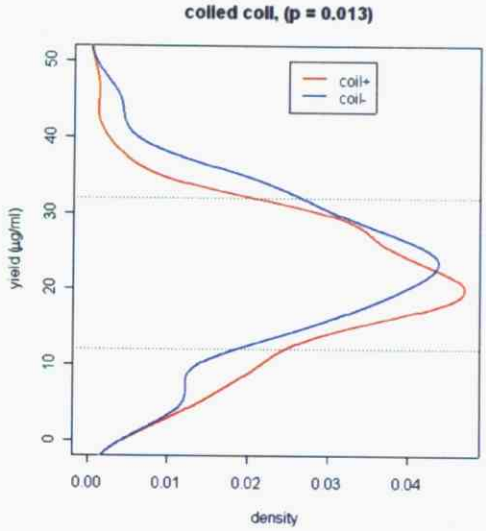

図 10　coiled-coil 構造の有無に基づく合成量分布の違い
coiled-coil 構造をとると予測されたタンパク質を coil+(赤線)．そうではないものを coil-(青線)とした場合の密度分布．縦軸は合成量である．

特に合成量が多いグループ(32µg/ml 以上)の中で coiled-coil 構造を形成しているタンパク質が少ない(13% = 7/55)ことが観察された．しかし，分布のずれは統計的に有意であっても図 10 の通り 2 つの分布はほとんど重なっているので，coiled-coil 構造のみから翻訳効率を判断するのは難しいと思われる．

また別の特徴量として，β ストランド間で形成されうる構造，β 凝集(β aggregation)について調べた．β 凝集とはアミロイドタンパク質などのβストランド間で形成される相互作用で，TANGO プログラム(Fernandez-Escamilla *et al.* 2004)を用いて予測した．し

かし，有意な関連性は観測できず(corr = 0.090)，合成量に影響を与えるような因子ではないと考えられる．また，TANGO は短いアミロイドタンパク質の凝集を予測するプログラムであるため，今回のような通常のタンパク質の予測には応用できていない可能性もある．

3.6. 4 つの翻訳効率決定因子の関係

今回の解析で，翻訳効率に影響を与える因子として以下の 4 つが同定された．(i) 開始コドン付近の塩基対形成，(ii) N 末側領域の熱力学的安定性，(iii) N 末側領域の disorder 残基の割合，(iv) coiled-coil 構造の有無．

因子はそれぞれ独立であれば望ましいが，実際は(ii)と(iii)の間で高い相関がある(corr = -0.756)．この傾向は生物学的意義に由来するよりはむしろ，因子の元となる配列の探索領域が似ているところから生じているのだと考えられる．(ii)の探索領域は 11 残基目から 200 残基目で，(iii)は 20 残基目から 206 残基目であり,ほぼ重複している．では，因子を軸にとった空間に合成量はどのように分布しているのであろうか？coiled-coil は 2 値データであるため除き，残りの 3 因子を軸に 3 次元プロットを行ったのが図 11 である．

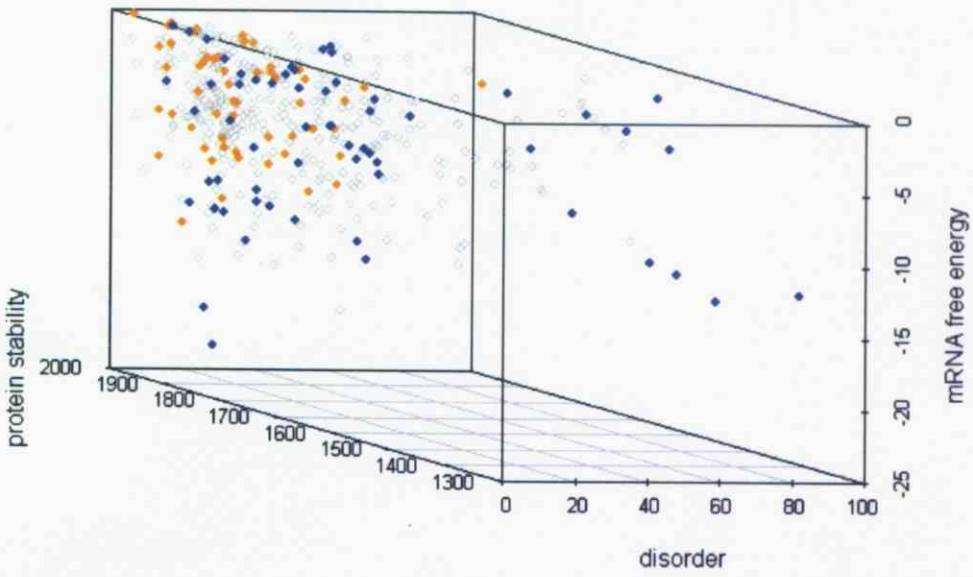

図 11　翻訳効率決定因子を用いた 3 次元プロット

disorder の軸は DISPRED の予測結果, mRNA free energy の軸は RNAfold の予測結果, protein stability の軸は AAindex から算出した値を基にしている．各点はタンパク質に対応し，high，middle，low グループ(図 1 参照)をそれぞれオレンジ，灰色，青と色分けして表示している．

因子の個別の傾向から，図の左上で合成量が高い傾向にあり，図の右下で合成量が低いということが予想される．実際に，合成量が高いタンパク質(high グループ)は左上に集まり，合成量が低いタンパク質(low グループ)は全体に広がっているようである．よって軸の傾向に一致するように分布していることが確認できる．しかしながら，high グループと low グループでさえまだまだ重なりが大きく，この 3 因子で合成効率全てを説明できたとは言えない．

3.7.　4 因子組み合わせの評価

では，今回発見した 4 つの翻訳効率決定因子で翻訳効率をどのくらい説明できるであろうか．それを検証するために図 1 で定義した，合成量の多いグループ(high)，少ないグループ(low)を検出する回帰木を構築した．使用したプログラムは JMP パッケージ(SAS Institute Inc.)である．

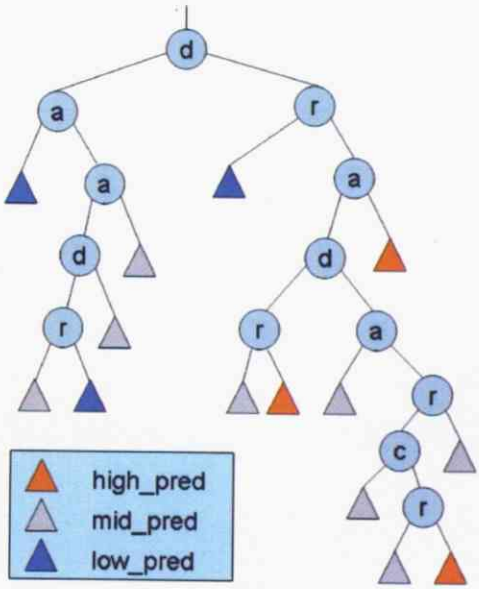

図 12　回帰木による分類

構築された回帰木．内部ノードはどの因子による分類か示す．d は disorder, a は AAindex, r は RNAfold, c は coiled-coil である．また，各内部ノード(翻訳効率決定因子)において，合成量が低下する傾向にあるグループを左の子に，増加する方を右の子に分類した．

木の分岐は合成量の値に対する分岐統計量を最大にすることを基準にした．つまり，分割後の 2 グループの合成量の平均値の差が最も大きくなるような指標を選択する．また，各葉は 20 以上のエントリーを持つべきであるという制限を設けた．その条件で木

を構築した結果，14 個の葉(グループ)に分類された．その中で合成量が多いほうに分布した 3 つを合成量が高いと予測されたグループ(high_pred, 72 個)，少ないほうに分布している 3 つを低いグループ(low_pred, 86 個)，残りを mid_pred, 265 個として，実際の合成量ラベル(図 1 参照)との比較を行った．

表 4　予測値と実測値のクロス統計表

	high_pred	mid_pred	low_pred	
high	24	28	3	55
middle	45	210	53	308
low	3	27	30	60
	72	265	86	423

予測の正答率は 62.6%(265/423)である．事前分布に基づく判別では high, low 共に 14%程度の正答率しか望めないのに比べ，4 つの因子を組み合わせることで予測精度が改善されている．具体的には，真の合成量が少ないタンパク質のうちの 50%(30/60)が正しく予測され，合成量が高いタンパク質は 44%(24/55)が正しく予測された．また，この予測では high が low に(3/55)，または low が high に(3/60)という極端な予測エラーの割合が少ない点もよい．ただし，木が深い，複雑すぎるルールは普遍性がない可能性が考えられ，また,予測精度には改善の余地がある．ただし，現状の 6 割程度の予測性能では，予測プログラムを改良することよりも，新たな因子を探索することの方が改善が見込まれる．さらに，無細胞合成系の改良を目指す本研究としても，実験系に応用しやすい新しい因子を見つけることが重要である．

合成量低下ルールに「disorder が高く AAindex が低い」があり，合成量増加ルールに「disorder が低く，AAindex が高く mRNA energy が高い」というものがある．これらは，各因子の傾向を反映するものであった．また，「disorder は低いが，mRNA energy が低い」という合成量低下ルールがあった．Disorder 因子のみの傾向とは相反することから，このルールは mRNA energy 1 因子による決定の影響が大きいと考えられる．

この回帰木作成の過程において coil の寄与があまりなかったことから，今回のルールは図 11 の 3 次元プロットを軸に平行に格子状に分割することに相当している．また，因子は個別に見ると合成量低下に関連するもので．単独では合成量の増加に関する知見は得られなかったが(ただし，AAindex を除く)，因子を組み合わせることにより合成量が増加する傾向も見出すことができた．合成量低下要因群の否定集合が合成量増加グループとして判別できたのである．

3.8. 翻訳効率予測の可能性

前節では合成効率決定因子の有効性を検討したが，これから一般的な機械学習問題について考えてみる．つまり，配列由来の特徴量（説明変数）から合成量（目的変数）がどの程度予測できるのか，その際に，どの機械学習プログラムが適しているのかを議論する．ただし,今回のデータセットは 3.9.節で述べるように，配列が酷似しているが合成量が異なるペアが存在するため，それらを区別して予測することは不可能である．

説明変数には，今回同定した合成効率決定4因子に加え，アミノ酸配列長，タンパク質二次構造含量，コドン使用頻度，AAindex 由来の因子 5 つ，計 77 個の特徴量を使用した．AAindex 由来の 5 因子は，表 3 に掲載されている指標を用いた．目的変数は数値属性である合成量そのもので，実測値と予測値の間の誤差二乗平均を予測の精度とし，5-fold 交差検定で評価した．回帰木，重回帰分析，ニューラルネットの手法を用いて予測精度を検証した．

表 5　予測精度

	誤差二乗平均	
	訓練セット	テストセット
回帰木	66.8	85.3
重回帰分析	58.0	102.3
ニューラルネットワーク(全て)	17.1	211.8
ニューラルネットワーク(選択)	53.0	94.5

合成量の実測値と予測値の差を誤差とし，5-fold 交差検定で評価した．ニューラルネットワークは全ての特徴量を用いた場合と，選択した場合で評価した．

まず，回帰木において，各葉は 40 以上のエントリーを持つべきであるという制限を設けて木を構築した．予測精度は表 5 に示す．今回用いた手法の中では，一番良い精度であった．しかしながら，以下の性質から回帰木は，合成量が高いものあるいは低いものをよりすぐって予測する，という目的には不向きであると考えられる．回帰木の予測値には，各葉に所属する訓練セットの平均値が割り当てられる．よって,平均をとることから，この手法では極端に合成量が多い少ないものを予測することができない．加え

て，テストセットの予測率向上との兼ね合いから，詳細に分割された木を構築することができないため，各葉には数多くのエントリーがある．そのため,結果的に全体の平均値に近づくような値を返し極端なはずれ値へのあてはめが悪くなる．

重回帰分析では，77 特徴量全てを用い最小二乗法で回帰モデルを作成した．この手法では，説明変数の次元が大きいほど予測値の範囲が大きくなる傾向にあった．少ない特徴量から回帰モデルを作成すると，ほとんどのものが 15～30μg/ml の範囲に収まって極端な外れ値を予測するのは難しい．これは，個々の特徴量の相関，寄与度が弱いことが理由であると考えられる．

同様にニューラルネットワークでも全ての特徴量を用いて予測を行った．隠れ層は 1 層で 3 ユニットからなり，オーバーフィットペナルティは 0.001 にした．この場合，表 5 にあるとおりテストセットでの予測が悪く典型的な過学習に陥っている．ペナルティを小さくして過学習を緩和させようと試みたがあまり改善が見られなかった．これは，特徴量を全て用いていることが原因だと思われる．合成量に影響のない特徴(コドン使用頻度など)を数多く用いることで精度が悪くなっている可能性が考えられる．

よって，少ない説明変数でニューラルネットを構築した．説明変数には，今回同定した合成効率決定 4 因子に加え，アミノ酸配列長，ランダムコイルの二次構造含量，コドン aaa,act,caa,gaa の使用頻度，最も負の相関のあった AAindex，計 11 個の特徴量を使用した．その結果，単純な重回帰モデルより精度の高い予測が可能になった．

今回の解析で，どの手法でも予測値の範囲が狭く同じような値を返す傾向にあった．よって今後は，モデル構築と精度評価の両方の段階で，極端な合成量をうまく予測するように重みをつける必要がある．また，根本的な問題として，用いた配列由来の特徴量が合成量と関連性が薄いものが多かったため予測精度が悪くなったと考えられ，機械学習の観点からも，さらなる特徴量を抽出することは重要である．

3.9. 配列のより局所的な特徴の重要性

今回，発見した因子は配列の局所的な傾向に基づくものであった(coiled-coil は全体的な傾向)．またファミリー分類の確認(図 3)から，配列の全体的な類似性からは合成効率に関する情報を得られないことも示唆された．よって，配列の局所的な条件が合成量を決定している可能性がある．

我々のデータセットには興味深い例がある．配列が酷似しているのに合成量が大いに異なるというペアがいくつか存在するのである．データセットに対し，ALIGN (Pearson *et al.* 1992)を用いて全てのペアの組み合わせでペアワイズアラインメントを行った結果，32 ペアが 80%以上の配列一致度を持ち，またその中で 9 ペアの合成量がそれぞれ 2 倍以上異なっていた．再現性の実験から，観測された合成量の違いは有意であると考えられるため，その差には実験誤差以上の意味があると思われる．そこで，やはり配列の異なる部分にその意味が隠されているのではないかと考えた．しかし，それらのペアの相

違には共通性は見られなかった．例えば配列の長短は合成量の大小と無関係であり,またプロリン置換も方向性があるわけではなかった．図 13 に最も類似度の高かったペアのアラインメント結果を示す．

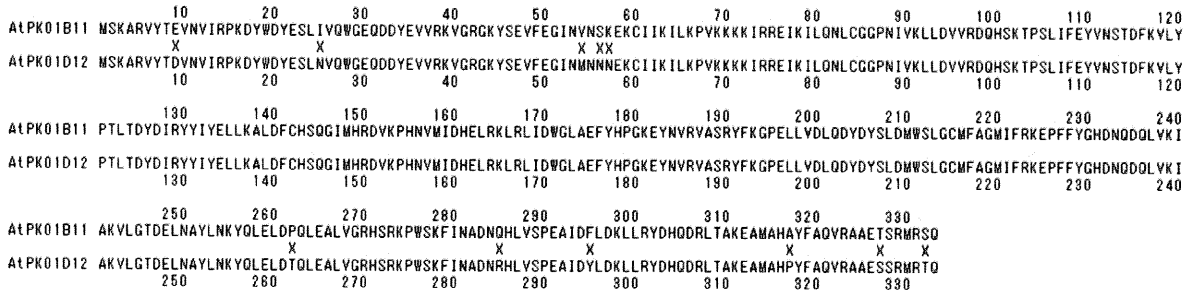

図 13　最も類似度の高いペア

AtPK01B11 の合成量は 6.64 μg/ml，AtPK01D12 は 21.02 μg/ml である．これらのアラインメント結果を示す．長さは共に 333 残基である．

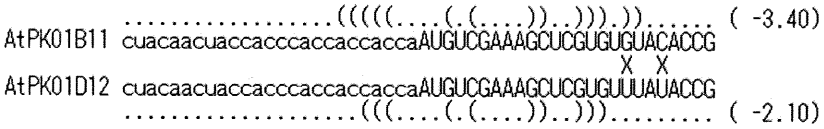

図 14　塩基配列の相違と二次構造変化

最も類似度の高いペアの，開始コドン周り 50 塩基のアラインメント結果と，二次構造予測結果を示す．また，カッコ内の数字は二次構造に対応する自由エネルギー(kcal/mol)である．

アミノ酸配列の一致度は 97%(核酸配列は 87%)であるが，合成量は 3.2 倍も異なる(21.02 と 6.64 μg/ml)ペアである．言い換えると，わずか 11 残基の変異で合成量が 3 倍以上も変動していることになる．これだけ酷似しているペアに対しては，今回同定した翻訳効率決定因子を使って予測することはできない．例えば，確かに図 14 のようにミスマッチによって mRNA 二次構造の自由エネルギーが変化しているが，図 6 と照らし合わせてみるとその差は些細な違いであり，自由エネルギーのみから合成量を説明する

ことはできなかった．したがって，今後はより局所的な特徴を抽出し，このようなわずかな差異に基づく因子を同定することが重要であると考えられる．

3.10. 転写因子での検証

これまで，タンパク質リン酸化酵素 423 個に対して因子探索を行ってきた．転写因子 661 個に対しても網羅的に合成量を計測したデータが入手できたので同様の傾向が言えるのかを解析した．反応時間は前回と同じく 17 時間であるが，細胞抽出液の改良等，条件が少し異なる．まず，転写因子はタンパク質リン酸化酵素と比べ，分子量が低く，またファミリー内での配列類似度は低かった．

転写因子は，リン酸化酵素と同じく開始コドン付近の塩基対形成が合成量に影響しているという傾向が観察された(図 15)．CDS の両端領域と比べて，UTR＋CDS の領域の mRNA 自由エネルギーと合成量が有意に相関していたのである(corr = 0.104, p = 0.007)．しかしながら，アミノ酸配列に関する 3 因子は転写因子では関係が観察されなかった．転写因子ファミリーはあまり類似性のない短いタンパク質の集合であるため，数百残基を比較することで傾向を見出すことは難しいことも考えられる．

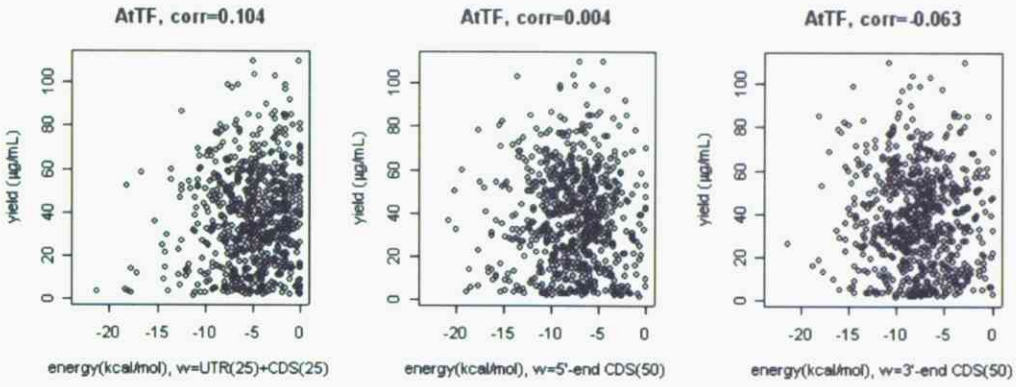

図 15　mRNA の安定化エネルギーと合成量との関係(転写因子)

3 つの散布図は，図 5 の 3 つの条件における，ウィンドウ内の mRNA 安定化エネルギー(横軸)と合成量(縦軸)の関係を示す．1 つの点が 1 つのタンパク質に対応する．左上の散布図が，ウィンドウ UTR(25)＋CDS(25)の場合であり，左下が 5'-end CDS(50)，右下が 5'-end CDS(50)である．

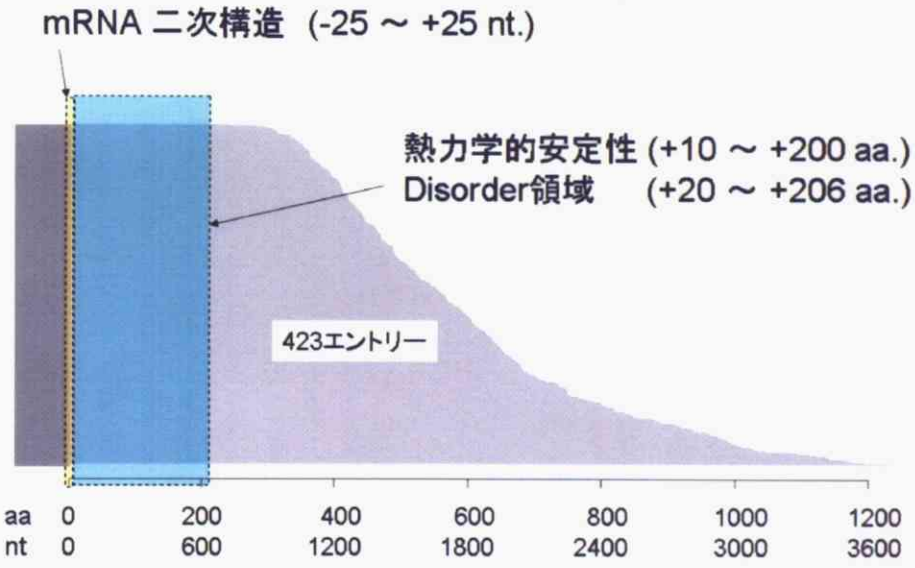

図 16　翻訳効率決定要因の位置

データセット AtPK の 423 配列の 5'-UTR(濃い灰色)と CDS(薄い灰色)の長さを示す．短いタンパク質から順番に上から並べて表示した．aa.はアミノ酸残基，nt.は塩基の位置を示す．

4. まとめ

本研究では，コムギ胚芽無細胞タンパク質合成系におけるシロイヌナズナのタンパク質の翻訳効率とその配列上の特徴との相関を探索した．その結果，開始コドン下流の塩基対形成が合成量に影響を及ぼすという原核生物での報告と一致する傾向を見つけた．またタンパク質リン酸化酵素の配列に関し，N 末側領域の特徴が合成量に影響を及ぼす可能性が示唆された(図 16)．N 末端からの絶対的な位置に基づく領域で顕著であったことは興味深く，翻訳過程の開始，あるいは前半が翻訳効率に重要であると考えられる．

同一条件で数百のタンパク質を合成した量を扱った初めての解析であり，今回の知見はタンパク質の大量生産に役立つものと考えられる．特に，mRNA の自由エネルギーに影響される合成量の低下は同義コドンで改善できる可能性があり，同義置換した PCR プライマーの作成し合成量がどの程度変化するかという実験的検証を予定している．

5. 謝辞

本研究を行うに当たって，多忙にもかかわらず熱心に指導して下さった，指導教官の高木利久教授に感謝致します．また，議論の場を与えて下さった，高木研究室の皆様に感謝します．共同研究者である,愛媛大学の遠藤弥重太教授，澤崎達也助教授,理研の関原明先生，篠崎一雄先生には貴重なデータの提供と，実験系からの議論や提案をして頂き御礼申し上げます．また，細やかな指導と鋭い指摘をして下さった山下理宇助手，木下賢吾助教授，両先生方に感謝致します．そして，研究の場を与えてくれた，東京大学医科学研究所の中井研究室の皆様に感謝いたします．最後になりましたが，数多くのご指導ご鞭撻を頂いた中井謙太教授に厚く御礼申し上げます．

6. 参考文献

Blaschke, U. K., J. Silberstein, et al. (2000). "Protein engineering by expressed protein ligation." Methods Enzymol **328**: 478-96.

Chenna, R., H. Sugawara, et al. (2003). "Multiple sequence alignment with the Clustal series of programs." Nucleic Acids Res **31**(13): 3497-500.

Chrunyk, B. A., J. Evans, et al. (1993). "Inclusion body formation and protein stability in sequence variants of interleukin-1 beta." J Biol Chem **268**(24): 18053-61.

Fernandez-Escamilla, A. M., F. Rousseau, et al. (2004). "Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins." Nat Biotechnol **22**(10): 1302-6.

Gallie, D. R. (2002). "The 5'-leader of tobacco mosaic virus promotes translation through enhanced recruitment of eIF4F." Nucleic Acids Res **30**(15): 3401-11.

Hofacker, I. L. (2003). "Vienna RNA secondary structure server." Nucleic Acids Res **31**(13): 3429-31.

Kamura, N., T. Sawasaki, et al. (2005). "Selection of 5'-untranslated sequences that enhance initiation of translation in a cell-free protein synthesis system from wheat embryos." Bioorg Med Chem Lett.

Kawashima, S. and M. Kanehisa (2000). "AAindex: amino acid index database." Nucleic Acids Res **28**(1): 374.

Kim, D. M., T. Kigawa, et al. (1996). "A highly efficient cell-free protein synthesis system from Escherichia coli." Eur J Biochem **239**(3): 881-6.

Kozak, M. (2005). "Regulation of translation via mRNA structure in prokaryotes and eukaryotes." Gene **361**: 13-37.

Lengyel, P. and D. Soll (1969). "Mechanism of protein biosynthesis." Bacteriol Rev **33**(2): 264-301.

Lukaszewicz, M., M. Feuermann, et al. (2000). "In vivo evaluation of the context sequence of the translation initiation codon in plants." Plant Science **154**(1): 89-98.

Lupas, A., M. Van Dyke, et al. (1991). "Predicting coiled-coils from protein sequences." Science **252**(5010): 1162-4.

Madin, K., T. Sawasaki, et al. (2000). "A highly efficient and robust cell-free protein synthesis system prepared from wheat embryos: plants apparently contain a suicide system directed at ribosomes." Proc Natl Acad Sci U S A **97**(2): 559-64.

McGuffin, L. J., K. Bryson, et al. (2000). "The PSIPRED protein structure prediction server." Bioinformatics **16**(4): 404-5.

Ohashi, Y., A. Yamashiro, et al. (2005). "In silico diagnosis of inherently inhibited gene

expression focusing on initial codon combinations." Gene **347**(1): 11-9.

Pearson, W. R. and W. Miller (1992). "Dynamic programming algorithms for biological sequence comparison." Methods Enzymol **210**: 575-601.

Punta, M. and A. Maritan (2003). "A knowledge-based scale for amino acid membrane propensity." Proteins **50**(1): 114-21.

Ringner, M. and M. Krogh (2005). "Folding Free Energies of 5'-UTRs Impact Post-Transcriptional Regulation on a Genomic Scale in Yeast." PLoS Comput Biol **1**(7): e72.

Ryabova, L. A., S. A. Ortlepp, et al. (1989). "Preparative synthesis of globin in a continuous cell-free translation system from rabbit reticulocytes." Nucleic Acids Res **17**(11): 4412.

Sawasaki, T., Y. Hasegawa, et al. (2004). "Genome-scale, biochemical annotation method based on the wheat germ cell-free protein synthesis system." Phytochemistry **65**(11): 1549-55.

Sawasaki, T., T. Ogasawara, et al. (2002). "A cell-free protein synthesis system for high-throughput proteomics." Proc Natl Acad Sci U S A **99**(23): 14652-7.

Spirin, A. S. (2004). "High-throughput cell-free systems for synthesis of functionally active proteins." Trends Biotechnol **22**(10): 538-45.

Spirin, A. S., V. I. Baranov, et al. (1988). "A continuous cell-free translation system capable of producing polypeptides in high yield." Science **242**(4882): 1162-4.

Stenstrom, C. M., H. Jin, et al. (2001). "Codon bias at the 3'-side of the initiation codon is correlated with translation initiation efficiency in Escherichia coli." Gene **263**(1-2): 273-84.

von Heijne, G. and C. Blomberg (1979). "Trans-membrane translocation of proteins. The direct transfer model." Eur J Biochem **97**(1): 175-81.

Ward, J. J., J. S. Sodhi, et al. (2004). "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life." J Mol Biol **337**(3): 635-45.

Wright, S. I., C. B. Yau, et al. (2004). "Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata." Mol Biol Evol **21**(9): 1719-26.

Yokoyama, S. (2003). "Protein expression systems for structural genomics and proteomics." Curr Opin Chem Biol **7**(1): 39-43.

Yutani, K., K. Ogasahara, et al. (1987). "Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit." Proc Natl Acad Sci U S A **84**(13): 4441-4.

7. 付録

7.1. コムギ胚芽無細胞タンパク質合成系

コムギ胚芽の精製は以下の手順である．小麦の実を製粉器で挽き，710-850 mm の網でふるいにかけた．小麦胚芽は Erickson and Blobel の溶媒浮遊法で選出した．その溶媒にはシクロヘキサンと四塩化炭素を含む(240:600, vol/vol)．傷ついた胚芽と混入物質は除き，無傷の胚芽のみをドラフト内で一晩乾かした.混入している胚乳を除くため，胚芽を 3 回，10 vol の水で強力に撹拌，洗浄した．その後，Bronson model 2210 超音波粉砕機を使って Nonident P-40 の 0.5%溶媒の中で 3 分間超音波分解した．最後に超音波粉砕機内でもう一度，滅菌水で小麦胚芽を洗浄した．

次に，コムギ胚芽抽出液を作成する．用いた手法は Erickson と Blobel の方法を少し修正したものである．洗浄後の胚芽を液体窒素の中で挽き細かい粉末にした．その粉末 5g を 2 倍の buffer A に加えた．Buffer A : 40mM Hepes, pH 7.6 / 100mM potassium acetate /5mM magnesium acetate / 2mM calcium chloride / 4mM DTT / 0.3mM of each of the 20 amino acids. その混合物を少し撹拌，そして 30000*g で 30 分遠心分離した．その結果，上清を buffer A で 200 A260/ml に調整し，少量ずつに分割し使用時まで液体窒素内に保存した．

生合成システム．RIKEN Arabidopsis full-length protein kinase clones に対する split-primer 法のためのユニークなプライマーは，そのデータベースに登録されている配

列に従ってデザインされた．STA 配列，SP6 プロモーターの 3'部分に GAAΩが続く 2 本鎖 DNA, の導入のために STA の ORF と 5'-CCACCCACCACCACCA を PCR で増幅，精製後 primer-2 の代わりに使用した(0.2mM)．96 well プレートで一晩培養した 3ml の E. coli 懸濁液を 60 ml PCR に対し使用した．PCR の増幅反応は飽和水準まで達し，結果の DNA は上記のように転写された．マイクロタイタープレートの各 well 内の転写物はエタノールで沈殿され，Hitachi R10H rotor を用いプレートの遠心分離で沈降させ,沈殿物を回収した．洗浄した mRNA(通常 30-35μg)を 50ml の翻訳混合物の中に移し，重層モード(Sawasaki *et al.* 2002)で反応を行った．下層にあたる翻訳混合物は抽出液を $3A_{260}$ ユニット含み(A_{260}/A_{280}=1.55), 翻訳溶媒に含まれる様々な成分は以下の通りである．24 mM Hepes/KOH, pH 7.8, 1.2 mM ATP, 0.25 mM GTP, 16 mM creatine phosphate, 2 μg creatine kinase, 2 mM DTT, 0.4 mM spermidine, 0.3 mM の各 20 種のアミノ酸，2.8 mM magnesium acetate, 100 mM potassium acetate, 0.005% NaN3 (基質溶媒と見なされる)．上層として，125 μl の基質溶媒を翻訳混合物の上に乗せた．そして 17 時間 26℃で培養した．合成機械 PIM は転写からタンパク質生産までの全工程を行うことが可能である．mRNA の精製，その後の翻訳のためのピペッティング，混合，培養，遠心分離等の基本的な作業も備え付けてある．

合成量の測定はロイシンの同義体置換で計測した．^{14}C でラベルされたロイシンを反応系に加えると，無細胞合成系であるため加えたロイシンは目的のタンパク質のみに組

み込まれる．トータル溶液と遠心後の上澄みをそれぞれ5μlずつワットマンの濾紙にスポットして乾燥後，ホットTCA法でタンパク質に取り込まれなかったアミノ酸を除去した．液体シンチレーションで測定したものはタンパク質に組み込まれたロイシンの値となり，タンパク質中のロイシン頻度から合成量を算出した．今回，上澄み部分の可溶化タンパク質の値を合成量として用いる．また，上澄みとトータル溶液の比率を可溶化率とする．