Title	ゲノム情報を利用した新規RNA結合蛋白質推定とその構造パターン予測
Sub Title	
Author	藤島, 皓介(Fujishima, Kosuke) 金井, 昭夫(Kanai, Akio)
Publisher	慶應義塾大学湘南藤沢学会
Publication year	2004-06
Jtitle	研究プロジェクト優秀論文
JaLC DOI	
Abstract	本書は、生命情報工学を用いた新たな2つの蛋白質機能推定法について報告している。機能が既知のタンパク質に現れるアミノ酸の種類やその頻度、周期性をコンピュータ解析により詳細に検討する事で、これまでにない２つの新しいタンパク質の機能推定法を考案し、さらにこの方法を駆使する事で機能未知であったタンパク質の中から、核酸制御に関わる新規の遺伝子産物を指定していく事に成功した。
Notes	冨田・内藤・中山研究プロジェクト 2003年秋学期
Genre	Technical Report
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=0302-0000-0483

SFC-SWP 2003-A-005

研究プロジェクト優秀論文集

ゲ

Keio University Shonan Fujisawa Academic Society

ノム情報を利用した新規RNA結合蛋白質推定とその構造パターン予測

2003年　秋学期
AUTUMN

藤島　皓介　環境情報学部 3年

冨田・内藤・中山 研究プロジェクト

慶應義塾大学湘南藤沢学会

「研究プロジェクト」優秀論文推薦のことば

近年、生命科学の分野では、バクテリアからヒトに至るまで、いろいろな種でそのゲノム配列(遺伝子DNAの塩基配列の並び方)が決定されました。今世紀はこれら遺伝子の機能を明らかにすることに研究の中心が移っていますが、ゲノムにコードされている全タンパク質（プロテオーム）の約半分は機能が未知のままです。藤島君は機能が既知のタンパク質に現れるアミノ酸の種類やその頻度、周期性をコンピュータ解析により、詳細に検討することで、これまでにない、新しいタンパク質の機能推定法を考案いたしました。

さらに、この方法を駆使することで、機能未知であったタンパク質の中から、核酸制御に関わる新規の遺伝子産物を推定していくことに成功いたしました。これは、学部の学生が成し遂げた成果として、目をみはるものがあります。

また、彼の研究に対する態度は極めて実直で、その解析データは信頼のおけるものです。この意味でも、今回の藤島君の研究論文を湘南藤沢学会の優秀論文として推薦いたします。

慶應義塾大学　環境情報学部助教授
（同・先端生命科学研究所助教授）
金井　昭夫

ゲノム情報を利用した新規 RNA 結合蛋白質推定とその構造パターン予測

環境情報学部　藤島　皓介

目次

I 要旨

生命情報工学（バイオインフォマティクス）を用いた新たな2つの蛋白質機能推定法に関して報告したい．ここで，実験生物学での実証を鑑み，ゲノム情報が既知であり，かつ生化学的な実験に適した超好熱性古細菌 ***Pyrococcus furiosus*** をモデル生物に選んだ．また，機能性蛋白質として新規の **RNA** 結合蛋白質を推定することを目的にした．まず，プロテオームを構成するアミノ酸の使用頻度，分子量，疎水度および特定のアミノ酸の周期性に関して解析し，これらの指標を段階的に用いることで，ある種の機能性蛋白質（特に膜蛋白質，リボソーム蛋白質，遺伝子制御蛋白質など）を分画可能なことを示す．本法を用いて，特に新規の **RNA** 結合蛋白質候補 **29** 種を挙げることが出来た．このうち **2** つの候補について遺伝子工学を用いた実験をおこない，その **RNA** 結合性を確認した．次に，蛋白質を構成するアミノ酸が有する電荷の周期性に着目した解析を行なった．この周期性をウェーブレット変換することで，無駄なノイズを除去し，電荷情報の再構築を行った．この結果，大腸菌 ***Escherichia coli*** と ***P. furiosus*** に共通に保存されている **RNA** 分解酵素 **RNase HII** において，アミノ酸配列の相同性が約 **30**％程度しかないにもかかわらず，再構築した波形が酷似した．以上は，本稿で提唱した2つの方法が蛋白質の機能推定法として極めて有用であることを示唆している．

キーワード： バイオインフォマティクス，プロテオーム，蛋白質機能予測，周期性，RNA結合蛋白質， 超好熱性古細菌

Ⅱ 序論

2. 1 背景

20 世紀の終わりから今世紀にかけて，大腸菌から酵母，線虫，マウス，ヒトに至るまで多種多様な生物の全ゲノム配列が相次いで決定された．時代は今，ポストゲノムの幕開けを迎えている．現在までに微生物と高等生物併せて 177 種，ウィルスゲノムでは約 1000 種にも及ぶ生物種の全ゲノムがデータベースとして存在している（1/31/2004 updated NCBI database）．また，この状況に伴いゲノム情報が規定している蛋白質の機能予測も重要視されてきた．従来の代表的な機能予測法は 1990 年に考案された BLAST [1]を用いて，機能既知蛋白質とのアミノ酸配列の相同性（ホモロジー）を比較し，機能を推定する方法である．また，最近では DNA ポリメラーゼサブユニットやリボソーム蛋白質群のように複数の蛋白質が協調的に作用するもの同士が似た性質，機能を持つことを利用した蛋白質間相互作用（Protein-Protein Interaction）が注目されており，データベースや viewer なども開発されている[2]．さらに，オーソログといい，進化的関連性がある遺伝子同士のデータベースも存在する．同じオーソログジーンの中には配列に相同性が見られないものもいくつか含まれているため，これもプロテオームの機能分類法として近年注目を浴びている．しかし原核生物，真核生物，古細菌にかかわらず，ゲノム上に存在する約 50%の遺伝子（蛋白質）はいまだに機能が同定されておらず，蛋白質機能推定法の大半は配列の相同性解析に依存しているのが現状である．

2. 2 新規機能推定法

我々はこれまでに配列の相同性に依存しない解析手法をいくつか模索してきた．主にアミノ酸の疎水度と分子量を用いた解析，使用頻度解析，周期性解析の３方法である．それぞれの解析はある程度の基準をもうけてやれば，機能推定に有効であることが示唆されている．したがって本稿ではそれぞれの解析について詳しく述べるとともに，３つの解析を段階的に用いることで，新規の RNA 結合蛋白質を高い精度で予測することを可能とした結果についてまとめた．また我々が提唱する 2 つ目の新規蛋白質機能推定法はアミノ酸の一次配列上に分布する電荷の増減からノイズを除去し，高次構造を予測する方法である．

2. 3 超好熱性古細菌 *Pyrococcus furiosus*

本研究に用いた生物種は超好熱性古細菌 *Pyrococcus furiosus* であり， 2001 年に全ゲノムシーケンスが公開された深海に生息する菌である．ゲノムサイズも 2Mbp と小さく（大腸菌の約半分）で，推定される遺伝子は 2065 個である．また遺伝子産物の多くが耐熱性を獲得していると考えられており，生化学的に非常に扱いやすいことが挙げられる[3][4].

2. 4 RNA と RNA 結合蛋白質

RNA は DNA から転写される核酸物質で生命現象と深い結びつきがある．RNA は遺伝子発現の段階では mRNA として DNA の転写に携わり，翻訳段階では rRNA や tRNA などが蛋白質との複合体を形成し，アミノ酸をつなげて蛋白質を合成する．近年 RNA 干渉といい miRNA や siRNA などの小さい RNA 断片が複数の mRNA に部分結合することで翻訳レベル

で遺伝子を制御する現象も発見された．このようにRNAを中心とした生命現象は今後さらに発展していく兆しをみせている．同様にRNA結合蛋白質も生体内で重要な役割を果たしていることがわかっている．例えばRNAポリメラーゼがなければRNAは伸長することができず，RNA分解酵素がなければRNAが細胞内に溢れかえってしまう．したがって我々はRNA結合蛋白質を網羅的に予測するという方向性から，RNAを中心としたダイナミックな系の解明，しいては生命の起源の謎を解き明かそうと考えている．そこで本稿では超好熱性古細菌 *P. furiosus* のゲノム情報を用いて，新規のRNA結合蛋白質推定法を確立することを具体的な目標にすえた．

Ⅲ　ゲノム情報による蛋白質機能解析

3. 1 疎水度分布解析

3. 1. 1 目的

疎水度(hydropathy)とはアミノ酸の持つ性質の一つで，1982年にKyteとDoolittleによって数値化されたことにより定量的な扱いが可能となった[5]．また，この数値は主に蛋白質折りたたみ時のトポロジー計算などに使われており[6]，重要な指標の一つである．ここで生体で用いられている20種のアミノ酸のうち9種が疎水性アミノ酸である(TABLE1).

アミノ酸	3文字略号	1文字略号	側鎖、性質	分子量	等電点	Hydropathy
グリシン	Gly	G	脂肪族、親水性	75	5.97	−0.4
アラニン	Ala	A	脂肪族、疎水性	89	6.00	1.8
バリン	val	V	脂肪族、疎水性	117	5.96	4.2
ロイシン	Leu	L	脂肪族、疎水性	131	5.98	3.8
イソロイシン	Ile	I	脂肪族、疎水性	131	6.02	4.5
プロリン	Pro	P	イミノ酸、疎水性	115	6.30	−1.6
セリン	Ser	S	水酸基、親水性	105	5.68	−0.8
スレオニン	Thr	T	水酸基、親水性	119	6.16	−0.7
システイン	Cys	C	含硫、親水性	121	5.07	2.5
メチオニン	Met	M	含硫、疎水性	149	5.74	1.9
アスパラギン酸	Asp	D	酸性、親水性	133	2.77	−3.5
アスパラギン	Asn	N	アミド、親水性	132	5.41	−3.5
グルタミン酸	Glu	E	酸性、親水性	147	3.22	−3.5
グルタミン	Gln	Q	アミド、親水性	146	5.65	−3.5
リジン	Lys	K	塩基性、親水性	146	9.74	−3.9
アルギニン	Arg	R	塩基性、親水性	174	10.76	−4.5
ヒスチジン	His	H	塩基性、親水性	155	7.59	−3.2
フェニルアラニン	Phe	F	芳香族、疎水性	165	5.48	2.8
チロシン	Tyr	Y	芳香族(水酸基)、疎水性	181	5.66	−1.3
トリプトファン	Trp	W	芳香族、疎水性	204	5.89	−0.9

TABLE 1: **各アミノ酸についてのパラメーター詳細**

慶應義塾大学先端生命科学研究所の鈴木治夫氏はアミノ酸の変動の第一要因が疎水度だということを，主成分分析を用いて導きだした[H. Suzuki, unpublished work]．そこで我々は蛋白質の機能が疎水度に一部依存しているのではないかという仮説を立て，疎水度と分子量の相関図に機能既知蛋白質をプロットすることで機能ごとに分類される領域があるか調べた．

3. 1. 2 方法

まずNCBIで公開されている超好熱性古細菌 *Pyrococcus furiosus* (NC_003413) のgenbank fileから全蛋白質2065個のアミノ酸配列を抽出した．予想される機能の内訳をTABLE2に示す．次に機能アノテーションにConserved hypothetical（機能不定）が付いているものをすべて取り除いたところ，1031個の蛋白質が残った．1031個の蛋白質それぞれについて疎水度（Hydropathy）と分子量を算出した．さらに疎水度を縦軸に，分子量を横軸に用いた2変量の相関グラフを作成し，1031個の蛋白質をすべてプロットした．比較のため，代表的な原核生物3種類についても同様の解析を行った．さらに，似た機能を持つ蛋白質同士がグラフ上にどのように分布しているかを調べた．

機能群	個数
Amino Acid Biosynthesis	91
Autotrophic Metabolism	3
Biosynthesis	56
Cell Envelope	45
CellularProcesses	57
Central Intermediary	51
Conserved Hypothetical	1024
Energy Metabolism	120
Fatty Acid	12
Other Categories	48
Purines,Pyrimidines	45
Regulatory functions	51
Replication	39
Transcription	39
Translation	157
Transport	132
Unknown	95
Total	2065

TABLE 2: *P. furiosus* の全蛋白質の機能内訳

3. 1. 3 結果

図1(a)に示すように1031個の蛋白質がやじりの形に分布する特徴的なパターンが，種を超えて保存されていることが確認された．このうち *P. furiosus* の解析結果を用いて複数の蛋白質群について分布領域を調べた結果，幾種類かの機能蛋白質はグラフ上の特定の領域に集合することを見出した．図1(c)に傾向が顕著に見られた蛋白質群9種を示す．また，一部重なってはいるものの，明らかに異なる分布を示した膜蛋白質，RNA結合蛋白質，リボソーム蛋白質の3種について図1(b)に示した．

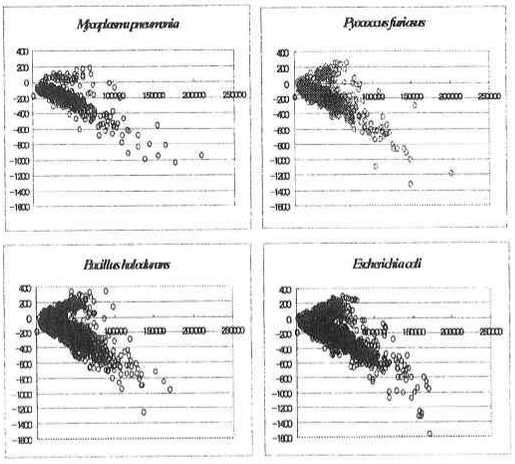


図1(a) 疎水度/分子量プロット(多種間比較)

生物種4種に対して疎水度/分子量プロットを描いた

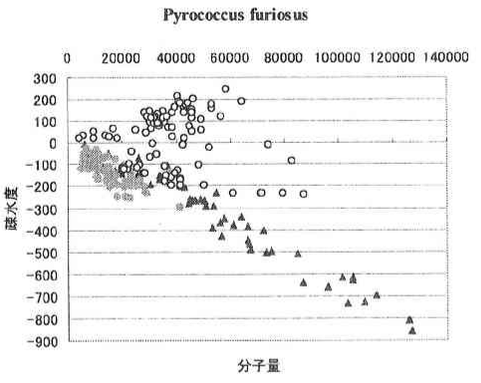


図1(b) 疎水度/分子量プロット(機能群3種)

P. furiosus において領域の異なる3種の蛋白質群

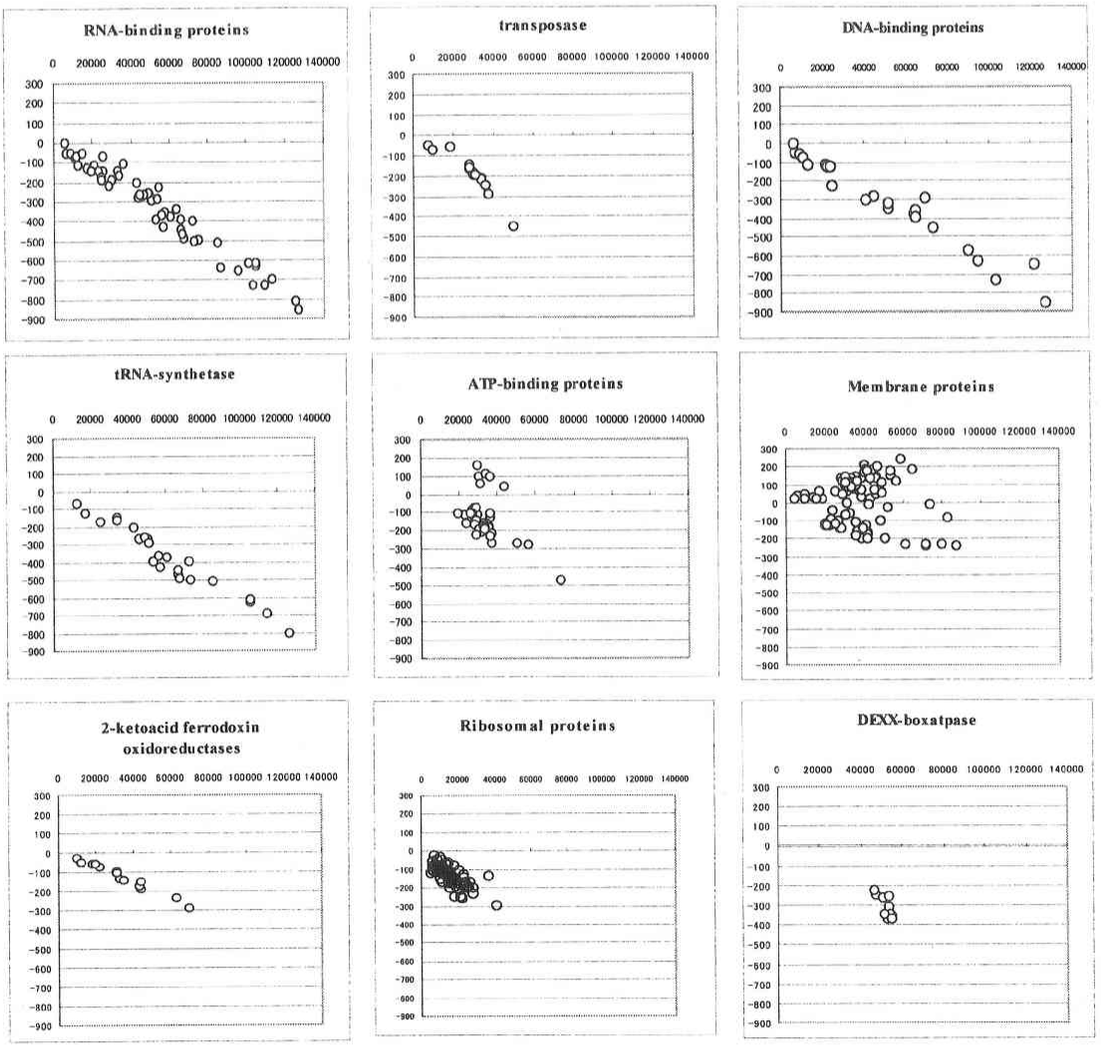


図 1.(c) 疎水度/分子量プロット(機能群9種)

9種類の蛋白質群について横軸に分子量, 縦軸に疎水度をとったグラフを同スケールで表示した.

3. 1. 4 考察

これらの結果から機能既知蛋白質を疎水度/分子量で表した分布図が種を超えて保存されていることが確認された．*P. furiosus* において蛋白質を機能群ごとに分類した結果，少なくともある種の蛋白質群（膜蛋白質，リボソーム蛋白質など）に関しては特定の領域に集合した．このことは蛋白質の有する疎水度はその生物が生育する環境や種に依存せずに，機能分類のための指標となることを示唆している．

3. 2 アミノ酸頻度解析

3. 2. 1 目的

アミノ酸頻度解析の歴史は深く，1983 年に 300 個あまりの蛋白質をアミノ酸使用頻度によって分類することが行われていた．しかし当時は機能が既知なサンプルが非常に少なく，大まかに 4 つの機能群に分類するのが精一杯であった[7]． 我々はアミノ酸 20 種類の使用頻度を用いて，*P. furiosus* における全蛋白質クラスタリングを行い，より詳細な機能分類を目指した．

3. 2. 2 方法

本解析には前述の疎水度分布解析(3．1)で用いた *P. furiosus* の 1031 個の蛋白質を使用した．1 つの蛋白質につき，アミノ酸 1 種類の数をアミノ酸の総数で割り，20 種類すべてのアミノ酸について使用頻度を算出した．しかし蛋白質を構成する２０種類のアミノ酸は必ずしも均等に使われているわけではない．そこで各々のアミノ酸の使用頻度と全蛋白質の平均との差をもってアミノ酸使用頻度と定義した．さらに 20 種のアミノ酸のスコアを一つの蛋白質が持つ要素とし，1031 個の蛋白質に対して米 Stanford 大学． の Eisen 研究室が開発したクラスタリングソフトウェア Cluster と，解析結果を視覚化するソフトウェア Treeview を用いて階層的クラスタリングを行なった．

3. 2. 3 結果

1031 個の蛋白質による階層的クラスタリングの結果，tRNA-synthetase， リボソーム蛋白質，transposase， DEXX-box ATPase は一群にクラスタリングされることが明らかになった．そのうちの 2 クラスタを図 2．a に示す．一方で ATP 結合蛋白質や RNA 結合蛋白質などは複数箇所で小クラスタを形成した．まず膜蛋白質は電荷を持ったアミノ酸 D，E，R，K の使用頻度が極端に低かった．中でもグルタミン酸（E），リジン(K）が特に顕著であった(図 2．b の 1 レーン目)． リボソーム蛋白質は塩基性アミノ酸であるアルギニン（R），リジン(K)を極端に好んで使用している半面，疎水性アミノ酸のロイシン（L），イソロイシン（I）の使用頻度は極端に低かった(図 2．b の 2 レーン目)． RNA に結合する蛋白質は総じて電荷を持つアミノ酸を多く持つ傾向にあった．特にリジン（K）の割合が高かった．一方，同じ正電荷アミノ酸であるアルギニン(R)は全蛋白質平均と同程度の使用頻度だった(図 2．b の 3-6 レーン目)．

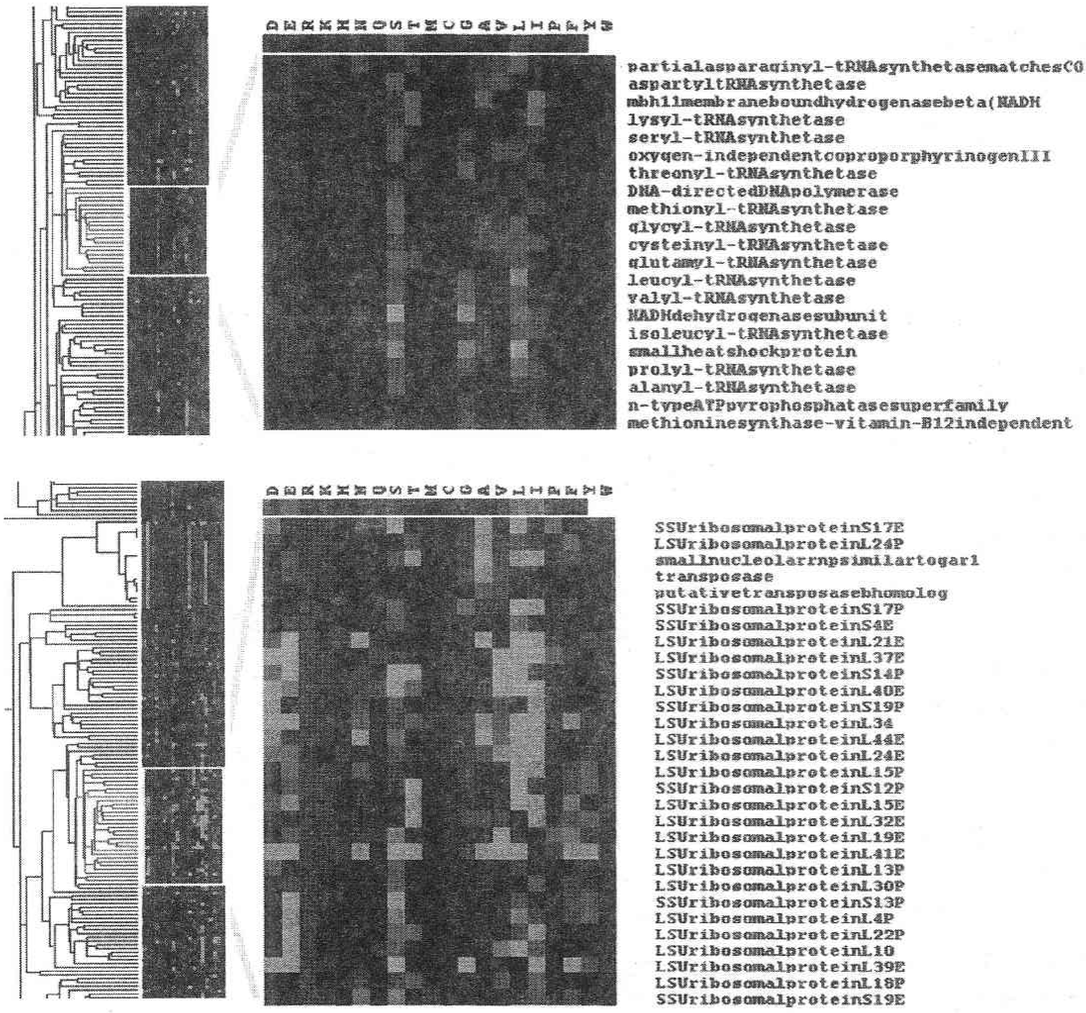


図 2.a 機能別アミノ酸頻度クラスタリング（顕著なもの）

図の上半分は tRNA-synthetase 群を，下半分はリボソーム蛋白質群がクラスタを形成していることを表している．

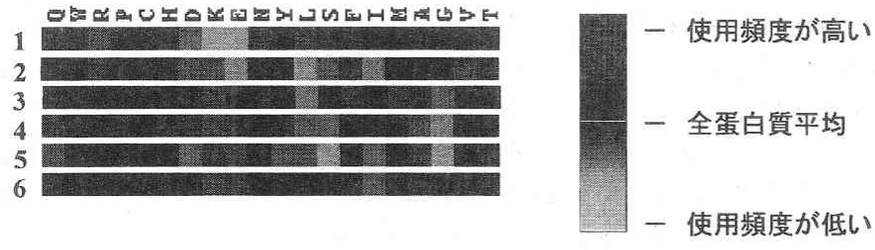


図 2.b 機能別アミノ酸頻度クラスタリング

20種類のアミノ酸を横軸にとり，それぞれの使用頻度を色で示した．赤いマスは該当するアミノ酸が全体平均よりも高い頻度で使われていることを示し，緑のマスは逆に使用頻度が低いことを示している．1 Membrane Protein 2 Ribosomal Protein 3 RNase HII 4 mRNA end-processing factor 5 DNAdirected RNA-polymerase subunitl, 6 tRNA nucleotidyl transferase

3. 2. 4 考察

アミノ酸使用頻度の偏りを 1 つの指標として捉えることで，また別の切り口からの蛋白質機能推定が可能となった．リボソーム蛋白質を生化学的な分野からみると負電荷をもつ RNA に効率よく結合するために正電荷を豊富に含んでいると理解することができる．しかし同じ RNA 結合特性を持つ蛋白質でも，DNA 依存で RNA を伸張させるもの，RNA を分解するもの，複数の RNA をつなげるものなど細胞内で広い用途で使用されているがために，その多様性がゲノム配列，しいてはアミノ酸の頻度の差を生み，小クラスタが複数できたと考えられる．しかし一方でリジン（K）のような正電荷アミノ酸が RNA 結合蛋白質全般にわたって強く保存されていることを確認することができた．これらのアミノ酸が一次配列上でどのように配置されているかを調べることも重要である．

3. 3 周期性解析

3. 3. 1 目的

慶應義塾大学先端生命科学研究所の金井昭夫氏らはある種のＲＮＡ結合蛋白質にはその一次配列上に電荷を有したアミノ酸残基が周期的に出現することに気がついた[8]．そこで我々は電荷を有したアミノ酸の周期性をプロテオームレベルで解析すれば新規の RNA 結合蛋白質候補を見出せると考えた．本研究の目的は周期を抽出するような独自の解析プログラムを作成し，その周期を指標としたクラスタリングで RNA 結合蛋白質を抽出することである．

3. 3. 2　アミノ酸周期性探索プログラム(APS)の作成

この目的のため，まず，アミノ酸周期性探索プログラム（**A**mino acid **P**eriodicity **S**earch）を作成した．**APS** は指定した蛋白質，アミノ酸，周期にマッチしたものを自動的に抽出する．Output として周期の長さ，配列上のポジションを，オプションとして配列の長さ，アミノ酸配列のカラー表示などを行うことが可能である．このプログラムは以下のように実行する．[ ]の中はユーザーが指定する条件である．

. /period_search. pl [FILE 名] [アミノ酸] [周期] [蛋白質名 or 遺伝子番号] [誤差]

[FILE 名]･･･現在は genbank file から生成したアミノ酸配列のリストを与えているが，近いうちに genbank file で実行可能にする予定．

[アミノ酸]･･･周期を見たいアミノ酸を指定する．例えば電荷を有するアミノ酸 D,E,R,K.H など

[周期]･･･主に 3 から 20 の間で行なう．両親媒性アルファ-ヘリックス上での周期ならば 3.6 ピッチ，リピートドメインなどの場合には 9～20 程度の周期で探索するのが良い．

[蛋白質名 or 遺伝子番号]･･･特定の遺伝子群を網羅的に調べるなら機能アノテーションをいれればよい．特定の蛋白質だけを解析したければその遺伝子番号を入れる．

[誤差]･･･通常は 1. 必ずしも特定のアミノ酸が正確な周期で存在しているわけではないのでそれを考慮にいれた誤差のことである．

3. 3. 3　方法

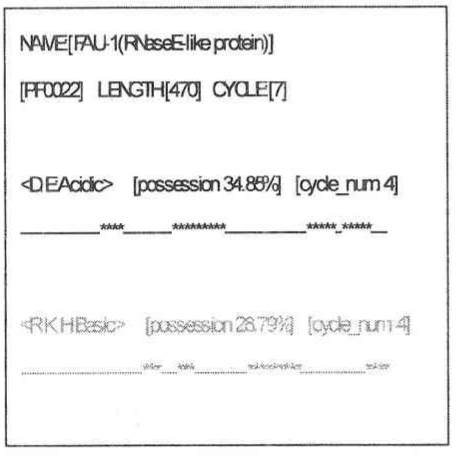
上記のプログラムを用い，まず 1031 個の機能既知蛋白質について酸性/塩基性アミノ酸の出現する周期が 3-28 までのもの (誤差 1). という条件で解析し，周期が存在する部分の割合(これを possession と定義する)を算出した. 例えばアルファ-ヘリックスという構造においては，アミノ酸平均 3．6 残基で構造的に一周期を成すことがよく知られている．本解析に用いた 3-28 周期というのはアルファ-ヘリックスが持つそれの約 8 倍長までに相当する．また *P. furiosus* における平均的な蛋白質は約 282 アミノ酸残基により構成されることを考慮すれば，その長さの約 1/10 までの周期をここでは検討することになる.

さらに新規の RNA 結合蛋白質候補を抽出するため，機能既知蛋白質ばかりでなく Conserved hypothetical protein を加えた全 2065 個の蛋白質に対して同様の解析を行った．ここで，アミノ酸頻度解析と同様に，大規模なクラスタリングでは抽出したい対象が複数のクラスタに分散する傾向がある．そこで類似度が高いものから中心近傍にそろえる centered というオプションを使用した.

3. 3. 4 結果

本プログラムを RNA 結合蛋白質 FAU-1(RNaseE like protein)[8]に対して行った結果を図 3. a に示す．ここでは 7 cycle で解析した結果について示した．図中の酸性アミノ酸と塩基性アミノ酸のクラスタ領域が可視化されているのがわかる.

RNA/DNA 結合蛋白質が中心領域に集まってクラスタを形成した. 核酸結合蛋白質のクラスタ分布をヒストグラムで確認したところ，77 個の RNA 結合蛋白質のうち 30 個が中心領域に存在していることが明らかとなった．DNA 結合蛋白質もほぼ同じ形で分布しているものの，RNA 結合蛋白質に比べて中心近傍に収束する強い傾向は見えなかった(図 3.b).



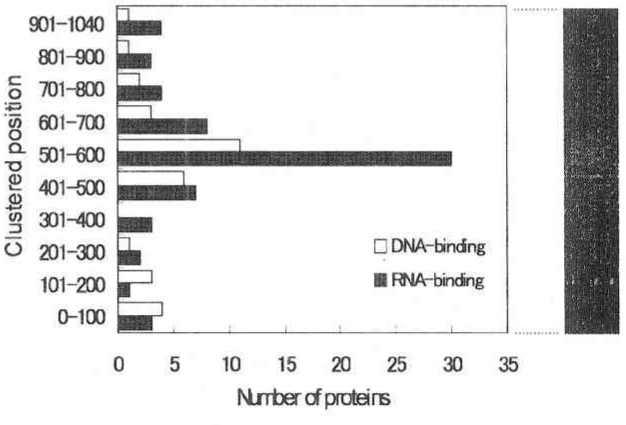


図 3.a FAU-1 蛋白質の APS 実行結果　　図 3.b 核酸結合蛋白質の局在を表すヒストグラム

図 3. a 金井氏らによって新規に発見された RNA 結合蛋白質 FAU-1 の Amino acid Period Search プログラム実行結果．出力結果は上から順に名前，遺伝子番号，長さ，周期，アミノ酸，配列上に占める割合，周期が存在する箇所の数，そして一次配列上でのポジションである．周期 7 の酸性と塩基性の領域が交互に存在していることが確認できる.

図 3. b 縦軸に機能既知蛋白質 1031 個が並んでおり，501-600 の中心領域に 30 個の RNA 結合蛋白質が集合した．中心領域の RNA 結合蛋白質の数は他の領域とくらべて 5～10 倍の数存在する．一方，DNA 結合蛋白質は2～7倍程度であった.

図 3.c 全蛋白質の周期性を用いたクラスタリング結果

左の棒はクラスタリングされた 2065 個の蛋白質を列挙したもの. 右側の蛋白質名は中心近傍を拡大したものである. 機能未知蛋白質は hypothetical protein として表示している. 赤の下線が RNA 結合蛋白質, 黄緑の下線が DNA 結合蛋白質. 丸い記号がついている3つの蛋白質は実験で検証したもの.

3. 3. 5 考察

本解析においてRNA結合蛋白質とリボソーム蛋白質に関しては区別化に成功したといえる．使用しているアミノ酸及びその周期パターンに大きな違いが見られたため，完全に別のクラスタを形成したと考えられる．一方，RNA と DNA 結合蛋白質に関しては完全な分類は不可能だった．RNA と DNA 両方とも核酸であるため，負電荷を帯びているので，認識する蛋白質の性質も類似してくるものと思われる．また中にはRNaseHII のようにDNA 依存のRNA 結合蛋白質も存在するため，厳密な区別をつけるのは難しい，しかしそれでもRNAだけを認識するドメインが存在するように，RNA に特化した周期的なアミノ酸の配置が存在していても不思議ではない．本解析はそのような隠れたルールの一部を浮き彫りにしたと考えてよいと思われる．

機能推定に用いた際にはcentered オプションは避け，周期データも3-20 に減らした．過剰な周期はかえってクラスタリングの際に逆効果になるからである．また，配列長が近いもの同士でクラスタリングさせることで誤差を失くそうと考えた．さらにアミノ酸頻度とpossession を同時にクラスタリング要素としてした場合，互いに傾向を打ち消しあってしまうことが確認された．以上の理由からそれぞれの解析を段階的に使用する新規RNA 結合蛋白質抽出法を考案した．

3. 4 統合解析

3. 4. 1 目的

先の疎水度分布解析(4．1)，使用頻度解析(4．2)，周期性解析(4．3)はそれぞれ蛋白質機能分類に対してある程度の基準を設けてやれば有効であることがわかった．全蛋白質を網羅的に扱うプロテオーム解析においては個々の解析手法を単独で使用するよりも，それらを統合することで新規のRNA 結合蛋白質を見つける可能性が高いと考えた．そこで我々は先の3 種の解析手法を段階的に用いることで新規のRNA 結合蛋白質候補の抽出を行った．

3. 4. 2 先の解析との変更点

疎水度分布解析についてはRNA 結合性を有する候補を高い精度で抽出するため，疎水度あたりの分子量（相関プロット上の傾き）を数値化した．そして転写翻訳に関わるRNA/DNA結合蛋白質群の値でヒストグラムを作成し，機能未知蛋白質をふるいにかけるための閾値を定めた．

アミノ酸頻度解析ではRNA 結合蛋白質に特化した解析を行うため，アミノ酸を20 種類から 7 つのグループに分類した．アミノ酸を性質別に分類することでアミノ酸配列から機能に関わる部位を高い精度で予測したという研究が報告されている[9]．そこで我々は各グループごとに使用頻度を算出し，その値をクラスタリング要素として用いた．

周期性解析は周期3-20 で行った．またクラスタリングは誤差を最小限に抑えるために，アミノ酸配列長の近いもの同士で行った．

3. 4. 3 手法 ～新規 RNA 結合蛋白質の抽出～

3. 4. 3a データセット

蛋白質の機能アノテーションにputative，unknown，Conserved hypothetical がついているもの

をすべて除去した842個をモデルデータセットとした．機能アノテーションにputativeと記述されている蛋白質には予測された蛋白質の機能名が付加されているが，信頼性が低いため本解析では扱わないことにした．また候補蛋白質抽出のために1024個の機能未知蛋白質を用意した．

3. 4. 3b 疎水度分布解析を用いたフィルタリング

まず機能アノテーションにTranslationもしくはTranscriptionと明記されている転写・翻訳制御蛋白質の合計196個がRNA結合蛋白質の疎水度特性を十分に模倣していると仮定した．次に196個の蛋白質に対して疎水度/分子量のスコアを用いたヒストグラムを作成する．最終的に196個の蛋白質の約9割を内包する値を上限と下限に閾値として設定し，候補蛋白質をフィルタリングした．

3. 4. 3c アミノ酸頻度解析を用いた候補の絞込み

20種類のアミノ酸を次の性質の異なる7グループに分類した．DE-酸性，RKH-塩基性，NQ-アミド，FYW-芳香族，GAVLIP-脂肪族，ST-水酸基，CM-含硫
それぞれの使用頻度を計算し，群平均法を用いて先の解析でフィルタリングされた機能既知蛋白質に対してクラスタリングを行った．既知のRNA結合蛋白質と隣接する機能未知蛋白質は新規のRNA結合蛋白質の可能性が高いと仮定し，周期性解析用のデータとして抽出した．その際，解析の信頼性を調べるためにtrue positive，false negativeを算出した．

3. 4. 3d 群平均法

群平均法とはクラスタ間の非類似度を対象間の平均的な値で定義しようという考え方である．メリットとしてリボソームや膜蛋白質のようにあるアミノ酸を多量に持つ特異的な蛋白質が多少そのクラスタに含まれてようと，それら少数の"はずれ値"によってクラスタの傾向が変わる心配がない．逆に最短距離法や最長距離法はクラスタ内の対象間の類似度の最も高い値，低い値を用いるため，極端な値を持った蛋白質が存在する場合，その蛋白質の傾向が色濃くでてしまう可能性がある．

3. 4. 3e 周期性解析を用いた最終候補の抽出

アミノ酸頻度解析によって抽出した候補蛋白質の精度を高めるために周期性解析で最終候補を選択した．周期性解析(4. 3)で用いたpossessionを配列長で割ることで正規化を行った．さらにアミノ酸配列が200-300のようにアミノ酸100残基ごとにクラスタリングを実行することで誤差を最小限に抑えるようにした．最終的に既知のRNA結合蛋白質に隣接するようにクラスタリングされた未知の蛋白質を最終的に新規のRNA結合蛋白質候補として選択した．

3. 4. 4 結果

*P. furiossus*のゲノム中に存在した196個の転写・翻訳制御蛋白質は疎水度/分子量のヒストグラムから，その値が-0．506～0．232の範囲に約88%存在することが明らかとなった．そこでこの範囲内に該当する機能未知蛋白質はRNA結合性を有する可能性が高いと仮定しフ

ィルタリングを行った．その結果 268 個が候補外となり 1024 個から 268 個を除去した 756 個の蛋白質をアミノ酸頻度解析用のデータセットとして残した．機能既知蛋白質も 128 個除外され，714 個が残った．除外された蛋白質の大半は Transport protein(輸送蛋白質，膜状に多数存在する) が占めていた．

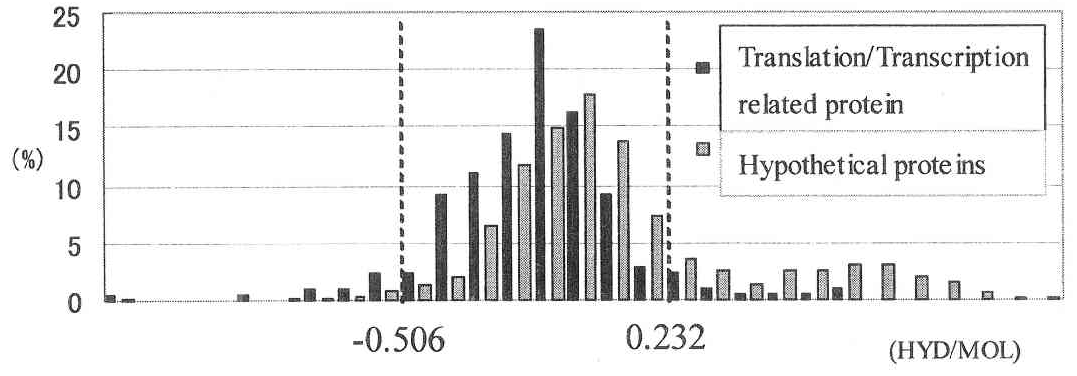


図 4. 疎水度/分子量スコアのヒストグラム

濃い棒が転写翻訳制御蛋白質，薄い棒が機能未知蛋白質である．点線に挟まれた範囲が転写翻訳制御蛋白質の約 88%が存在している領域．

残った機能既知蛋白質 714 個を群平均法によって階層クラスタに分割したところ，転写翻訳関連蛋白質が偏っている大クラスタが存在することが確認できた．機能未知蛋白質 756 個を含めた 1478 個の蛋白質に対して同様のクラスタリングを行った結果，RNA 結合蛋白質と同じクラスタ内に複数の機能未知蛋白質が集合したため，それらの中から特に相関が高いもの 133 個を抽出した．解析結果の信頼性は以下に示す(TABLE3).

Protein Function	Total			Sensitivity(%)	Cluster node	Possibility(%)
RNA-binding	77	true positive	61	79.2	225	61/225 = 27.1
		false negative	16	20.8		
DNA-binding	31	true positive	23	74.2	225	23/225 = 10.2
		false negative	8	25.8		
Ribosomal	48	true positive	28	58.3	38	28/38 = 73.6
		false negative	20	41.7		

TABLE 3: クラスタリング結果の信頼性（蛋白質機能別）

* true positive: 本解析によって同機能群が集合したクラスタに属し，且つ RNA 結合性を持つ蛋白質の個数．
* false negative: 本解析によって同機能群が集合したクラスタに属し，且つ RNA 結合性を持つ蛋白質の個数

最終的に 133 個の候補蛋白質に対して先に述べた条件(3．4．3e)で周期性解析を行なったところ，既知の RNA 結合蛋白質と類似性の高い周期を持つ蛋白質が 29 個確認されたので，それらを最終的な候補としてリストアップした．(TABLE 4)

Gene number	Length (aa)	Predicted Protein Function*
PF0007	496	tRNA modification
PF0029	478	tRNA synthetase
PF0037	565	RNA processing
PF0051	220	translation initiation factor
PF0269	89	translation elongation factor
PF0470	114	transcriptional regulator
PF0493	660	tRNA synthetase
PF0545	98	protein translation factor
PF0590	243	RNA processing
PF0656	133	DNAdirectedRNApolymerase
PF0741	101	translation initiation factor
PF0838	84	DNAdirectedRNApolymerase
PF0859	253	ribosomal protein modofication
PF0908	280	tRNA synthetase
PF0979	126	translation initiation factor
PF1129	872	DNAdirectedRNApolymerase
PF1307	140	translation initiation factor
PF1310	497	RNA processing
PF1312	236	Ribonuclease
PF1353	74	protein translation factor
PF1483	243	tRNA synthetase
PF1570	237	tRNA synthetase
PF1577	416	tRNA synthetase
PF1588	208	translation initiation factor
PF1724	132	translation elongation factor
PF1725	448	tRNA synthetase
PF1894	101	rRNA-binding
PF1912	426	RNA/DNA replication
PF1915	391	translation initiation factor

*近接した既知蛋白質の機能

TABLE4: **新規 RNA 結合蛋白質の候補**

3. 5 実験的検証

我々は先の周期性解析(3．3)で挙げた RNA 結合性を有するであろう候補蛋白質のリストの中からランダムに 3 つを選び，分子生物学的な検証を行った．その候補は PF0029，PF0547，PF1912 の各蛋白質である．そのうち PF0029，PF1912 は統合解析(3．4)の最終候補に選ばれている．本検証は慶應義塾大学先端生命科学研究所助教授の金井昭夫氏と同技術員/非常勤講師である佐藤朝子氏によって行われた．

3. 5. 1 組換え体蛋白質の発現と精製

まず，*Pyrococcus furiosus* のゲノム DNA から，目的遺伝子を PCR 法により増幅し，pET-23b 発現ベクターにサブクローニングした．このベクターは蛋白質 C 末端側に 6 つのヒスチジン(His-tag)を付加することが可能であり，組換え体蛋白質の精製の際にアフィニティタグとして用いることが出来る．組換え体蛋白質はファージ T7 プロモーターの制御下に *E. coli* BL21(DE3)株にて過剰発現させ，Ni2+-Sepharose カラムを用いることで，SDS ポリアクリルアミドゲル電気泳動上で主な成分になるまで精製した（図 5）．

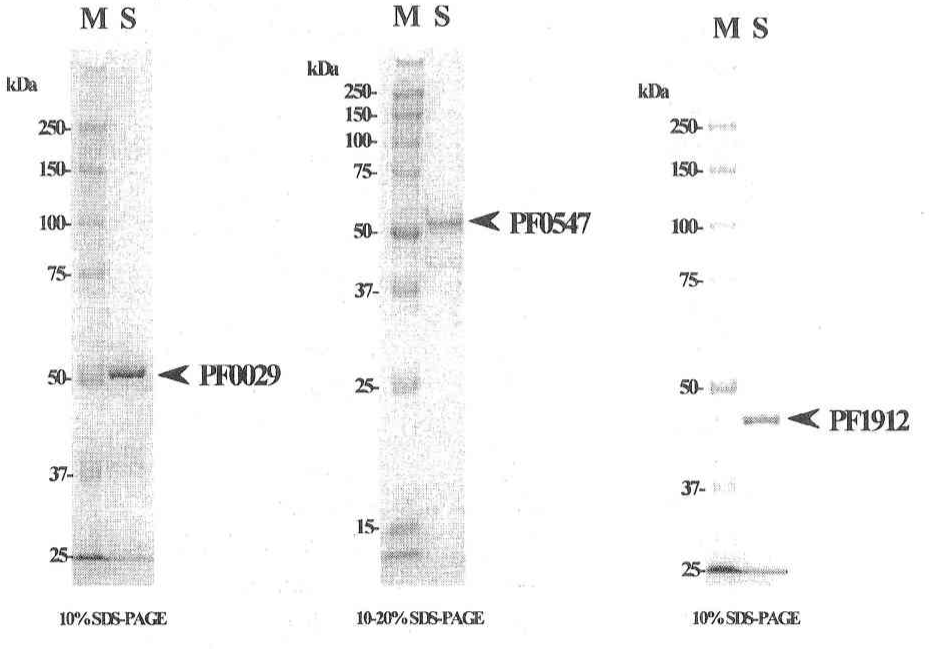


図 5. 精製した候補蛋白質の SDS-ポリアクリルアミドゲル電気泳動による確認

CBB 染色により蛋白質を検出した．目的の蛋白質を矢頭にて示す．

3. 5. 2 ゲルシフト法による RNA 結合性の解析

精製した組換え体蛋白質が RNA 結合性を有するかどうかに関して，ゲルシフト法を用いて検討した．この際，ステムループ構造を持つ合成 RNA である 5'-FAM-end-labeled オリゴヌクレオチド S-5 をプローブとして用いた．この S-5 プローブと精製蛋白質とを 75℃にて 15 分間インキュベートすることで RNA-蛋白質複合体の形成反応を行なった．その結果，少なくとも PF0029 と PF1912 の 2 種の蛋白質においては，精製した蛋白質の容量依存的に未変

性ゲル電気泳動上でシフトしたバンドが検出され，これらの蛋白質はRNA結合活性を有すると結論した(図6).

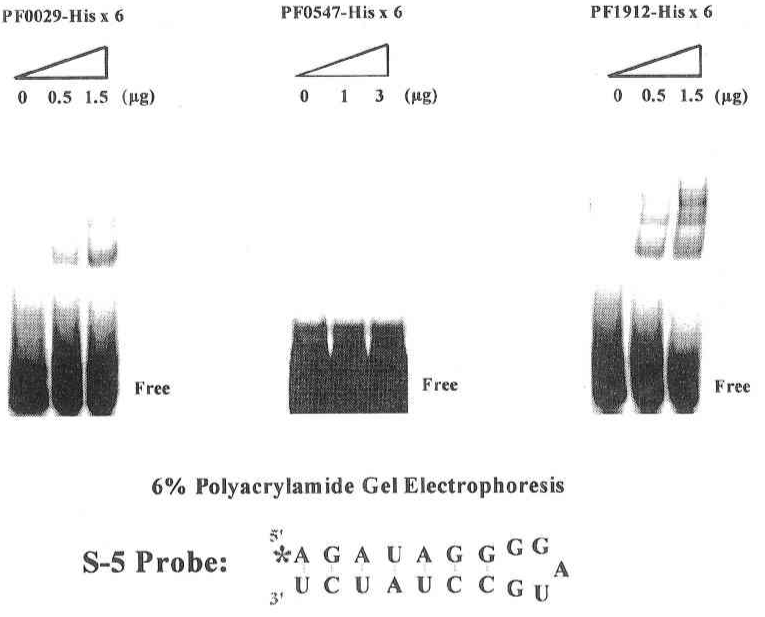


図6. ゲルシフト法によるRNA結合性の確認

図の下部分にあるループ構造を持つRNAがプローブ．RNA結合性をゲルシフト法で確認した．PF0029, PF1912の2つの蛋白質がRNA結合性を示した．

3. 5. 3 考察

実験的検証の結果から，PF0029， PF1912二つの蛋白質に関してS-5プローブの濃度を0～1. 5μgと増やしていくにつれてシフトしたバンドが見受けられた．このことは蛋白質がRNA結合性を有することを強く支持する結果である．PF0547はRNA今回の実験ではRNA結合性を確認することができなかった．だが今回用いたプローブとは別のRNA構造を認識している可能性もある．したがって今回調べた例数は少ないものの，本方法が新規RNA結合蛋白質を推定する上で非常に有効な手段であることを実験的に立証することに成功した．

Ⅳ　ウェーブレット変換を用いた新規機能推定法

4. 1 概論

これまでの章ではアミノ酸の疎水度，頻度そして周期性と機能の関連性について述べてきた．ここでは周期性に特化し，配列上に存在する様々な周期を，数学的手法を用いて解く方法について述べる．

フーリエはすべての周期的な信号はフーリエ級数というサイン及びコサインの和に分解

して表現することが可能であることを19世紀に証明した．また分解した成分から元の信号を再構築することも可能である．従ってフーリエ変換により，周期的な時系列データを振動で表現することが可能となった． フーリエ変換は波長をすべて正弦波に分解するため，定常的な信号ほど分解しやすく，高い効果を発揮する．一方で短い信号，特にノイズが混じった非定常的な信号を正弦波で表示するには限界があり効果が薄い．この問題を解決するためにウェーブレット変換を用いることが可能である．ウェーブレット変換の強みはサイン，コサインの代わりに独自の分析関数を定めるところにある．つまり一つの波形が与えられた時にそれを表現することができる分析関数（基底ウェーブレット）を定め，それを用いて波形を表現することができる．ウェーブレット変換には１０種類以上の基底ウェーブレットが用意されており，分析したい波形にあわせて選択することができる利点がある．さらに基底ウェーブレットを伸ばしたり縮めたりすることで低周波から高周波まで幅広い域（スケール）において詳細な分析が可能である[10]．従って周期性解析(4.3)を本方法で解析することで，よりノイズの少ない周期的特徴を抽出することができると考える．

4. 2 アミノ酸機能予測への応用

4. 2. 1 生物学的アプローチ

ここで蛋白質のアミノ酸一次配列において特定のアミノ酸が周期的に存在する場合，それは配列が周期的信号を保有しているとみなすことができる．したがって，アミノ酸配列における空間振動数をフーリエ変換で導出することができるということになる．実際にフーリエ変換はDNAレベルでクロマチン構造のようなフラクタル性の解析[11]やリピート配列の検出に用いられている[12]． またプロテオームレベルでも膜蛋白質の膜貫通領域の両親媒性アルファ-ヘリックスの予測などに用いられている[13][14]． ここ数年，数多くの蛋白質の立体構造が同定されるにつれ，一次配列の相同性ではなく構造上の類似性によって機能を推定する手法も考案された[15]． これからは一次配列データからより多くの情報を抽出する方法の開発が望まれる．しかし蛋白質の立体構造をアミノ酸の一次配列から予測することは困難であるといわれている，それは蛋白質の取り得るコンフォメーション空間が膨大であるからである(Levinthal's paradox)． そこで本稿では立体構造を正確に予測するのではなく，電荷の増減で表現される構造特性の類似性から機能予測をおこなった．

4. 2. 2 先行研究と問題点

アミノ酸の一次配列から蛋白質の高次構造を予測する手法としては，これまでにアミノ酸配列に相同性を有する蛋白質同士を多種間で比較することにより種を超えて保存されている部位を特定し，機能に関連する構造を予測するという研究が報告されている[16]． 同論文中ではマルチプルアライメントを用いることで複数のアミノ酸配列を相同性を有する部位が最大限マッチするように列挙し，次に情報の複雑度を表現するシャノン・エントロピーを用いて各ポジションごとのアミノ酸残基の多様性を数値化する．最終的に離散フーリエ変換を用いて数値の増減から構造パターンを抽出するという手法である．しかし問題点として*P. furiosus*をはじめとする好熱性古細菌のアミノ酸配列は原核生物や真核生物に比べ特徴的なアミノ酸の使い方をしていることが確認されており[17]，大量の類似配列を用いるような多種間比較は困難である． 従って単一のアミノ酸配列から構造にかかわる情報だ

けを抽出する新規の手法が求められている.

4. 3 目的

本研究では単独のアミノ酸の一次配列から構造パターンを予測するため，ウェーブレット変換を解析に用いた独自の手法を考案した．解析対象として先の周期性解析(4．3)では有意にクラスタリングされなかった RNA 結合蛋白質 RNaseHII に対して構造に関わる特徴の抽出を試みた．*P. furiosus* の RNaseHII は実験的に機能が特定されており[18]，　将来的に構造と機能の関係を調べる上で非常に有効だと考えた．また，抽出した構造特徴が一次配列の相同性に依存しないことを証明するために，*E. coli* の RNaseHII に対しても同様の解析を行なった．*E. coli* と *P. furiosus* の配列はアライメント結果で 30%程度の類似性しかないことが知られている. 本解析には解析ソフト SPLUS のウェーブレット変換モジュールを使用した.

4. 4 解析手法

RNA 結合性を有する蛋白質の立体構造の表面には広域にまたがって正負の電荷ポテンシャルが存在し，それらが機能に深く関わっていることが知られている[19]，そこで RNA 結合蛋白質の一次配列上での電荷の分布を用いて蛋白質の高次構造の情報を抽出する手法を考案した．以下にその解析手順を示す.

まず *P. furiosus* と *E. coli* について RNaseHII のアミノ酸配列を用意する．そしてそれぞれの蛋白質についてアミノ酸が持つ静電ポテンシャル（等電点）の値をアミノ酸残基ごとにプロットし，グラフを作成する.

グラフの波形データに対してフーリエ変換を行い，特徴的な周期が存在するか確認する.

＊アミノ酸の電荷の増減は非定常的な信号であり，配列長が短いため，ここでは白色雑音（波形が特異的なピークをまったくもたない，限りなくフラットにゆらいでいる状態）ではないことだけ確認できれば良い.

ピークが多数存在することを確認し，ウェーブレット変換によって 5 段階のスケール（解像度）ごとに波形を分解し，ウェーブレット係数を得る.

係数のピークの大きさはすなわちその信号への影響力を表す．従ってウェーブレット係数が大きいもの上位 10 個だけを残し，それ以外をノイズとしてすべて取り除く．最終的に残った係数で元の波形を再構築することで隠れていた高次構造の波形を浮き彫りにする.

4. 5 結果

P. furiosus 及び *E. coli* 由来の RNaseHII に関して本解析を行った．まず，各蛋白質におけるアミノ酸の静電ポテンシャルの分布を図 7 に示す. 両蛋白質ともに多数のピークが存在し，これらのパターンから共通性や異質性を抽出するのは困難であった．また，電荷の増減に周期性が存在するかどうかをフーリエ変換によって確認した結果，周期的に現れる電荷のピークが多数存在することは確認できたが，構造に関わるような特徴を抽出することは出来なかった（data not shown）.

そこで，フーリエ変換の代わりにウェーブレット変換を用いて波形の再構築を行なったところ，再構築された波形が両蛋白質において酷似することが確認された（図 8）.

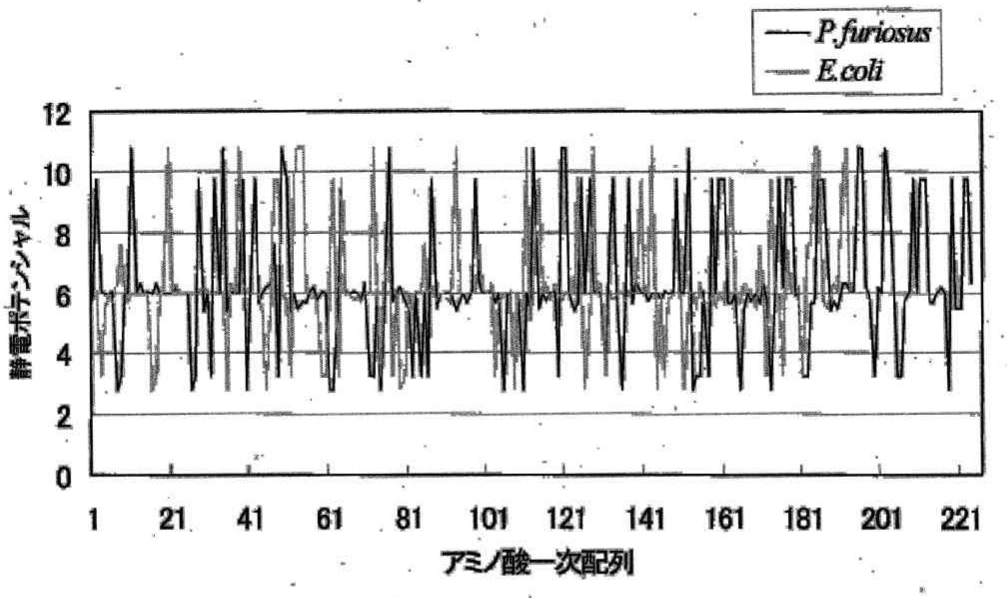


図 7.. 2 種間での RNaseHII のアミノ酸一次配列における静電ポテンシャル比較

E. coli と *P. furiosus* 2種の RNaseHII におけるアミノ酸配列にそって残基の静電ポテンシャル(等電点)をプロットし, 線でつないだもの. アミノ酸配列の相同性は約 30%と低い.

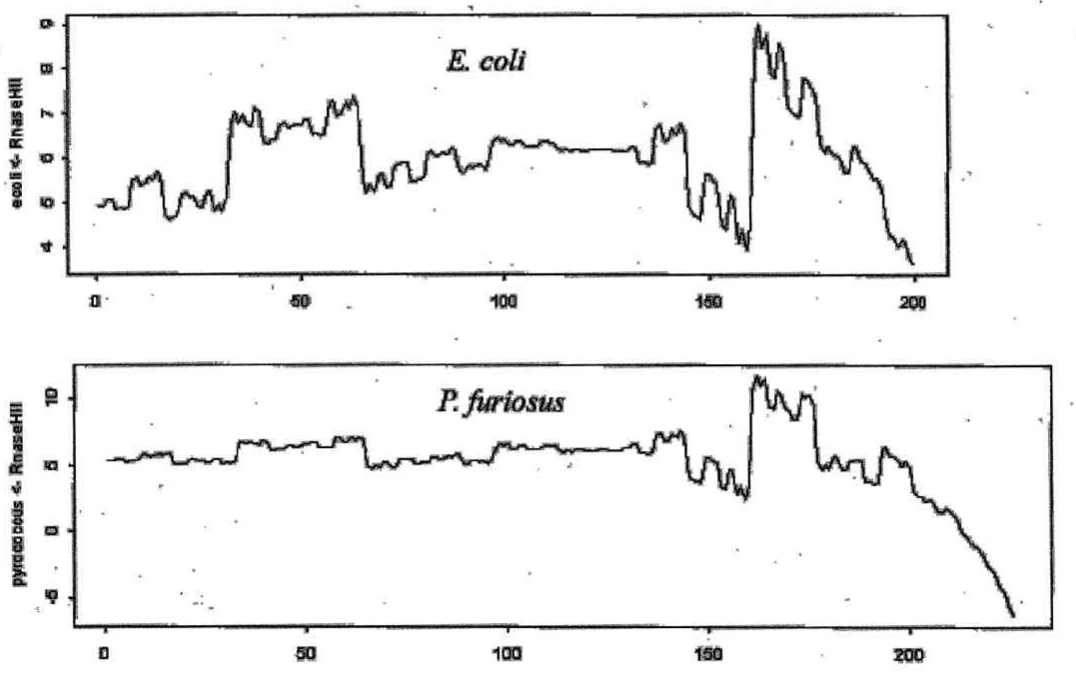


図 8. ウェーブレット変換を用いて再構築した 2 種における RNaseHII の波形

E. coli と *P. furiosus* 2 種の RNaseHII における電荷の信号からノイズを除去し波形を再構築した結果. 酷似するパターンが見受け

られた. .

V 総合考察

本研究の前半で提唱した新規機能推定法（統合解析 3．4）はアミノ酸の持つ性質や配列上における“周期”を定義することで機能予測を行った新規の解析手法であるといえる．この手法の有効性に関してはリストアップした候補蛋白質（TABLEⅣ）のいくつかをランダムに選択し，実験によって検証することで確認できると考える．そして新規の RNA 結合蛋白質であると確認できたものの割合が既知の true positive で予測した割合に近ければ，本手法の有効性を支持する結果となりうる．またオーソログ，BLAST，PPI などの従来のバイオインフォマティクスを用いた手法で予測した結果と比較することで本解析の信頼性の指標を得ることができると考える．

一方で解析手法はこれで完成したわけではなく，改良の余地が十分にあると考える．精度を高める具体案としてフィルタリングの順番を替える，クラスタリングに使用する周期の最適な組み合わせを模索する，SVM(support vector machine)を用いたクラスタリング手法を試みる，などが挙げられる．今後 29 個の候補蛋白質について候補蛋白質のリストの予測結果と比較検討していきたいと考える．また実験を用いた候補蛋白質の機能検証を行っていきたい．

ウェーブレット解析に関しては，従来のフーリエ変換では抽出することができなかった周期的特徴を得ることが可能となった．仮に前半部分に低周波数，後半部分に高周波数のを持つような信号がある場合，この信号をフーリエ変換しても前半と後半部分の特徴が打ち消しあってノイズとして埋もれてしまう．逆に Wavelet 変換では配列特性を逃すことなく表現することが可能であることが今回の解析ではある程度確認できたといえる．一方で再構築した波形の類似性がなにに依存しているかを調べることが今後の課題だといえる．本研究で用いた RNaseHII の場合，一次配列の約 30%の類似性の特徴が色濃く反映されたのか，それともアルファ-ヘリックスやベータ-シートのような高次構造に依存した結果なのか，つまりどの次元の情報の影響を再現しているのかを調べていく必要性がある．

VI 謝辞

慶應義塾大学先端生命科学研究所の金井昭夫助教授，同技術員/非常勤講師の佐藤朝子氏，同メディア政策研究科博士課程の藤森茂雄氏，鈴木治夫氏，沼田興治氏，同環境情報学部の谷内江望氏には有用なアドバイスを頂き大変感謝している．同プロジェクトの井元淳氏，根岸義輝氏とは有益なディスカッションをさせていただいた．また慶應義塾大学環境情報学部の冨田勝教授にはこの研究環境を与えてくださったことに深く感謝している．

Ⅶ 参考文献

[1] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ; Basic local alignment search tool. *J Mol Biol.* 5;215(3):403-10. 2003

[2] Suzuki H, Saito R, Kanamori M, Kai C, Schonbach C, Nagashima T, Hosaka J, Hayashizaki Y; The mammalian protein-protein interaction database and its viewing system that is linked to the main FANTOM2 viewer. *Genome Res.* Jun;13(6B):1534-41. 2003

[3] iala, G, Stetter, K. O. Pyrococcus furiosus sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. Arch. Microbiol. 145:56-61. 1986

[4] Maeder DL, Brown JR, DiRuggiero J, Stump MD, Yeh RK, Weiss RB, DunnDM. (2001); Genomic sequence of hyperthermophile, Pyrococcus furiosus: implications for physiology and enzymology. Methods *Enzymol* 300, 134-157.

[5] Kyte, J. and Doolittle, R. 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157: 105-132.

[6] White SH, Jacobs RE. Observations concerning topology and locations of helix ends of membrane proteins of known structure. J Membr Biol. 1990 May;115(2):145-58.

[7] Nishikawa K, Kubota Y, Ooi T. Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. J Biochem (Tokyo). 1983 Sep;94(3):981-95.

[8] Kanai A, Oida H, Matsuura N, Doi H. Expression cloning and characterization of a novel gene that encodes the RNA-binding protein FAU-1 from Pyrococcus furiosus. Biochem J. May 15;372(Pt 1):253-61. 2003

[9] Innis CA, Anand AP, Sowdhamini R. Prediction of functional sites in proteins using conserved functional group analysis. J Mol Biol. 2004 Apr 2;337(4):1053-68.

[10] B. B ハバード著 山田道夫/西野操 訳 ウェーブレット入門 2003 年 2 月 20 日初版

[11] Nagai N, Kuwata K, Hayashi T, Kuwata H, Era S. Evolution of the periodicity and the self-similarity in DNA sequence: a Fourier transform analysis. Jpn J *Physiol.* Apr;51(2):159-68. 2001

[12] Chechetkin VR, Turygin AYu. Search of hidden periodicities in DNA sequences. *J Theor Biol.* 1995 Aug 21;175(4):477-94.

[13] Taylor WR, Heringa J, Baud F, Flores TP(2002); A Fourier analysis of symmetry in protein structure. *Protein Eng.* Feb;15(2):79-89.

[14] Donnelly D, Overington JP, Ruffle SV, Nugent JH, Blundell TL. Modeling alpha-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues. *Protein Sci.* 1993 Jan;2(1):55-70.

[15] Sunyaev SR, Bogopolsky GA, Oleynikova NV, Vlasov PK, Finkelstein AV, Roytberg MA.

From analysis of protein structural alignments toward a novel approach to align protein sequences. *Proteins*. 2004 Feb 15;54(3):569-82.

[16] Yasuo Yonezawa, Takayuki Kamei Structure informational properties of protein sequence data 7, 115 - 122, 1997

[17] Singer GA, Hickey DA. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. Gene. 2003 Oct 23;317(1-2):39-47.

[18] Sato A, Kanai A, Itaya M, Tomita M. Cooperative regulation for Okazaki fragment processing by RNase HII and FEN-1 purified from a hyperthermophilic archaeon, Pyrococcus furiosus. *Biochem Biophys Res Commun*. 2003 Sep 12;309(1):247-52.

[19] Miyanoiri Y, Kobayashi H, Imai T, Watanabe M, Nagata T, Uesugi S, Okano H, Katahira M. Origin of higher affinity to RNA of the N-terminal RNA-binding domain than that of the C-terminal one of a mouse neural protein, musashi1, as revealed by comparison of their structures, modes of interaction, surface electrostatic potentials, and backbone dynamics. *J Biol Chem*. 2003 Oct 17;278(42):41309-15. Epub 2003 Aug 07.

ゲノム情報を利用した新規RNA結合蛋白質推定とその構造パターン予測

2004年 6 月10日　　初版発行

著者　藤島皓介

監修　金井昭夫

発行　慶應義塾大学 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-SWP 2003-A-005

■本論文は研究プロジェクトにおいて優秀と認められ、出版されたものです。