モチーフ探索のための断片ペプチドライブラリー　FRAP の整備と特徴
Fragment Peptide Library(FRAP) and Analysis of Motifs.

瀬戸保彦、磯山正治、池内義典、西川建(1)、川北栄継(1)、金久実(2)
（財）蛋白質研究奨励会　(1)蛋白工学研　(2)京大化研
Y.Seto, M.Isoyama, Y.Ikeuti, K.Nishikawa(1), Y.Kawakita(1), and M.Kanehisa(2)
Protein Res.Found., (1)Protein Eng.Res.Inst., and (2)Kyoto Univ. Inst.Chem.Res.

1．モチーフ候補としての断片ペプチド収集の背景

ひとつの生物の中で同じような機能を示す様々のタンパク質が知られている。例えばヒトの血液凝固系、線溶系、消化液、タンパク質のプロセシングなどで、いづれもセリンプロテアーゼとしてしられるペプチド結合の加水分解酵素が関与している。異なる生物の間で同じ機能を持つタンパク質が知られている。それらの配列を比較すると、短い共通の配列、つまり保存された配列が通常見いだされる。

タンパク質はアミノ酸が数十或いは数百個つながり、特有の立体構造を持った分子であるが、この分子が機能を発揮するのは別の分子と相互作用をすることによる。その作用に直接関与しているアミノ酸残基は基本的にはせいぜい数残基であろう。このような保存された配列或いは機能的に重要な短い配列をモチーフと呼ぶ。

ところで、配列を決めると似た配列を持ったタンパク質を検索する。このような作業が日常的になってきた。実際には似ているという事を厳密に、一般に通用するように言い表す事は難しい。そこで研究者は自分の持つ知識を利用する。その知識を背景として似ているタンパク質のセットを選んで比較する。文献に記載されているそのような配列比較データから我々は幾つかの基準の元に保存配列、つまりモチーフ（正確にはモチーフ候補）を収集し断片ペプチドライブラリー(FRAP)を作成して来た(1)。

本報告ではFRAPに蓄積された配列、特に保存配列に焦点を絞り解析する事によって、モチーフとはどの様なものなのか、如何に整備しているのかそしてそれを何に利用できるのかについて述べる。

2．モチーフの収集、解析と考察

モチーフは、配列データベースを作成する過程で調査する配列決定の文献を利用して配列比較データから抽出された。モチーフは次に示す規範に従って抽出されている(1)。

(1) 5から20残基の長さ　　(2) ひとつのタンパク質に3個程度あるだろう
(3) 複数の機能領域(ドメイン)を持つタンパク質はドメインごとに考える。

FRAPに蓄積した配列は合計約8000件、その中保存配列は7540件(17.15版)で、3502件の文献より抽出された。モチーフの長さとエントリー数(括弧の中)を次に示す。

-------- LENGTH OF MOTIF:AMINO ACID RESIDUES(ENTRIES) --------
5(2077), 6(2406), 7(1646), 8(870), 9(318), 10(113), 11-20(110)

＊＊＊モチーフには様々なパターンがある＊＊＊

様々な配列比較データから保存した断片配列を収集する過程でいわゆるモチーフは単に特徴的な短い配列だけではなく、様々なパターンがある事が分かってきた。配列の並び、組成、２次構造、アミノ酸の官能基などに着目して同族タンパク質に特有の１次構造上の短い領域の特徴となっている。これらは次の(1)-(4)のように分ける事ができる。広い意味では(5)の比較的長い領域に使われるドメインをもモチーフと呼んでいる。

(1) ５から１０残基の引き続く配列で幾つかの残基に任意性を許す
(2) それぞれ２－４残基からなる互いに離れている２、３個の集まりによるパターン
(3) 配列というよりは組成の特徴としてのパターン
(4) 繰り返し配列
(5) ２０から５０残基の特徴的なドメインと呼ばれる配列

------------------ EXAMPLES OF PATTERN FOR MOTIF --------------------
(1) Ser protease GDSGG, superoxide dismutase WEHAYY, homeo box WFQNRR, amino acid-tRNA synthetase KMSKS
(2) GTP binding motif GXXXXGK--DXXG--NKXD, Zn finger CX3C--CX2C, Leu zipper LX6LX6LX6L
(3) Leu rich amphiphilic helix motif
(4) repeat motif:RNA polymerase YSPTSPS, collagen PPG, histone GGK
(5) helix turn helix/EF hand domain:tramsmembrane domain:Cys rich EGF domain homeo box:Ig domain:calmodulin domain:sushi domain:lectin domain

＊＊＊モチーフのパターンは2次構造に依存する＊＊＊

アミノ酸７残基のモチーフABCDEFGにおいてアミノ酸A-Gのいづれもがタンパク質にとり重要だろうか。機能に直接関与しているモチーフの場合、モチーフはリガンドや基質等の分子と弱い相互作用をしている。モチーフを構成するアミノ酸の側鎖が相互作用をすると共に、その特異性を決めている。個々のアミノ酸は表記上は1次元のつながりであるが、立体的な配置を背景に相互作用が起こる。

2次構造の観点からモチーフを見た場合、その配列は次に示すようにヘリックス、ベータシート、ループのいづれかの上にある。ヘリックスの場合アミノ酸残基は3.6残基で1回転する。従って立体的に特定の面に向いているアミノ酸は例えばA,D,Gである。その意味でモチーフはAXXDXXGのパターンになる事が予想される。リガンドの形によってはABXXXFGとなる。一方ループの場合はAからGの比較的多くのアミノ酸が必須となることが予想される。

これはモチーフ配列において、アミノ酸残基にあまり違いを許さない型、つまり強いモチーフと幾つかの残基に違いを許す型とがある事に対応している。具体例を次に示す。このような観点から現在モチーフのリファインメントを行っている。強く保存されているモチーフは実際に立体構造の分かっているタンパク質で調べてみると多くはループ構造にある。FRAPの現在の配列でXを含まないものは5384件、Xを含むものは2156件である。

```
-------------------- MOTIF AND SECONDARY STRUCTURE ------------------
thymidine synthase:MALPPCH(loop),RSXDXXLG(helix)
Fe superoxide dismutase:WEHAYY(loop),HXXKHHXXYV(helix),QXXNHXXYW(helix)
RNase H:TDSQYV(loop),ras p21:DTAGQE(loop),nuclease:GNA1ER(loop)
```

＊＊＊遠く離れたモチーフが共同作業をする＊＊＊

タンパク質が機能を発揮する上での立体構造の重要性という点から、ひとつのモチーフというよりは立体的に近いいくつかのモチーフが共同でその機能を発揮していると考えられる。２或いは３個のモチーフが相手分子を挟み込むといった姿が相互作用として考えられるわけである。その意味で直接に重要なアミノ酸はひとつのモチーフの中で１、２或いは３個であろう。５個や７個の引き続く配列のアミノ酸全てが立体的に相手分子と相互作用をしている事は考えにくい。このようなモチーフ数個が集まってより高次元のパターンとしてのモチーフとなっている事も予想される。共同作業と言う意味で一方のモチーフの変異を別のモチーフが補完する事も有り得る。

同族タンパク質が同じ生物或いは様々な異なる生物で見いだされ、その成分が増えるに従って、モチーフのゆらぎも大きくなる。プロテインキナーゼの３個のモチーフが弱いパターンとして現れる例を次に示す。

```
-------- FLUCTUATION IN MOTIF SET ----------
protein kinase   HRDLKPEN---DFGLAR---KWTAPE
                 HXDLXXXN---DFG----------APE
                 HXXLXXXN---DXGL------WTXPE
```

このようなモチーフ探しは、アミノ酸50から100残基の長い配列全体から弱い保存配列パターンを見つける事を意味する(2)。膨大な計算時間が必要となりそうである。ゲノム解析プロジェクトで生ずる大量の情報は先ず何よりも１次元の文字情報である。立体構造情報はない。従ってこのような弱いモチーフ探しを可能にする方法論が必要となる。我々は現在、立体構造既知のタンパク質についてひとまず共同作業をするモチーフの抽出を進めている。

＊＊＊モチーフのゆらぎは多様なタンパク質を生成した＊＊＊

モチーフといえども変化する。セリンプロテアーゼはペプチド結合を加水分解する酵素でモチーフGDSGGP を持つ。脂質を加水分解するリパーゼはモチーフGXSXG を持つ。これらのモチーフはそれぞれのタンパク質のconvergent evolution により似た配列を持つようになったというよりは共通のモチーフがdivergent evolution を起こしたものであろう。プロテインキナーゼのモチーフHRDLKPEN,HRDLAARN もキナーゼという点では同じ族のタンパ

ク質がそれぞれのモチーフに分散進化して階層的により下位の族に分かれていったのであろう。
このようにモチーフを詳細に解析する事により、進化によって生じたタンパク質の階層構造が明らかになる。米国の配列データベースPIRにおいては、タンパク質をスーパーファミリー、ファミリーという階層に分類しているがそのPIRに対してモチーフ配列を評価した。複数のスーパーファミリーに共通なモチーフ、特定のスーパーファミリーにユニークなもの、スーパーファミリーの中のあるファミリーにユニークなものというように分類上特定の階層に対応するモチーフが認められた(1)。
モチーフがどの階層に対応するものかを考慮しないとそのユニーク性を議論できない点がモチーフのリファインメントを難しくしている。

＊＊＊モチーフはひとつのタンパク質に複数のセットのあるものがある＊＊＊
配列データベースにあるタンパク質の平均アミノ酸残基数は320である。そして多くのタンパク質が500を越える残基を持つ。長いタンパク質の多くは次に示すような複数の機能領域、マルチドメイン構造を持つ。括弧内は残基数である。

```
-------------------- PROTEINS WITH MULTIPLE DOMAIN --------------------
ANF receptor(903):ligand binding domain,cyclase domain,kinase domain
Factor C(1019):EGF domain,sushi domain,lectin domain,Ser proteas domain
```

また酵素には２役酵素や多機能酵素が知られている。このようにモチーフ抽出において、タンパク質一つあたり多くのモチーフ候補を収集しておく必要があった。この事が逆にモチーフの整備を難しくしている。どの機能或いはドメインに対応するモチーフであるか念頭において整備しているが、知識データベースの充実が待たれる。

＊＊＊FRAPには生のデータがある＊＊＊
我々のライブラリーではモチーフを抽出した文献をエントリーの単位として配列を入力している。文献において著者はいくつかのタンパク質を整列表記して議論しているわけだが、どのタンパク質を選んで整列するかは個々の著者の知的な背景、意図或いは利用したデータベースに依存している。
タンパク質の類似性には段階があり、階層的である。モチーフ抽出をした論文での議論の背景に応じて、配列の整列表記はその階層性の特定のレベルを見ている事になる。モチーフの整理を現在進めているが、統合していない原因のひとつはここにある。各々のモチーフで見ているタンパク質の階層は必ずしも同じではない。
スーパーファミリー、ファミリー或いはサブファミリー等に対応する様々なレベルの整列表記がされていて、そこから抽出したモチーフも当然それぞれの階層に応じたグループに特徴的なものである。議論しているタンパク質に関連した全体の階層が見通せないと、モチーフの統合は困難である。ひとつのタンパク質でも、進化的階層に応じていくつかのモチーフのグループが残っていると考えられる。このためにモチーフの統合化はせず、初めに抽出したまま、多面的に整備できるように残してある。

＊＊＊モチーフは知識データによってリファインされる＊＊＊
モチーフの相互作用する相手は何であるか。金属イオン、ATP等のヌクレオチド、燐酸等のような構造的なゆらぎのないものであれば堅いモチーフとなる。或いはペプチド、タンパク質のような立体的に柔らかいものがリガンドであればモチーフのパターンは複雑で柔らかな配列となる。
あるタンパク質についていくつものモチーフがひとまず得られたときにリガンドが何であるか或いは相互作用する配列上の位置の情報があれば、より良いモチーフを選ぶ事ができるだろう。このような点からもタンパク質或いはその配列に関連した知識データベースの充実を進めている。

３．モチーフで何が分かるのか
＊＊＊モチーフはタンパク質のアイデンティティを決める＊＊＊
モチーフは同族タンパク質に特徴的な短い配列パターンである。従ってモチーフを探し出すためには同族となるタンパク質のグループが必要である。タンパク質の配列が続々と配列データベースに蓄積される一方、それらの配列のグループ分けは遅々として進まない。米国NBRFのBarker,W.らがこつこつと長年進めているだけである。そこで我々は昨年から配列の分類を始めた。

ふたつのタンパク質の配列を似るように並べたとき、つまりアラインメントしたとき、50パーセントのアミノ酸が同じであれば殆ど間違いなく同族である。同じアミノ酸が20パーセント前後の場合は同じグループともそうでないともいえるような関係である。これをDoolittleがTwilightと呼んでいるのにちなんで、たそがれ（誰そ彼）の間柄と呼ぶ。

分類をするには指標を何にするかによって様々な規範がありえる。配列のアミノ酸組成、長さ、機能、安定性、細胞における存在場所等々である。我々は基本的には配列以外の情報はないとの前提で分類する事にした。ゲノム解析が順調に進むにつれてまさにそのような状況が生まれる。もっとも、分類する時点で得る事のできる知識情報は実際には利用する。

たそがれの関係にあるふたつのタンパク質はFRAPライブラリーにある配列が共有されていれば同族とする。タンパク質を分類する事自体は相対的な意味を持つわけだが、一方個々のタンパク質の帰属をする事、つまり絶対的なアイデンティティを決める事でもある。その意味に置いてモチーフがアイデンティティを決めている。

＊＊＊モチーフによりいろいろな生物の同族タンパク質を探す＊＊＊

生物の中には塩酸溶液や原油の中といった、人にとっては考えられないような化学的、物理的環境で生活しているものがある。特定の生物的機能が突出していたり、まったくなかったりする生物がある。このような生物が研究者の興味をかき立て、核酸やタンパク質の配列が決められつつある。一方、ヒトのタンパク質の配列が決められるとそれらの生物における、対応するタンパク質の配列を決めるというような関係でデータの蓄積が進んでいる。

モチーフはこのようなさまざまな生物の核酸やタンパク質配列の間の架け橋である。オリゴヌクレオチドプローブやPCR増幅、抗ペプチド抗体の利用といった技術が開発された現在、保存されたモチーフ配列を利用して生物の間の同族のゲノムやタンパク質を抽出する事ができる。

我々のFRAPライブラリーにある断片配列は正確にはモチーフ候補である。これらの配列情報を利用した実験を積み重ねる事によって、候補のふるい分けができ、真にモチーフと呼べるものを選び出す事ができる。実験研究者による評価と情報処理によるモチーフのリファインメントは表離一体のものといえる。

＊＊＊モチーフでゲノムのタンパク質地図をつくる＊＊＊

大腸菌ゲノムは450万の塩基からなるという。全て翻訳されるとアミノ酸残基で150万、平均300残基で一つのタンパク質をつくるとして5000個のタンパク質となる。同族タンパク質は5個のタンパク質を含むとして、1000の族つまり、生物を構成するタンパク質のアイデンティティは1000かけるいくつかのモチーフで決める事ができる計算になる。

例えばモチーフに対応するオリゴヌクレオチドによりゲノムのうえにタンパク質族のマークをいれる事が可能となる。モチーフが多様なタンパク質をつくってきたと言う意味で、ゆらぎのあるモチーフがゲノムのうえで分かれば、発現していない配列の正体の一面が解明されるだろう。モチーフの変化は機能変異や病態との関連が予想される。

4．最後に

本報告では整備されたモチーフの提案までには至っていない。我々の関心は日々発表される配列データからモチーフ候補を抽出し、FRAPライブラリーを充実することにある。タンパク質はカオスのような姿を持つ。タンパク質を20の要素からなる文字列と考え、初期配列から要素を1個づつ変えてできる時系列変化を眺めてみる。これは36億年の進化である。初期条件とアミノ酸置換の規則とからでは予測できないようなフラクタル図形に似たタンパク質ができあがってゆく。非線形の世界である。

FRAPのモチーフは変異データベースの情報と共にタンパク質のカオスとしての姿を解きあかす鍵であると我々は考えている。FRAPのデータを多くの領域の研究者が活用されることを期待したい。FRAPを活用したい方にはフロッピーをお送り下されば提供します。モチーフの検索プログラムもついています。またゲノムネットによりオンラインでの入手も可能です。

参考文献

(1)Seto, Y., et al.:Proteins:structure, function, and genetics., 8, 341(1990)
(2)Smith, H.O., et al.:Proc. Natl. Acad. Sci. USA, 87, 826(1990)