類似タンパク質グループを特徴付ける配列モチーフの辞書の自動作成

Automatic construction of a dictionary of sequence motifs that characterize groups of similar proteins

京都大学化学研究所　荻原淳・内山郁夫・金久實
Atsushi Ogiwara, Ikuo Uchiyama, Minoru Kanehisa
Institure for Cheminal Research, Kyoto University

1．はじめに

ゲノム計画の推進によって大量の配列データが得られるようになってくると、これらをいかに妥当に、しかも効率よく生物的な特徴付けできるかが問題になってくる。我々は特徴付けの指標の１つとして配列モチーフを捕えている。従来配列モチーフは実験的に同定された機能部位などから、どのような配列パターンが重要であるか手作業で定義されることが多かった。しかし、これからさらに配列データが増大していく状況のもとでは、与えられた配列から何らかの手段で自動的に特徴を抽出することが必要になってくる。我々はこのような目的で、配列データの集合からモチーフを自動的に定義するための手続きについて研究してきた。ここではモチーフ抽出の手続きを定式化し、実際のデータに適用した結果について発表し、このようなモチーフの意味づけの問題を考察する。

2．配列グループからのユニークペプチドの抽出

まず始めにデータベース中で、ある共通の性質をもつ１つの配列グループに着目する。グループ内では保存され、かつそのグループに特異的な配列断片を、ユニークペプチドと名付ける。ユニークペプチドを探すために、長さ l の可能なすべての配列パターン（20^l 通り）について、データベース全体で配列パターンを含むエントリーの出現頻度 $\mathcal{N}p$ と、着目しているグループでの出現頻度 $np(g)$ を計数する。このグループにおけるエントリーの総数を $e(g)$ とすると、以下の条件に合うものがユニークペプチドとなる。

(1) Uniqueness: $np(g) = \mathcal{N}p$. かつ

(2) Conservation: $np(g) > \theta\ e(g)$.

ここで θ は保存されているエントリー数の割合で、0.7 または 0.8 を採用した。

次にユニークペプチドの定義を拡張し、複数のグループにわたって保存し、かつ特異的に現われる配列断片も探索した。あるグループ群 $G = \{ g_1, ..., g_k \}$ に対し、各グループでのある配列パターンの出現頻度 $np(g_i)$、各グループのエントリー数 $e(g_i)$ とすると、

(1) Uniqueness: $np(g_1) + ... + np(g_k) = \mathcal{N}p$. かつ

(2) Conservation: $np(g_i) > \theta\ e(g_i)$ for all $i \in \{ 1, ..., k \}$.

実際には、我々は配列データベースとして PIR を使用し、共通配列グループとしては PIR のスーパーファミリー分類を利用した。ただし保存性を言うためにはそのグル

ープがある程度の大きさをもたなくてはならないと考え、構成エントリー数が一定数（3ないし5）未満のものは切り捨てた。データ収集に利用したリリースはRelease 26である。また、探した断片の長さlは4～6とし、保存性のパラメータθは0.7および0.8について調べた。また複数グループでのユニークペプチドの探索には、構成グループ数kが5以下のものに限って探した。

3．ユニークペプチドからモチーフ文脈の構成

前節で述べた手続きでは、得られたユニークペプチドが配列中のどこに出現するのか分からない。実際の配列に即して場所を調べることで、我々がモチーフ文脈と呼んでいるものを作った。モチーフ文脈とはアミノ酸配列パターンおよび間隔指定子の並びで、配列データを特徴的な配列パターンの並びだけで表現したものである。前節で得られたユニークペプチドの配列パターンを再度各配列中で探す際には、1文字までの置換を許した。これはグループ内での小さな変異を許容するためである。見つかった配列断片のそれぞれの出現位置を相対位置関係に直し、配列パターンが相対位置でつながれた構造が、モチーフ文脈構造となる。

このモチーフ文脈はグループ内の各エントリー毎に得られるもので、つぎにこれをグループ全体で統合しなくてはならない。統合されたものをコンセンサスモチーフ文脈と呼ぶ。この作成手続きは配列パターンのマルチプルアライメント問題と見做せる。配列が繰り返し構造をもつ場合や、1文字置換を許したために起こるノイズを拾った場合などは、一つの配列中に同じパターンが複数回出現することもあるため、パターンの出現順序や相対位置関係がもっとも良く保存されるような最適アライメントを求める必要がある。そこで、ペアワイズにダイナミックプログラミング法を繰り返し用いて最適なマルチプルアライメントを求めることにした。

モチーフ文脈のi番目のパターン$\mathcal{P}_i$と、コンセンサス文脈のj番目のパターン$\mathcal{P}cons_j$について、$\mathcal{P}_i = \mathcal{P}cons_j$であるとき、スコア$S_{ij}$は

$$S_{ij} = \max(S_{kl} - \text{g.p.}(i,j,k,l) + \text{w}(i,j)) \qquad (k < i,\ l < j,\ \mathcal{P}_k = \mathcal{P}cons_l) \qquad (*)$$

で定義される。ここで、w(i,j)はモチーフ文脈のパターンiと、コンセンサス文脈のパターンjの一致によるスコアである。また、g.p.はモチーフ文脈のi番目とk番目のパターンとの間と、コンセンサス文脈のj番目とl番目のパターンの間にあるギャップのペナルティーであって、一般に、モチーフ文脈のi番目とk番目のパターンの相対距離とコンセンサス文脈のj番目とl番目のパターンの相対位置の差、diff(i,j,k,l)の関数で表される。

$$\text{g.p.}(i,j,k,l) = f(\text{diff}(i,j,k,l))$$

ただし

$$\text{diff}(i,j,k,l) = \begin{cases} \text{min\_relpos}(j,l) - \text{relpos}(i,k) & (\text{relpos} < \text{min\_relpos}(j,l)) \\ 0 & (\text{min\_relpos}(j,l) \le \text{relpos}(i,k) \le \text{max\_relpos}(j,l)) \\ \text{relpos}(i,k) - \text{max\_relpos}(j,l) & (\text{relpos} > \text{max\_relpos}(j,l)) \end{cases}$$

(*) 式をすべてのi,jについて順に計算することにより、最高のスコアを持つアライメントを求めることができる。これをグループ内のすべてのエントリーについて繰り返すことによって、グループ全体のコンセンサス文脈を作ることができる。このうち、半分以上のエントリーで、置換を含まないパターンが見つかっているパターンのみを

残し、他はノイズとして除く。最後に重なりを持ったパターンをつなぎあわせて、次のような形式でコンセンサス文脈を表す。

<Motif Block1> [Min_spacer, Max_spacer] <MotifBlock2> ...

ここで、つながれてひとかたまりとなった配列パターンをモチーフブロックと呼ぶ。

4．モチーフ辞書とその意味づけ

コンセンサス文脈は1残基までの置換パターンを許して作成したが、実際にどんな置換があったかの記録は行っていない。また2残基以上の置換がある場合や挿入・削除がある場合も見落とされている。そこで、コンセンサス文脈をもう一度グループ内の各配列と照合して、すべての置換パターンを探しだした。

照合は、モチーフ文脈中でモチーフブロックごとにNeedleman & Wunschタイプのダイナミックプログラミング法によってアライメントをとり、ブロック間の間隔をコンセンサス文脈のものに合わせるようにしてつなぎあわせることにより行なった。置換した残基の数がブロックを構成する残基数の4割以下の場合に置換パターンを記録し、4割を超える場合には、そのエントリーはそのブロックを持たないものとした。

このようにして作成したコンセンサスモチーフ文脈の例を示す。PIR データベース Release 26 の 2350 のスーパーファミリーのうちエントリー数が3以上のものが 521 あるが、$\theta = 0.7$ の場合 324 のスーパーファミリーに対しては、自動的にコンセンサスモチーフ文脈を定義できた。これを収集したものをモチーフ辞書と呼ぶ。

```
コンセンサスモチーフ文脈の例：スーパーファミリーによるグループ化
< 50 > glyceraldehyde-3-phosphate dehydrogenase
     +GFGR{I|-}GR=[129,134]=+SNASCTTN{C|S}LAP=[14,14]=
     +{L|M}MTTVH=[30,31]=+TGAA{K|R}A{V|T}=[92,95]=+{S|A}WYDNE
< 60 > acyl-CoA oxidase
     +{T|A}{V|I}GDIG=[10,21]=-RFFM=[153,159]=+ACGGHG
< 68 > glutamate dehydrogenase (NAD(P)+)
     +AEG{A|S}N=[24,31]=+N{A|C}GGV
< 83 > NADH dehydrogenase (ubiquinone) chain 2
     +LS{L|M}GGLPP
< 96 > cytochrome-c oxidase polypeptide I
     -{Q|E}HLFWFFGHPEVYI=[70,127]=-VV{A|G}HFH{Y|L}{V|M}L{S|A}
< 97 > cytochrome-c oxidase polypeptide II
     +G{H|F|Y}QWYW=[83,91]=+YG{Q|A}CSE{I|L}
< 98 > cytochrome-c oxidase polypeptide III
     +SP{W|S}P{L|I}=[111,113]=+PLLNT=[105,106]=-YWHFVDV
```

コンセンサスモチーフ文脈はスーパーファミリーという機能的にある程度独立したグループに特異的でかつその中で保存しているという性質から得られたものであるから、得られたアミノ酸配列パターンは生物学的な機能部位を多く含むことが期待される。この検証として PIR データベースの FEATURE テーブルにある活性部位などと、どの程度対応するか調べた。

NBRF Feature	#Entries	#Sites	#Hit	%Hit
Active site	111	253	120	47.43
Binding site	370	1059	119	11.24
Inhibitory site	2	2	0	0.00
Modification site	127	248	90	36.29

データ数は必ずしも多くないが、活性部位や修飾部位についてはある程度の対応が見られた。

つぎに、モチーフ辞書を配列データのスーパーファミリー分類の指標として利用した例を示す。PIR データベース Release 29 のうちスーパーファミリー化されていない PIR2 の配列の分類を試みた。モチーフの各ブロックの配列パターンが８０％以上一致しているかどうかの判定と、グローバルアライメントによりホモロジーを調べる判定を比較したところ、９３例の分類が一致したが、７例はアライメントとして検出できないのにグループとして認識され、逆に４６例ではアライメントで検出できているのにモチーフ辞書ではできなかった。つまり、false negative となるものがかなり存在しており、モチーフ辞書の作成方法、およびそれを利用した分類方法に今後の改良が必要である。

本手法では配列データが事前にグループ化されている必要があるが、ユニーク性を複数グループに拡張することにより、グループ間の関連や、より大きなグループ化を調べることができる。また逆にそこから例えばスーパーファミリーというグループ分けの妥当性を検討することができる。２節で述べた方法により、複数のグループでしか存在しない共通配列を網羅的に調べたところ、例えば HPDKKG というパターンが T antigen の３つのスーパーファミリー（large, middle, small）でのみに見られた。このように本来類縁性が高いと考えられるが別々のスーパーファミリーとなっている場合に共通のパターンがいくつか見つかった。これ以外にも多数のパターンが得られたが、それが実際に生物的な関連を示すのか、小規模なスーパーファミリーでの単なる確率的な一致なのか検討中である。

５．まとめ

我々が開発した配列から自動的にモチーフを定義する方法は手続き的には安定しているが、得られたモチーフ辞書が配列データの同定にどの程度有効であるかは、さらに検討が必要である。また、スーパーファミリーといっても実際にはメンバーが１つか２つしかないグループが大多数であるので、グループの定義についても再考を要する。この手続きは保存性など以外の規範によるデータにも使えるため、様々な規範のデータを取り込むことでより充実したモチーフ辞書を作成していくことができると考えている。