古くて新しい問題：タンパク質二次構造ブレイカ

今井賢一郎 [1)2)]，美宅成樹 [2)]

1. 豊田理化学研究所，2. 名古屋大学大学院工学研究科

Secondary structure breaker in soluble protein

Kenichiro Imai[1, 2] and Shigeki Mitaku[2]

1. Toyota Physical and Chemical Research Institute

2. Department of Applied Physics, Graduate School of Engineering, Nagoya University

プロフィール

今井賢一郎　豊田理化学研究所　奨励研究員

E-mail: imai@bp.nuap.nagoya-u.ac.jp

464-8603 名古屋市千種区不老町名古屋大学大学院マテリアル理工学専攻応用物理分野美宅研究室

美宅成樹　名古屋大学　教授

E-mail: mitaku@nuap.nagoya-u.ac.jp

464-8603 名古屋市千種区不老町名古屋大学大学院マテリアル理工学専攻応用物理分野

1. はじめに

タンパク質科学の大きな目標の一つは，タンパク質立体構造形成のプロセスを物理化学的に理解し，コンピュータ上（*in silico*）においてアミノ酸配列情報から立体構造を再現することである．アミノ酸配列には，タンパク質の立体構造を形成するための情報が含まれている．しかし，タンパク質の立体構造形成プロセスを簡単な物理化学的ルールによって理解できるかどうかは，まだはっきりしていない．歴史的には，一見もっとも手軽な問題である二次構造予測が精力的に行われてきたが，この問題はまだ解決されていない．アミノ酸配列情報のみからタンパク質の立体構造を予測することは非常に難しいのが現状なのである．しかし，最近ゲノム解析や構造ゲノミクスの進展によって豊富なアミノ酸配列情報や立体構造情報が得られるようになり，研究の環境が大きく変わりつつある．それらの豊富な情報をもとに，もう一度アミノ酸配列の持つ物性量の分布と立体構造との相関を調べ直し，アミノ酸配列に含まれるタンパク質の二次構造形成のシグナルを明らかにすることが，古くて新しい課題として再登場してきていると著者らは考えた．

アミノ酸配列に含まれるタンパク質の構造の特徴として，筆者らは二次構造自体ではなく，そのブレイカに注目した．二次構造ブレイカとは，二次構造を壊し，ループ構造に変えるものであり，二次構造をつなぎ三次構造を形成するものとして重要な構造的特徴となっている．以前より，アミノ酸の出現傾向の解析から，α-ヘリックス，βストランドにおいてそれぞれ，二次構造を壊す傾向の強い残基が知られている[1),2),3),4)]．これらの残基では，二次構造のN端，C端で出現の偏りがあるものもある[3),4)]．しかし，これらの残基は互いの二次構造中にも存在し，任意のアミノ酸配列から二次構造ブレイカとなっているものを見つけることは難しい．実際，過去の研究では，α-ヘリックス領域だけ，あるいはβストランド領域だけに注目してルールを論議しているので，任意のアミノ酸配列に対応した予測ができるものになっていない．二次構造ブレイカを見つけるためには，これらの残基が本当に二次構造ブレイカとなる条件をさらに明確にすることが必要なのである．また，二次構造を壊す傾向の強い残基では，二次構造の種類によらず共通しているものが多い．そこで，著者らは，二次構造ブレイカの位置が二次構造形成以前に決まっている（したがって，二次構造の種類によらず予測できる）という可能性を検討すべく，より一般的なブレイク領域の解析を行った．そのとき，共通の二次構造ブレイカとして，4つのタイプのアミノ酸残基クラスタに注目し，それらについて検討した．そして，それら残基の周辺のアミノ酸配列の物性解析を行い，二次構造ブレイカとして働く条件を明確にすることができた．本稿では，二次構造ブレイカのメカニズムと筆者らが開発した二次構造ブレイカ予測システムについて紹介する．

2. 二次構造ブレイカ

2.1 4つのタイプの二次構造ブレイカ

プロリン，グリシンまたは，それらのクラスタは，二次構造を壊す傾向が強いものとしてよく知られている[1), 2) 3), 4)]．プロリンは，側鎖が主鎖構造の形成に制限を与えることで二次構造の規則性を壊す．一方，グリシンは，側鎖の体積が非常に小さいため，主鎖構造を柔軟にし，二次構造からループ構造への構造変化を起こしやすい．しかし，プロリンとグリシンが全ての二次構造ブレイカとなっているわけではない．そこで，筆者らは，新しいタイプの二次構造ブレイカとして，側鎖の小さい極性残基クラスタ，両親媒性残基クラスタに注目した．

セリン，スレオニン，アスパラギン，アスパラギン酸は，側鎖の体積が小さく，極性を持つ残基である．これらの側鎖の小さい極性残基は，プロリンやグリシンほどではないが，二次構造の端やループ領域に現れやすい傾向がある[1), 2) 3), 4)]．側鎖の小さい極性残基のクラスタは，グリシンと同様に主鎖構造に柔軟性を与え，二次構造ブレイカとして働く可能性がある．

リジン，アルギニン，グルタミン，グルタミン酸，ヒスチジンは，その側鎖に長い炭化水素鎖と極性基を持っており，両親媒性の性質の特徴を示すアミノ酸残基である．近年では，タンパク質のディスオーダー領域が分子認識と密接な関係があるとして注目されている[5),6)]．ディスオーダー領域は，本質的に構造を形成しない領域のことであり，この領域には，これら5つの両親媒性残基がよく現れる[6)]．実際，これらの残基も二次構造の端に見られ[1), 2) 3), 4)]，構造をディスオーダーにさせる両親媒性残基のクラスタも二次構造ブレイカとなると考えられる．

2.2 4つのタイプの残基クラスタの出現傾向

二次構造ブレイカ候補として，プロリン，グリシン，側鎖の小さい極性残基，両親媒残基の4つのクラスタを挙げた．しかし，これら4つの残基の出現位置を調べてみると残基の性質だけで二次構造ブレイカが決まっているわけではないことがわかる．

Fig.1 (a)-(d) は，PDB[7)]から立体構造既知の1031個の水溶性のタンパク質(配列相同性30%以下)を選出し，それぞれの残基クラスタについて7残基以上の二次構造周辺における全残基に対する出現比を調べたものである．横軸は，二次構造の端からの位置を示しており，0は二次構造の端を，負の領域は二次構造領域内，正の領域はループ領域内となっている．二次構造領域は，PDB[7)]の記述に従っている．また，二次構造の端領域として二次構造の端±3残基の領域をブレイク領域と定義し，ブレイク領域以外の二次構造領域，ループ領域を二次構造コア，ループコアと定義した．短いループに関しては，コア領域はなく，ブレイク領域のみとして解析している．アミノ酸配列上におけるそれぞれの残基クラスタの位置は，プロリン，グリシン，側鎖の小さい極性残基のアミノ酸配列における位置と両親媒性

指標（A index）[8]を用いて算出した[9]．プロリンの出現位置を見てみると，二次構造コアには，ほとんど現れることなく，ブレイク領域，ループコアに良く現れる．プロリンは，二次構造ブレイカとしてよく機能していることがわかる．しかし，他の残基クラスタを見ると，グリシン，小さい極性残基クラスタは，ブレイク領域，ループコアに多く現れるが，二次構造コアに現れる割合も無視できない．さらに，両親媒性残基クラスタにおいては，二次構造コアに現れる割合とブレイク領域，ループコアに現れる割合はほとんど変わらない．グリシン，小さい極性残基，両親媒性残基の中には，二次構造ブレイカとなるものもあるが，二次構造の一部となっているものあり，これらの残基が二次構造ブレイカとなるかどうかは，その残基の性質だけで決まるわけでなく，その運命を決める何か他の要因があることを示している．

2.3　周辺の配列環境による二次構造ブレイカ判別

タンパク質のローカル構造の形成を考えた場合，周辺のアミノ酸配列環境を無視することはできない．当然，二次構造ブレイカにおいても，その周辺のアミノ酸配列環境は重要な因子となるはずである．グリシン，小さい極性残基，両親媒性残基のクラスタ周辺のアミノ酸配列の物性を調べてみると，ブレイク領域にある残基クラスタと二次構造コアにあって二次構造の一部となっているものとでは，周辺のアミノ酸配列の疎水性$\langle H(j)\rangle$，ヘリックス周期性スコア$\langle HPS(j)\rangle$，トリプトファン・チロシンの両親媒性指標[8]（両親媒性指標 A' index）$\langle A'(j)\rangle$，セリン・スレオニン密度$\langle ST(j)\rangle$，グリシン密度$\langle G(j)\rangle$といった物性パラメータにおいて差が見られた．$\langle H(j)\rangle$は疎水性指標（KD 指標）[10]をもとに，$\langle A'(j)\rangle$は A' index[8]をもとに，$\langle ST(j)\rangle$，$\langle G(j)\rangle$は，それぞれの残基のアミノ酸配列における存在をもとに式(1)に従ってj番目の残基の周辺 7 残基の平均を求めたものである．また，ヘリックス周期性スコア$\langle HPS(j)\rangle$は，疎水性指標（KD 指標）[10]を用いた以下の式（2）で定義される．

$$\langle X(j)\rangle = \left[\sum_{i=j-3}^{j+3} X(i)\right] \Big/ 7 \tag{1}$$

$$\langle HPS(j)\rangle = \max\{\left|\langle HP(j)\rangle\right|, \left|\langle HP(j-1)\rangle\right|\} \tag{2}$$

$$\langle HP(j)\rangle = \left[H(j+5) - H(j+3) + H(j+1) - H(j) + H(j-2) - H(j-4)\right] \Big/ 5$$

Fig.2 は，ブレイク領域にある両親媒性残基クラスタと二次構造コアにあって二次構造の一部となっているもので，すべての残基クラスタ周辺 15 残基のアミノ酸配列の物性パラメータの平均値を比較したものである．ブレイク領域にある残基クラスタの周辺では，二次構造コア内にあるものに対して，疎水性が低く，また，ヘリックスの周期性スコアも低い．一方，トリプトファン・チロシンの両親媒性，セリン・スレオニン密度は，ブレイク領域

にある残基の周辺の方が高い．ループ領域などの規則的な構造をとならい領域は，タンパク質の表面にあり，水との親和性を高めるために疎水性が低くなっていると考えられる．また，そのような領域では，ヘリックスの特有の周期性も失われていると考えられる．トリプトファン，フェニルアラニンといった側鎖の大きい極性残基が多く見られるのは，排除体積効果によって二次構造の規則性を失わせるからだと考えられる．また，セリン，スレオニンといった側鎖の体積が小さい残基の密度が高いのは，主鎖の自由度を高めることによって二次構造を形成しにくくしていると考えられる．このような周辺のアミノ酸配列の物性パラメータの差は，他の残基クラスタ周辺においても同様である．そこで，周辺のアミノ酸配列の物性の差から環境因子 $\Delta\langle H(l)\rangle$，$\Delta\langle HPS(l)\rangle$，$\Delta\langle A'(l)\rangle$，$\Delta\langle ST(l)\rangle$，$\Delta\langle G(l)\rangle$ を定義し，二次構造ブレイカの判別解析を行い，判別式を得た [9]．それぞれの判別に有効な環境因子に応じて，グリシンの判別には式（3）を，側鎖の小さい極性残基クラスタの判別には式（4）を，両親媒性残基クラスタの判別には式（5）を用いる． $\underline{l}$ は，残基クラスタの位置を表す．

$$Gscore(l) = 2.01\Delta\langle H(l)\rangle + 1.90\Delta\langle HPS(l)\rangle + 5.30\Delta\langle A'(l)\rangle + 32.90\Delta\langle ST(l)\rangle - 0.68 \tag{3}$$

$$SPscore(l) = 2.43\Delta\langle H(l)\rangle + 1.31\Delta\langle HPS(l)\rangle + 1.03\Delta\langle A'(l)\rangle + 178.06\Delta\langle G(l)\rangle \tag{4}$$

$$Ascore(l) = 3.15\Delta\langle H(l)\rangle + 1.36\Delta\langle HPS(l)\rangle + 2.40\Delta\langle A'(l)\rangle + 42.48\Delta\langle ST(l)\rangle - 1.71 \tag{5}$$

判別解析を行った後のグリシン，側鎖の小さい極性残基，両親媒性残基のクラスタの出現比を示したものが，Fig.3 (a)-(c)である．判別された各残基クラスタは，ブレイク領域やループアコアによく現れている．これは，周辺のアミノ酸配列環境を考慮することで，二次構造ブレイカとなっている残基クラスタを判別できたことを示している．つまり，二次構造ブレイカは，残基の性質だけで決まるわけではなく，周辺のアミノ酸配列環境との組み合わせで決まるのである．

最後にタンパク質の二次構造ブレイカについてまとめると以下のようになる．

(1) 二次構造ブレイカには，少なくとも 4 つのメカニズム（プロリン，グリシン，側鎖の小さい極性残基，両親媒性残基）がある．

(2) 二次構造ブレイカは，単純にその残基の性質だけで決まるわけではなく，周辺のアミノ酸配列の環境に依存する．

(3) 二次構造ブレイカになるかどうかを決める周辺のアミノ酸配列環境は，一つの物性ではなく，複数の物性の組み合わせによってできている．

3. 二次構造ブレイカ予測システム SOSUIbreaker

4 つのタイプの二次構造ブレイカ残基とそれら周辺の配列環境をもとに二次構造ブレイカ予測システム SOSUIbreake を開発し，ウェブサイトで公開している．Fig.3(d)は，二次構造予測結果の例である．α-ヘリックス，β-ストランドといった二次構造の種類に関わらず，予測された二次構造ブレイカは二次構造の端付近に位置している．また，テストデータ 352 アミノ酸配列（配列相同性 30%以下）に対して二次構造ブレイカ予測を行った結果，予測された二次構造ブレイカの 70%は，ブレイク領域に位置し，93%がループ領域（ブレイク領域＋ループコア）に位置していた．また，予測された二次構造ブレイカは，ループ領域の 80%をカバーしていた．Fig.4 は，これらの予測精度を模式的に表したものである．この二次構造ブレイカ予測システム SOSUIbreaker は，以下の URL から利用することができる．
http://bp.nuap.nagoya-u.ac.jp/sosui/sosuibreaker/sosuibreaker_submit.html

4. おわりに

これまで二次構造ブレイカに関しては明確な定義がなされていなかった．しかし，今回，我々は，二次構造ブレイカ傾向のある残基を再考し，その周辺の配列環境を調べることで 4 つのタイプの二次構造ブレイカを見出すことができた．

タンパク質の二次構造形成は長い研究の歴史にもかかわらず未だに理解されておらず，その予測精度は 80%を超えることができないでいる．しかし，今回，予測されたブレイク部位は、93%の精度でブレイク領域またはループコア領域にあり，二次構造コア領域にはほとんど見られず，二次構造領域との区別ができている．また，これらのブレイカ残基は，二次構造の種類によらない．これは，二次構造形成の前にブレイクの位置が決まっていることを示しているのではないだろうか．このように二次構造ブレイカから二次構造形成を見直すことで，二次構造予測のブレイクスルーとなるかもしれない．現在，筆者らは，二次構造ブレイカをもとにした二次構造予測について検討中である．また，二次構造ブレイカ予測システム SOSUIbreaker は，ウェブサイトで利用できるので，今後の研究に役立てていただけたら幸いである．

5. 文献

1) Levitt, M. (1978) *Biochemistry*. **17**, 4277-4285.
2) Chou, P.Y. and Fasman G.D. (1978) *Adv Enzymol*. **47**, 45-148.
3) Aurora, R and Rose, G.D. (1998) *Pro Sci*. **7**, 21-38.
4) Colloc'h, N and Cohen, F.E. (1991) *J Mol Biol*. **221**, 603-613.
5) Dyson H.J., Wright P.E. (2005), *Nat Rev Mol Cell Biol*. **6**, 197-208.
6) Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J.,

Campen, A.M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, W., Garner, E.C. and Obradovic, Z. (2001) *J Mol Graph Model*. **19**, 26-59.

7) The Protein Data Bank http://www.rcsb.org/pdb
8) Mitaku, S., Hirokawa, T., and Tsuji T. (2002) *Bioinformatics*. **18**, 608-616.
9) Imai, K. and Mitaku, S. (2005) *BIOPHYSICS*, **1**, 55-65
10) Kyte, J. and Doolittle, R.F. (1982) *J Mol Biol*. **157**, 105-132.

Figure caption

Fig. 1 Ratio of number of amino acid clusters to that of all amino acid residues: proline (a), glycine (b), small polar residues (c) and amphiphilic residues (d). The position of 0 on the horizontal represents the end of secondary structure, and the negative side and the positive side represents secondary structure region and loop region, respectively.

Fig. 2 Levels of averages of four kinds of properties, $\langle H(j)\rangle$, $\langle HSP(j)\rangle$, $\langle A'(j)\rangle$ and $\langle ST(j)\rangle$ in three regions were compared between cluster of amphiphilic residues in secondary structure core and break regions. The square and open triangle represent the properties in the break region and secondary structure core, respectively.

Fig. 3 Ratio of the number of discriminated secondary structure breakers to that of all amino acid residues: glycine (a), small polar residues (b), amphiphilic residues (c). Results of secondary structure breaker prediction for small G-protein Rap2A (PDB:1kao) (d). <P>, <G>, <A> and <SP> represents the breakers proline, glycine, amphiphilic residues and small polar residues, respectively.

Fig. 4 The diagram illustrating the accuracy of prediction of secondary structure breakers.

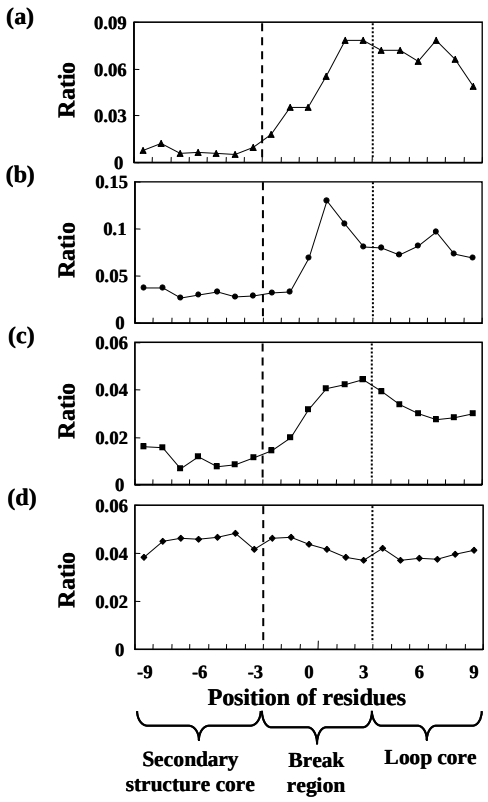

(a)
(b)
(c)
(d)
Ratio
0.09
0.06
0.03
0
0.15
0.1
0.05
0.04
0.02
-9
-6
-3
0
3
6
9
Position of residues
Secondary structure core
Break region
Loop core

Fig.1 (a)-(d)

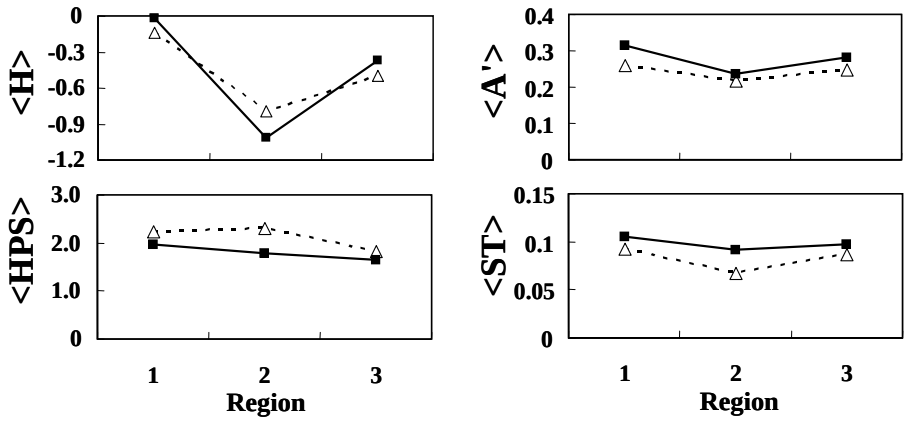
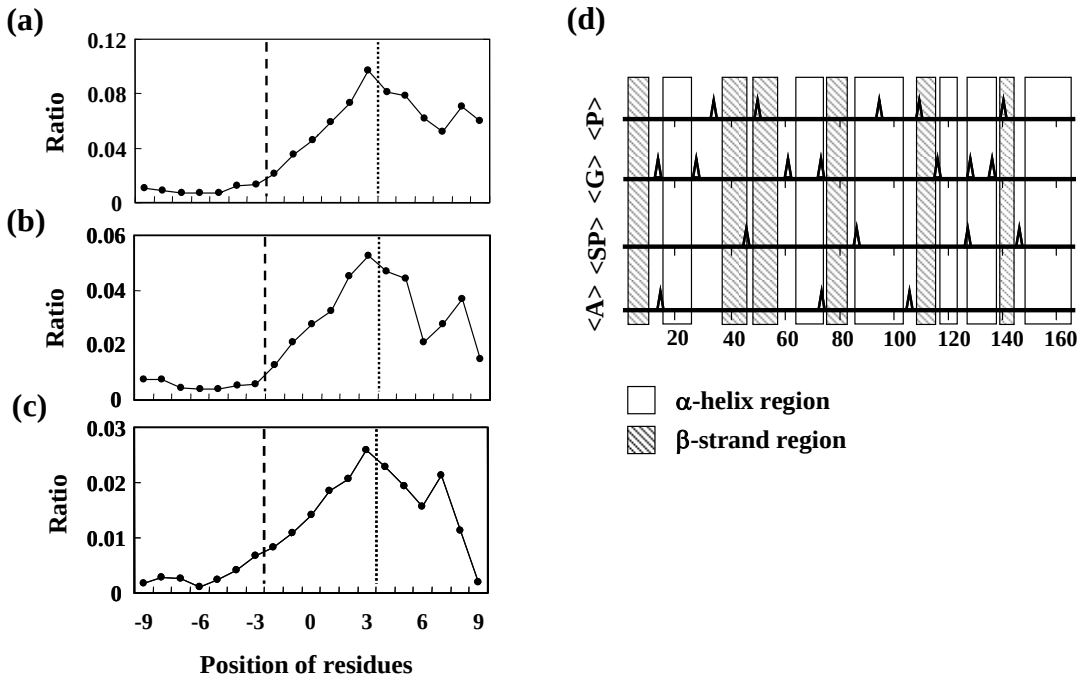
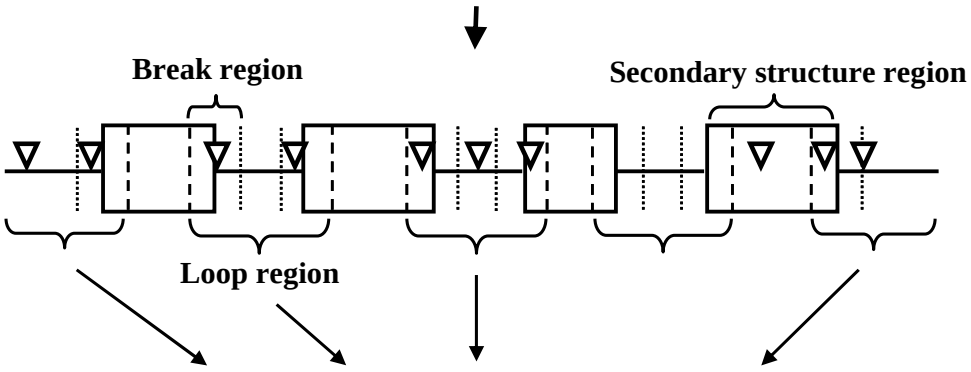

<H>
<HPS>
<A'>
<ST>
0
-0.3
-0.6
-0.9
-1.2
3.0
2.0
1.0
0.4
0.3
0.2
0.15
0.05
1
2
3
Region

Fig.2



70% of predicted breakers (all four types) were actually located in the break region, and 93% were located in the loop region.



Predicted breakers covered 80% of loop region.

▽ **Predicted breakers**

Fig. 4