コンビナトリアルタンパク質設計

日本原子力研究所・中性子利用研究センター・計算科学技術推進センター

河野秀俊

ペンシルベニア大学化学科 Jeffery G. Saven

Combinatorial protein libraries allow us to examine a huge number of sequences. Such methods are being used for *de novo* design of proteins and to investigate the determinants of protein folding. Since the diversity of possible sequences is beyond what can be explored in experiments, some restrictions have to be imposed during design of the library. Recently, theoretical tools have been developed to bias and characterize the ensembles of sequences that fold into a given structure. Such tools can be applied to the design and interpretation of combinatorial experiments and to the design of particular protein structures.

combinatorial library / protein design / sequence variability

1. はじめに

望みの機能、構造をもつタンパク質の設計・創製を試みることによって、我々ほどの程度タンパク質の物理化学を理解しているか知ることができる。また、その結果を解析することによって理解を深めることができる。タンパク質は、非常に大きい構造の自由度をもった複雑な高分子であるにもかかわらず、天然状態と変性状態の絶妙な自由エネルギーバランスで安定な構造をもち、固有の機能を発現している。さらに、20種類のアミノ酸から構成されるタンパク質の配列の数は天文学的な数字にのぼる。したがって、天然タンパク質のような自由エネルギーバランスをもった人工タンパク質を設計することは決して容易ではない。これまで、与えられた目的構造（多くの場合、天然タンパク質の構造が用いられている）に対して、アミノ酸側鎖のパッキング、疎水性残基の埋もれ具合、水素結合の数などが天然タンパク質と同じような特徴をもつように目的関数を最適化することによって配列の設計が行われてきた。しかし、対象となる配列の数は非常に多いため、1つ1つの配列を詳細に吟味することは不可能である。また、目的関数自体も計算時間を短縮するために単純化せざるをえ

ない。主鎖構造自体の設計もタンパク質設計の大きな問題であるが、この主鎖構造が与えられたとしても、非常に多くの配列とその側鎖構造を考慮する必要がある。そういった問題には、確率論的な扱い、すなわち統計力学的なアプローチが適していると思われる。本稿では、これまでの主鎖構造を仮定したタンパク質設計方法を簡単に述べ、我々の統計力学的なアプローチを紹介する。

2. エネルギー最小(極小)配列によるタンパク質設計

与えられた主鎖構造に対して、おもに側鎖構造のパッキング具合を目的関数にして、単純化したエネルギー関数のもとで最小エネルギーをもつ1つのアミノ酸配列を求めるアルゴリズムがこれまでにいくつか開発されている¹⁾³⁾。これらの方法により、Znフィンガーミミックやホメオドメイン構造などいくつかの成功例⁴⁾⁵⁾も報告されている。しかし、天然タンパク質と同様な熱力学的特性をもち、最終的に構造決定に至るものは非常に少ない。シミュレーティッドアニーリングや遺伝的アルゴリズムを用いた確率的な最適化方法では、決定される配列の初期条件依存性が強く、エネルギーが最小となる配列を探しだすことは非常に困難である。

Combinatorial Protein Design

Hidetoshi KONO* and Jeffery G. SAVEN**

* Neutron Science Research Center / Center for Promotion of Computational Science and Engineering,
Japan Atomic Energy Research Institute

**Department of Chemistry, University of Pennsylvania

また、エネルギー最小配列・構造の一部になりえない配列・側鎖構造を計算の対象から除いていく方法 (dead end elimination) では、ただだか30残基程度の小さいタンパク質に対してしか最適な配列を探しだせない⁶⁾。これらの方法で使われているエネルギー関数は、できるだけ多くの配列を考慮するために単純化されている。さらに、機能を発現するために、実在のタンパク質では安定性を犠牲にしている例もみられる⁷⁾。これらの点を考慮すると、単純化したエネルギー関数のもとで最小エネルギーをもつ配列を1本探すよりも、目的構造に矛盾しない複数の配列のライブラリーを作成し、そこから望みのタンパク質をスクリーニングするアプローチのほうが、タンパク質設計に適しているように思える。

3. コンビナトリアル設計

ファージディスプレイスクリーニング法に代表されるように、実験的にランダム配列を生成し、そこから望みのタンパク質を得るコンビナトリアルタンパク質設計は、エネルギー最小配列に基づく設計と相補的なアプローチである。コンビナトリアル設計では、鍵となる2つの要素がある。すなわち、意図したとおりの多様性をもったライブラリーの生成方法と、構造もしくは機能で生成したタンパク質をスクリーニングする方法である。現在、実験でスクリーニングできる配列空間の大きさは、 10^4 - 10^{12} といわれている⁸⁾。コンビナトリアル法では、望みのタンパク質を得るだけでなく、構造に適合する多くの新たな配列を得ることが期待されるので、配列と構造の関係がより一層明らかにされる可能性がある。また、進化的産物である天然タンパク質と人工的に得られたタンパク質の配列を比較することによって、タンパク質の折りたたみ機構やその他タンパク質の特性を理解できるかもしれない。

ライブラリーによるアプローチの例として、Hechtグループのバイナリ設計が挙げられる。このグループは、 α ヘリックスや β シートにみられる親水性、疎水性アミノ酸の出現パターンを用いてコンビナトリアル設計での探索空間を狭めることにより、4ヘリックスバンドルやアミロイドのように繊維状に凝集するタンパク質様構造の作製に成功している⁹⁾。彼らは、親水性と疎水性の単純なパターンで天然タンパク質に近いものを作製することに成功しているが、天然タンパク質のように熱変性の際に大きな ΔH をもつものはほとんどなく、NMRやX線で構造を決めるには至っていない¹⁾。これは、天然タンパク質にみられるように精密な側鎖のパッキングが実現されていないのが原因のように思える。

また、コンビナトリアル設計では、 10^4 - 10^{12} の配列を

探索することができるものの、この数字は100残基からなるタンパク質の全配列空間の 10^{130} と比べればはるかに小さい。したがって、探索空間を定量的に狭める方法論は、実験でのスクリーニング効率を上げるだけでなく、実験結果を解釈する上でも有効であると思われる。そこで、我々は統計力学的な扱いによって各残基位置でのアミノ酸の確率分布を計算し、その確率分布に従ったタンパク質設計方法を提案している。

4. 統計力学的な扱いによるタンパク質設計方法

統計力学的な扱いによるタンパク質設計方法では、特定の1つの配列を出力するのではなく、各残基位置におけるアミノ酸の確率を出力とする。すなわち、コンビナトリアルタンパク質設計でのライブラリーを理論的に設計する。もちろん、初期条件を変えてエネルギー最小化を繰り返して複数の配列を得ることによって同様に確率を求めることもできるが、計算された配列は初期条件依存性が強く、広い配列空間を探索するには非常に長時間の計算が必要である¹⁰⁾。我々の方法では、アミノ酸配列の組み合わせや側鎖の構造の組み合わせを1つずつ数えあげる必要がないので、部分的な配列設計に留まらず、タンパク質全体に対して配列を計算することができる。さらに、与えられた構造に対する配列空間の大きさやアミノ酸組成を推定することも可能である。以下、その方法論を概説する^{11)~13)}。

A. 構造エントロピー

計算の入力として、目的の主鎖構造とそれに対する配列の適合度を測る目的関数をまず与える。目的関数は、タンパク質の構造エネルギーや疎水性の指標などタンパク質全体に及ぶものから、特定のアミノ酸を特定の位置に配置させるような局所的なものが考えられる。これらの関数は、構造エントロピー最大化によってアミノ酸の確率を求める際に、拘束条件として扱われる。構造エントロピーは、残基位置 i でのアミノ酸 α が側鎖構造 (ロータマー: rotamer, rotational isomer からの造語) $r_k(\alpha)$ をとる確率を $w[\alpha, r_k(\alpha)]$ として、

$$S_c = - \sum_{i, \alpha, k} w_i[\alpha, r_k(\alpha)] \ln w_i[\alpha, r_k(\alpha)]$$

と定義する。後述する構造エネルギーなど拘束条件を f_i と表すと、上記のエントロピーを最大化するには、汎関数 V を導入して

$$\begin{aligned} V &= S_c - \beta_1(f_1 - f_1^0) - \beta_2(f_2 - f_2^0) \cdots \\ 0 &= \partial V / \partial w_i\{\{\alpha, r_k(\alpha)\}\} \\ f_m^0 &= f_m[\{w_i[\alpha, r_k(\alpha)]\}] \quad m = 1, 2, \dots \end{aligned}$$

となるアミノ酸の確率 $w[\alpha, r_k(\alpha)]$ を求めればよい。側鎖の構造は、主鎖構造依存のロータマライブラリー¹⁴⁾を使った。

B. エネルギー関数

我々は、拘束条件として2つのエネルギーを考慮した。1つは、いわゆる分子力学的なエネルギーで、静電相互作用、ファンデルワールス力、水素結合エネルギーをAMBERの力場によって計算した¹⁵⁾。ある配列 $(\alpha_1, \dots, \alpha_N)$ が $[r_1(\alpha_1), \dots, r_N(\alpha_N)]$ の構造をとるとき、構造エネルギーは、

$$E_c = \sum_i \varepsilon_i[\alpha, r_k(\alpha)] + \sum_{i,j>i} \varepsilon_{i,j}[\alpha, r_k(\alpha); \alpha', r_{k'}(\alpha')]$$

と表すことができる。ここで、 $\varepsilon[\alpha, r_k(\alpha)]$ は主鎖と側鎖の相互作用エネルギーを、 $\varepsilon_{i,j}[\alpha, r_k(\alpha); \alpha', r_{k'}(\alpha')]$ は側鎖どうしの相互作用エネルギーを、 $r_k(\alpha)$ 、 $r_{k'}(\alpha')$ はアミノ酸 α 、 α' それぞれの側鎖構造 (ロータマー) を表す。ここで、各ロータマーがまわりのアミノ酸から受けるエネルギーを平均場ポテンシャル

$$\begin{aligned} \bar{\varepsilon}_i[\alpha, r_k(\alpha)] &= \varepsilon_i[\alpha, r_k(\alpha)] \\ &+ \sum_{j \neq i, \alpha', k'} \varepsilon_{i,j}[\alpha, r_k(\alpha); \alpha', r_{k'}(\alpha')] w_j[\alpha', r_{k'}(\alpha')] \end{aligned}$$

の形で取り込むと、上記の構造エネルギーは

$$\begin{aligned} E_c \approx \bar{E}_c &= \sum_{i, \alpha, k} \bar{\varepsilon}_i[\alpha, r_k(\alpha)] w_i[\alpha, r_k(\alpha)] \\ &- \sum_{\substack{i,j>i \\ \alpha, \alpha' \\ k, k'}} \varepsilon_{i,j}[\alpha, r_k(\alpha); \alpha', r_{k'}(\alpha')] w_i[\alpha, r_k(\alpha)] w_j[\alpha', r_{k'}(\alpha')] \end{aligned}$$

と書き直すことができる。

2つ目の拘束条件として、溶媒との相互作用を考慮するために、疎水エネルギーを考慮した。アミノ酸のタンパク質内部への埋もれやすさ、逆に表面への露出しやすさを考慮するために、天然タンパク質の構造から経験的ポテンシャルを導出し、疎水エネルギーが天然タンパク質と同程度になるように拘束条件を加えた。このポテンシャルは、各アミノ酸のまわりの C_β 原子の数に依存した関数である。詳細は文献¹²⁾を参照していただきたい。

タンパク質の安定性は、天然状態と変性状態の自由エネルギー差で決まっている。したがって、何らかの形で変性状態のエネルギーを考慮する必要がある。当然ながら、1つ1つの可能な配列のそれぞれの状態の自由エネルギーを計算することは、その数の多さ、構造空間の広さから不可能である。そこで、第1次近似的に変性状態のエネルギーを見積もるために、*N*-アセチル- α -*N'*-メチ

ルアミド (α はアミノ酸) のペプチドの自由エネルギーを参照エネルギー $\gamma_{ref}(\alpha)$ として計算した。参照エネルギーは、Gly に対する相対エネルギーとして計算した。

$$\gamma_{ref}(\alpha, \beta_{ref}) = -\beta_{ref}^{-1} \ln[z_{ref}(\alpha, \beta_{ref}) / z_{ref}(G, \beta_{ref})]$$

$$z_{ref}(\alpha, \beta_{ref}) = \sum_{\phi, \varphi, k} \exp\{-\beta_{ref} \varepsilon_{ref}[\phi, \varphi, r_k(\alpha)]\}$$

ここで、 $\beta_{ref} = 1/k_B T$ 、 ϕ 、 φ はペプチドの二面角。温度 T とロータマーの数 k は前出の構造エネルギー計算の場合と同じにする必要があることを注意しておきたい。 $\gamma_{ref}(\alpha)$ を前出の E_c に加えて、最終的に構造エネルギーは以下の形になる。

$$\begin{aligned} E_c \approx \bar{E}_c &= \sum_{i, \alpha, k} \{\bar{\varepsilon}_i[\alpha, r_k(\alpha)] - \gamma_{ref}(\alpha, \beta_{ref})\} w_i[\alpha, r_k(\alpha)] \\ &- \sum_{\substack{i,j>i \\ \alpha, \alpha' \\ k, k'}} \varepsilon_{i,j}[\alpha, r_k(\alpha); \alpha', r_{k'}(\alpha')] w_i[\alpha, r_k(\alpha)] w_j[\alpha', r_{k'}(\alpha')] \end{aligned}$$

我々の理論では、これらのエネルギーの拘束条件のものと、構造エントロピーを最大化することによって、各アミノ酸のそれぞれのロータマーの確率を self-consistent に得る。各位置でのアミノ酸の確率は、各アミノ酸内でのロータマーの確率の和をとって以下のように計算される。

$$w_i(\alpha) = \sum_k w_i[\alpha, r_k(\alpha)]$$

5. 適用例

以下、我々の理論を protein L の IgG 結合ドメイン (嫌気性バクテリア由来) と c-Crk SH3 ドメイン (がん原遺伝子産物のアダプター分子) に適用した結果を示す。

A. protein L の IgG 結合ドメイン

Baker グループは、ファージディスプレイによってランダムに変異を入れた protein L の IgG 結合ドメインを発現させ、IgG との結合によって野生型と同じ立体構造をもつタンパク質をスクリーニングしている^{16), 17)}。ここで得られた配列は、ほとんどまったく進化的な影響を受けていない。protein L の場合、相同な配列がほとんどないため、実験的に得られた配列を進化的に得られた配列と比較することはできない。我々は、理論的に得られたアミノ酸の確率と実験的にスクリーニングされた配列の比較を行った。Fig.1, Fig.2 に示すように、計算、実験ともに、第1ストランド上の残基のうちタンパク質内部に向かって側鎖が伸びている残基では、許されるアミノ酸の種類および側鎖の構造が限定されるのに対し (赤色) タンパク質表面に露出する残基では多くの

Fig.1 Side chain conformations of β -strand 1 in protein L. The side chains of amino acid residues with high probability are rendered along with the wild type structure(white). Side chain conformations shown are most probable ones for each residue site. Color indicates the probability of the amino acids: for dark blue $w(\alpha) < 0.1$, and for dark red $w(\alpha) > 0.25$. For clarity strand 2 is not shown. The figure is prepared with Molscrip²¹⁾ and Raster3D²²⁾. (カラー図は電子ジャーナル参照.)

親水性アミノ酸が許され、アミノ酸の種類、構造に対する特異性は低い(青色)という結果が得られた。その傾向はアミノ酸を物理化学的特性に基づいて5つないし2つに分類することによって、より顕著にみることができる(Fig.3 a, b)。これは、実験的に選択された配列が、構造安定性を保つためには天然型配列と同じような物理化学的特性をもつことが必要であるということを示していると考えられる。また、残基の溶媒露出面積と実験的に得られたアミノ酸の多様性(エントロピー)が高い相関を示すことから、埋もれている残基は、露出している残基よりも構造に対する情報をより多くもっていることを示唆している^{12) 18)}。

B. SH3 ドメイン

我々の方法をタンパク質全体に対して適用した例として、SH3 ドメインの結果を示す。SH3 ドメインはキナーゼ、リパーゼ、アダプタータンパク質など非常に多様なタンパク質構造にみられる。SH3 ドメインに関しては、多くの立体構造、配列が既知であり、置換体などの安定性に関する実験的な情報も十分にある。c-Crk SH3 配列に対するマルチプルアライメントから得られたアミノ酸の確率分布と計算結果をFig.4に示した。タンパク

Fig.2 Probability of occurrence of amino acid residues for β -strand 1 of protein L. Calculated probabilities : open bars, experimentally observed probabilities : filled bars. Error bars show standard deviations of probabilities from 21 different backbones of protein L.

質のコアにあたる位置において、両者の確率は非常に良い一致を示し、コア位置においてアミノ酸選択性の高いことが示された。ここでも立体構造情報がコアにあたる残基により多く含まれていることが示唆された。このように、計算された確率とマルチプルアライメン

Fig.3 (a) Probabilities of occurrence of five groups of amino acid residues for β -strand 1 of protein L. Residues are classified into the following five groups : (A, C, I, L, M, V, F, W, Y), (H, S, T, N, Q), (K, R), (D, E), (G, P) and they correspond to bars from left to right at each site. (b) Hydrophobic positioning of amino acid residues for the calculated (open bar) and the experimentally obtained sequences (filled bar). Each "non-polar" probability is the sum of the probabilities of residues L, I, M, V and F at each site. In a similar way, the "polar" probability is the sum of the probabilities of residues R, D, N, Q, E, K and H.

トから得られた確率とを比較することによって、保存残基の立体構造上の重要性を推定することができる。

6. おわりに

以上、2つのタンパク質を例に、アミノ酸配列のライブラリーを設計する方法論を紹介した。この方法では、統計力学的な取り扱いにより、大量のアミノ酸配列を扱うことができ、側鎖を陽に扱うことによって天然タンパク質にみられるような精密な側鎖のパッキングを実現する配列を生成することができる。また、得られたアミノ酸確率に従って、モンテカルロ法による配列探索の採択・不採択を決定することにより、効率的に目的構造に矛盾しない配列を見つけることができる¹⁹⁾。また、計算されたアミノ酸確率を再現するコドンの設計方法も確立されているので、遺伝子の設計も可能である²⁰⁾。

しかしながら、まだ課題も多い。現在のほとんどの方法論はタンパク質の天然状態を最適にするアミノ酸配列の設計に中心をおいている。しかし、設計された配列にとって目的構造が最小エネルギー構造である保証はない。目的構造とは異なる構造では不安定になる、ネガティブ設計も必要であろう。この理論に基づいてライブラリーを設計し、タンパク質を作製する実験は、現在進行中である。その結果が出れば、この方法論の有効性および欠点がより明らかになるであろう。

文 献

- 1) Kraemer-Pecore, C. M., Wollacott, A. M. and Desjarlais, J. R. (2001) *Curr. Opin. Chem. Biol.* **5**, 690-695.
- 2) Street, A. G. and Mayo, S. L. (1999) *Structure* **7**, R105-R109.

Fig.4 Probability of occurrence of amino acid residues in the SH3 domain (1CKA). Calculated probabilities (filled bars) ; sequence alignment-based probabilities (open bars). F10, L18, F20 and L26 are located at hydrophobic core sites with solvent accessibility less than 20%.

- 3) Saven, J. G. (2001) *Chem. Rev.* **101**, 3113-3130.
- 4) Dahiyat, B. I. and Mayo, S. L. (1997) *Science* **278**, 82-87.
- 5) Marshall, S. A. and Mayo, S. L. (2001) *J. Mol. Biol.* **305**, 619-631.
- 6) Voigt, C. A., Gordon, D. B. and Mayo, S. L. (2000) *J. Mol. Biol.* **299**, 789-803.
- 7) Ogata, K., *et al.* (1996) *Nat. Struct. Biol.* **3**, 178-187.
- 8) Hoess, R. H. (2001) *Chem. Rev.* **101**, 3205-3218.
- 9) Moffet, D. A. and Hecht, M. H. (2001) *Chem. Rev.* **101**, 3191-3203.
- 10) Larson, S. M., *et al.* (2002) *Protein Sci.* **11**, 2804-2813.
- 11) Zou, J. and Saven, J. G. (2000) *J. Mol. Biol.* **296**, 281-294.
- 12) Kono, H. and Saven, J. G. (2001) *J. Mol. Biol.* **306**, 607-628.
- 13) Saven, J. G. and Wolynes, P. G. (1997) *J. Phys. Chem. B* **101**, 8375-8389.
- 14) Dunbrak, R. and Cohen, F. E. (1997) *Protein Sci.* **6**, 1661-1681.
- 15) Weiner, S. J., *et al.* (1984) *J. Am. Chem. Soc.* **106**, 765-784.
- 16) Kim, D. E., Gu, H. D. and Baker, D. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 4982-4986.
- 17) Gu, H., *et al.* (1999) *Protein Sci.* **8**, 2734-2741.
- 18) Reidhaar-Olson, J. F. and Sauer, R. T. (1988) *Science* **241**, 53-57.
- 19) Zou, J. and Saven, J. G. (2003) *J. Chem. Phys.* **118**, 3843-3854.
- 20) Wang, W. and Saven, J. G. (2002) *Nucleic Acids Res.* **30**, e120-e127.
- 21) Kraulis, P. J. (1991) *Appl. Crystallog.* **24**, 946-950.
- 22) Merritt, E. and Bacon, D. (1997) *Methods Enzymol.* **277**, 505-524.

河野秀俊

河野秀俊 (こうの ひでとし)

日本原子力研究所 研究員

東京大学大学院農学生命科学研究科応用生命工学専攻博士課程終了, 博士(農学) 理化学研究所基礎科学特別研究員, ペンシルベニア大学化学科ポスドクフェローを経て, 2002年7月より現職。

研究内容: タンパク質 - DNA 相互作用, タンパク質設計, モデリング, ゲノム情報解析など計算機を使って幅広くバイオインフォマティクス研究を行っていきたくと思っています。

連絡先: 〒619-0215 京都府相楽郡木津町梅美台 8-1

E-mail: kono@apr.jaeri.go.jp

Jeffery G. Saven (ジェフリー セイブン)

Department of Chemistry, University of Pennsylvania, Assistant Professor

研究内容: Statistical mechanics, molecular simulation, protein design, combinatorial libraries, protein folding, non-biological foldamers

連絡先: 231 South 34th Philadelphia, PA19104, USA

E-mail: saven@sas.upenn.edu

URL: <http://www.sas.upenn.edu/~saven/>