蛋白質のデザインおよび進化:
生命と物質の間のボトルネックを抜ける

大阪大学 サイバーメディアセンター 大規模計算科学研究部門　時田 恵一郎[1]

蛋白質等のヘテロ高分子を統計力学的に設計 (デザイン) する方法について報告する。ここで検討する手法は、Kurosky と Deutsch による設計基準 [1, 2]、すなわち、設計したいターゲット構造をとる確率を最大化するという設計基準に基づくものである。この新しい設計手法 (「設計方程式法 [3, 4]」と呼ぶ) の特徴は、もともと２０種類の値をとる離散的なアミノ酸配列変数を連続化して表現することにより、決定論的な力学方程式を用いた最適化を行うという点にある。また、各モノマーに対応する配列変数が並列に時間発展するので、従来手法に比べて高速に収束し、スケーラビリティーが期待できる。本稿では、まず設計問題自体についての解説を行い、統計力学に基づく代表的な設計手法を概説する。さらに、HP モデル [22, 24, 25] と呼ばれる単純化された蛋白質モデルにおける設計問題に設計方程式法を応用し、シミュレーテッド・アニーリングを用いた従来手法とのパフォーマンスの比較を行った結果を示す。

1　設計問題とは何か

蛋白質やヘテロ高分子の設計問題は、新薬開発、 分子機械設計、機能性高分子材料開発など、薬学、工学にもまたがる巨大な応用分野に関わっている。ここでいう「設計」とは、与えられた三次元「構造」へと折り畳むアミノ酸の「配列」を決定する問題である。これがどうして設計なのだろうか？　もしも、薬として作用する蛋白質や、ある形状や組織をもつ分子機械や高分子材料が必要になったとしたら、それらの作用、形状、組織を実現する三次元構造をとる高分子を作る必要がある。しかし、例えば、いきなり蛋白質を針金のように折り曲げて作るわけにはいかない。なぜならタンパク質の構造は、膨大に連なったアミノ酸モノマー (と水) の間に働く多様で複雑な相互作用による「折り畳み」を通じて創発するものだからである。だから、必要な立体構造を安定に保つような配列を見つけ出さなければならないのである。以下では、ここでの問題意識に関わる高分子を代表して蛋白質という言葉を用いるが、基本的なアイディアはヘテロ高分子一般に適用可能である。

蛋白質の設計問題は折り畳み問題の逆問題になっていることに注意しよう。折り畳み問題は物理化学研究の中心課題の一つであるが、折り畳みの全ステージを分子動力学シミュレーションで調べることは現在でもほとんど不可能である。さらに、ここで考える設計問題は、アルゴリズム論的には順問題である折り畳み問題を内側に含むため、計算コストが折り畳み問題にも増して高

[1] URL: http://wwwacty.phys.sci.osaka-u.a.jp/~tokita

い。また、原理的に「解がない」場合もある (与えられた立体構造を安定に保つようなアミノ酸配列が必ず存在するという保証はない。例えば、ひと筆書きで書いたあなたのサインの形の蛋白質を設計できるだろうか?)。むしろほとんどの場合には解がなくて、解を持つような特別な立体構造を私達は蛋白質と呼んでいるのである。そのため、現在では、現実的な分子量の蛋白質まるごと一個を計算機で設計することはほとんど不可能である。従来は、蛋白質の一部分に対する、発見的・経験論的な設計方針が研究されてきた [5, 6, 7, 8, 9, 10]。一方最近になって、(単純なモデルではあるが) 一個の蛋白質全体に対する設計方針が、統計物理、計算物理、組み合わせ最適化問題などの観点から研究されるようになってきた [11, 12, 13, 14, 15, 16, 1, 2, 17, 20, 3, 21, 4] 。以下では、「難しい順問題を内側に含む逆問題」という観点から蛋白質設計の最近の手法を分類・概観し、それらとは異なる新しいアプローチを紹介する。

ここでは蛋白質の設計問題を、逆温度 β において、与えられたターゲット構造 $\boldsymbol{R} = \{R_1, R_2, \ldots, R_N\}$(各 R_i は i 番目アミノ酸の位置ベクトル) をとる確率

$$P_\beta(\boldsymbol{R}|\boldsymbol{\sigma}) = \frac{1}{Z_\beta(\boldsymbol{\sigma})} \exp\left[-\beta E(\boldsymbol{R}, \boldsymbol{\sigma})\right] , \tag{1}$$

$$Z_\beta(\boldsymbol{\sigma}) = \sum_{\boldsymbol{r}} \exp\left[-\beta E(\boldsymbol{r}, \boldsymbol{\sigma})\right] \tag{2}$$

を最大にするアミノ酸配列 $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$($\sigma_j$ は j 番目アミノ酸のタイプで一般的には２０種のアミノ酸に対応して２０種の値をとる) を探す問題として定義する。N はアミノ酸の数、すなわちで蛋白質の鎖の長さを表す。$E(\boldsymbol{r}, \boldsymbol{\sigma})$ は、アミノ酸配列 $\boldsymbol{\sigma}$ からなる蛋白質が、ある構造 $\boldsymbol{r} = \{r_1, r_2, \ldots, r_N\}$($r_j$ は j 番目アミノ酸の位置ベクトル) をとったときのエネルギーである。和 $\sum_{\boldsymbol{r}}$ は可能なすべての構造についてとる。この設計方針を「ターゲット確率最大化 (Maximizing Target Probability (MTP)) 基準」と呼ぶことにする。MTP 基準は、折り畳んだ状態が熱的平衡にあることを用いた最も基本的な設計方針であるといえるが、この方針には計算論的な困難があることに注意しよう。というのも、探索すべき配列空間の各点 $\boldsymbol{\sigma}$ に依存した構造 $\boldsymbol{r}$ 空間での状態和 (2) の計算が必要なので、結局構造と配列の両方の空間中での最適化をしなければならないからである [20]。原理的には無限個の構造と 20^N 個の可能な配列があることから、いかに設計問題のコストが高いかがうかがえよう。状態和 (2) がわかると任意の構造の出現確率がエネルギーの計算コストのオーダ ($\sim N^2$) で即座にわかり、どの構造へと折り畳むかがわかる。つまり、状態和の計算が「折り畳み問題」を解くことと等価であり、この意味で設計問題は「折り畳み問題を含んでいる」のである。

設計問題には「解がない場合がある」と既に述べたが、複数ある場合もある。この解の個数を「デザイナビリティ (designability)」と呼ぶ。デザイナビリティは設計問題の難易度に関わっている。複数ある解のそれぞれは、いわゆる「相同蛋白質」であり、例えば同じ機能を果たすヘモグロビンの配列が生物種ごとに少しず異なっているような情況に対応している (異なる種のヘモグロビンが全く同じ立体構造をもつわけではないが、ここでは簡単に、機能が同じ蛋白質は立体構造も同じであるとする)。進化はアミノ酸の置換、すなわち配列空間上での変異とみなすことができるので、デザイナビリティは突然変異に対する蛋白質の機能の耐性の度合であり、進化的安定性とも関係がある。この蛋白質の進化的安定性が熱力学的安定性 (折り畳みやすさ) と相関している

というモデル研究もある [23]。このようにタンパク質の設計問題は、物理化学、進化生物学、計算科学と密接に関係している。

現実的なモノマー数（数百から数万）の蛋白質の設計問題を、全ての可能な配列、全ての可能な立体構造を枚挙することで解くことが不可能なのは明らかであろう。つまり、この問題に対しては、多かれ少なかれ発見的な手法が要求される。最もナイーブな方法は、ランダムにアミノ酸配列を生成して、それが解になっているかどうか折り畳ませてみて確認し、ダメなら一部のアミノ酸を変更して (=突然変異) 同じことを繰り返し、(1) 式の確率を大きくするようならばアクセプトするという方法であろう。これはまさにモンテカルロ法であり、突然変異率を連続的に下げるということを行えばシミュレーテッド・アニーリング法になる。ランダムな初期アミノ酸配列は、与えられた立体構造を安定にとることはほとんど期待できない。薬や酵素を設計しようという場合には、必要な立体構造をもつ蛋白質が細胞と相互作用する「生命物質」であるのに対して、ランダムなアミノ酸配列はほとんどの場合、基底状態が縮退していて単一の立体構造をとらない非生命物質であるといえる。つまり、設計問題は、非生命的な「物質」から「生命」的な物質への「進化」を実現することに他ならない。副題の「ボトルネック」の意味は、この「進化」の道筋が隘路になっているということである。

「蛋白質やヘテロ高分子を設計する」とは物理では聞きなれない表現かも知れない。「設計」という言葉が目的論的な響きを持つからだろう。しかし、例えば熱力学の誕生が熱機関の効率化についての考察を契機とするものであったことを思い出せば、同じ問題が科学の顔と工学の顔を同時に持つことが珍しくないということに気付くであろう。

2 設計モデル

現実的な蛋白質の設計を行うことは現状では大変困難であるが、ここ１０年程の間に、単純なモデルを使うことによって、設計問題を、経験論ではなくトップダウン的および理論的に扱う研究が現れてきた。この意味で、統計力学的な洞察と計算物理学的なテクニックがうまく融合した成功例が、Shakhnovich と Gutin らによる方法 (SG 法)[11, 12] であろう。彼らは、蛋白質がほどけた状態には自己平均性があること、すなわち、ほどけた状態のエネルギーの分布は個々の配列には依存せず、アミノ酸の組成にのみ依存することを全く異なる理論により導き、それを設計問題に応用した。それにより、アミノ酸組成が一定の条件下では、アミノ酸の置換に対して状態和 $Z_\beta(\boldsymbol{\sigma})$ が不変になるので、問題がターゲット構造の単純なエネルギー $E(\boldsymbol{R}|\boldsymbol{\sigma})$ の最小化に帰着される。HP モデル [22, 24, 25](２０種のアミノ酸を疎水性 (H) と親水性 (P) の２種類にわけ、σ_j が２種類の値をとるモデル) に対する SG 法は、統計力学の言葉でいえば、一定磁化のもとでの強磁性 Ising モデルと等価であり、高速に緩和する (律速である状態和の計算をまるっきりやらないので)。組成一定の条件は重要である。これがないとホモポリマー (全部のアミノ酸が疎水基) に収束してしまうからである。ホモポリマーは全ての構造のエネルギーが等しいので (全状態が縮退)、ターゲット構造以外へも折り畳んでしまうことになるからである。SG 法は様々な現実的な蛋白質設計にも応用されて成功している [13, 16, 18] が、(解があることが保証されているような

問題に対しても) 解に到達しない場合があることが指摘され、さらに最適な組成を先見的に知る方法がないことから、より精確な設計のためにはやはり状態和の計算は避けられないということになり、その後様々な改良アルゴリズムが提出されることになった。ここで注意したいのは、改良アルゴリズムを提案する論文においては、SG 法の正解率が低いことがしばしば強調されるけれども、「折り畳み」部分、すなわち設計問題において律速となる部分をを全く計算しないという、圧倒的な「収束の速さ」についてはほとんど指摘されないことである。次節でも述べるように発見的な解法にしても時間をかければ正解率を上げることは可能であるから、本来は、解を得るための計算コストを比較しなければならないが、そのような比較研究がほとんどなされないのは問題ではないだろうか。

MTP 基準の重要性を最初に指摘したのは、Kurosky と Deutsch (KD)[1, 2] である。彼らは、コンパクトな (=ほどけていない) 構造 $\boldsymbol{r}^c$ 群に等重率を仮定して、平均エネルギー $\sum_{\boldsymbol{r}^c} E(\boldsymbol{r}^c|\boldsymbol{\sigma})/\sum_{\boldsymbol{r}^c} 1$ によって、計算すべき自由エネルギー $F_\beta(\boldsymbol{\sigma}) = -\log Z\beta(\boldsymbol{\sigma})/\beta$ を近似する手法を提案した。これは、自由エネルギーのキュミュラント展開の最低次の項だけを取り出したことに対応する。DK 法は、SG 法では見つけることのできなかった解を見出したが、１００％解に到達することはなかった。続いて Morrissey と Shakhnovich[17] は、高次のキュミュラント項を考慮した手法を提案したが、解への収束のパフォーマンスについての従来手法との比較は行っていない。Seno らの方法 [15] は、MTP 基準に忠実に $P_\beta(\boldsymbol{R}|\boldsymbol{\sigma})$ そのものを最大化した最初の例である。状態和は鎖成長モンテカルロ法によるサンプリングで計算し、配列空間をシミュレーテッドアニーリングで探索する。この手法では、DK 法や SG 法が見つけられなかった解が見つかり、$N = 16$ の 2 次元格子 HP モデルに対しては１００％正解に到達すると報告されている。現在、最も洗練された方法は、Irbäck らによる「マルチ配列 MC[20, 21]」であろう。この手法では、配列空間と構造空間の探索を同時に行い、Seno らの方法のような「ナイーブな二重 MC」よりも、高速かつ正確に解を探索すると報告されている。ただし、MTP 基準に忠実にモンテカルロ法を用いる限り、状態和の計算が正確で、非常にゆっくりとアニーリングすれば正解には必ず達するのだから、正解率だけで手法の良し悪しを比較することにそれほど意味があるとは思われない (実際 Irbäck らは SG 法と彼らの方法の正解率と計算時間を表にしているが、(正解率/計算時間) は SG 法の方が大きかったりする)。フェアな比較のためには正解に到達するにはどれくらい計算コストがかかるかというファクターを考慮すべきである。これについてはあとで我々の提案する手法を含め検討する。

3 設計方程式

ここでは格子上の「一般化された HP モデル [22, 24, 25]」に対して「設計方程式」、すなわち力学的なアプローチを応用した例について解説する。ただし、これは決して設計方程式がこれ以外のモデルに適用できないという意味ではなく、より現実的なモデルにも応用可能である。ここではアミノ酸のタイプは疎水性 (H) と親水性 (P) に対応して、 $\sigma_i = \{1(\mathrm{H}), -1(\mathrm{P})\}$ と表現する。アミノ酸はある一定の距離の間にある場合 (「コンタクトしている」という)、$U(1,1) = \epsilon_1$, $U(1,-1) = U(-1,1) = \epsilon_2$, and $U(-1,-1) = \epsilon_3$ で表される相互作用 U をもつ ($\epsilon_1, \epsilon_2, \epsilon_3$ はパラ

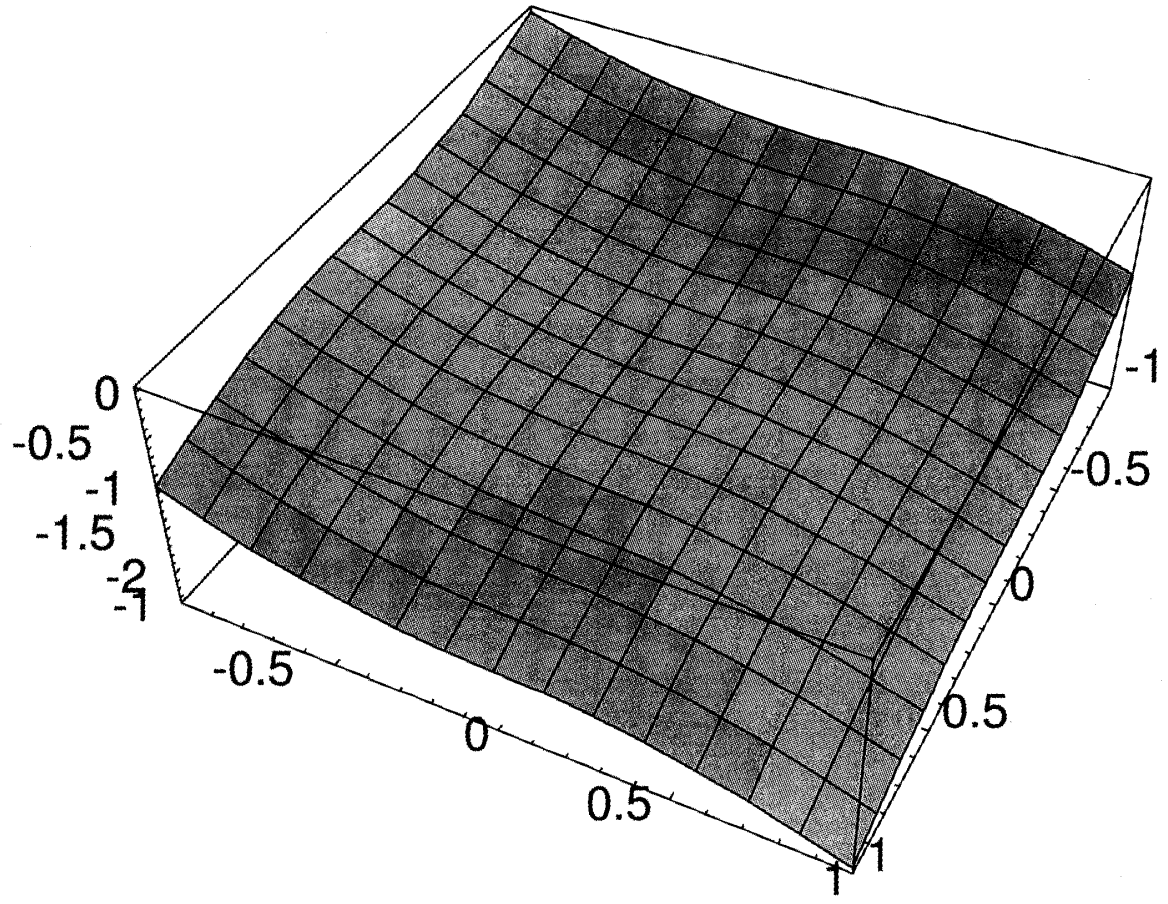

図 1: HP モデル ($\epsilon_1 = -1, \epsilon_2 = \epsilon_3 = 0$) に対する連続化コンタクトエネルギーの一つ $U^2(m_i, m_j)$。二つの水平軸は m_i および m_j を表し、垂直軸が $U^2(m_i, m_j)$ を表す。

メータであり、例えばオリジナルの HP モデルでは $\epsilon_1 = -1, \epsilon_2 = \epsilon_3 = 0$)。よって、配列 $\boldsymbol{\sigma}$ の蛋白質が構造 $\boldsymbol{r}$ をとっているときのエネルギーを

$$E(\boldsymbol{r}|\boldsymbol{\sigma}) = \frac{1}{2}\sum_{ij} U(\sigma_i, \sigma_j)\,\Delta(r_i - r_j)\ , \tag{3}$$

と書くことができる。$\Delta(r_i - r_j)$ は r_i と r_j がコンタクトしているときは 1 となり、それ以外では 0 となる (鎖にそって結合しているアミノ酸間も 0)。

ここで、Ising 変数である各 σ_i を仮想的に連続化して、新しい変数 $\boldsymbol{m} = \{m_i; -1 \le m_i \le 1\}$ で書くことにする。変数 m_i の非整数値には物理的な意味はないが、このような連続化が組合せ最適化問題に対して有効なツールになることが指摘されている [26] し、連続化がエネルギーランドスケープの凹凸を滑らかにする効果をもつことが統計力学的な解析によっても示されている [27, 28]。$\boldsymbol{\sigma}$ を「連続化モノマー変数」$\boldsymbol{m}$ で単純に置きかえて、エネルギー (3) を $E^*(\boldsymbol{r}|\boldsymbol{m}) = \frac{1}{2}\sum_{ij} U^*(m_i, m_j)\Delta(r_i - r_j)$ と書くことにする。この関数は、全ての i に対して $|m_i| = 1$ においてだけ元のエネルギー関数 (3)

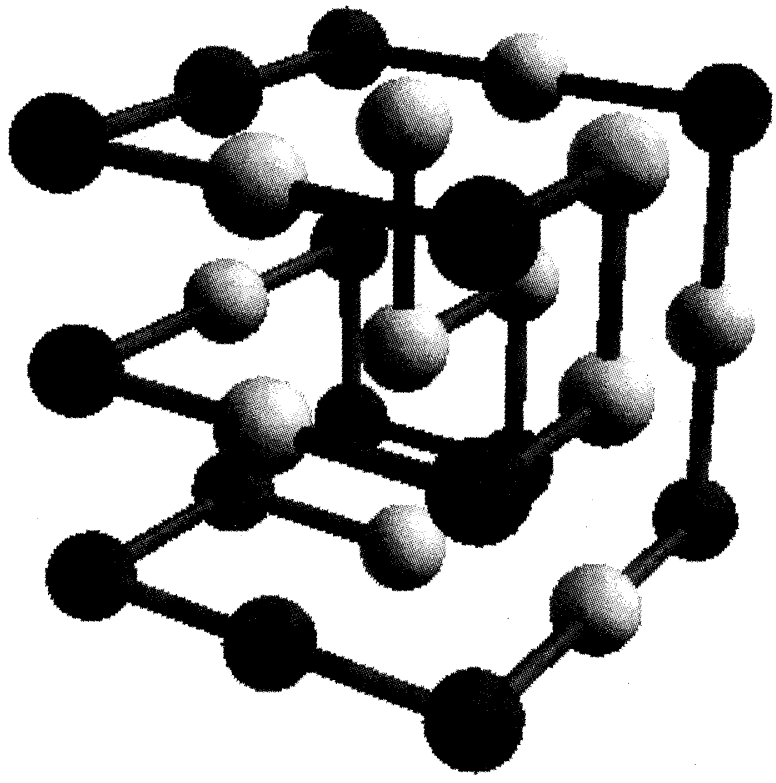

図 2: $3 \times 3 \times 3$ の格子上の一般化された HP モデル ($\epsilon_1 = -2.3, \epsilon_2 = -1, \epsilon_3 = 0$) で最も解の個数が多い構造 [23]。明るい (暗い) 球が疎水基 (親水基) を表す。

と等しくなればよい。ゆえに $E^*(\boldsymbol{r}|\boldsymbol{m})$ には不定性があって無限の「連続化エネルギー」の候補がある。ここでは、最も単純な 2 つの例

$$U^1(m_i, m_j) = \epsilon_a m_i m_j + \frac{\epsilon_b}{2}(m_i + m_j) + \epsilon_c, \tag{4}$$

$$U^2(m_i, m_j) = \epsilon_a m_i m_j + \frac{\epsilon_b}{2}(m_i|m_i| + m_j|m_j|) + \epsilon_c \tag{5}$$

を考えることにする。ここで、$\epsilon_a \equiv (\epsilon_1 - 2\epsilon_2 + \epsilon_3)/4$, $\epsilon_b \equiv (\epsilon_1 - \epsilon_3)/2$ and $\epsilon_c \equiv (\epsilon_1 + 2\epsilon_2 + \epsilon_3)/4$ とすると、上記の端点での条件 ($E^*(\boldsymbol{r}|\boldsymbol{m}) = E(\boldsymbol{r}|\boldsymbol{\sigma})$ for $\boldsymbol{m} = \boldsymbol{\sigma}$) が満たされる。$U^2(m_i, m_j)$ のエネルギーランドスケープを図 1 に示す。

連続化エネルギー関数 $E^*(r|m)$ を式 (1) に代入し、非物理的な $\{m_i\}$ の非整数値への収束を避けるための罰金項を加えた最小化すべきコスト関数を、

$$V(\boldsymbol{m}) = -\log P(\boldsymbol{R}|\boldsymbol{m}) + \frac{\lambda}{4}\sum_i (m_i^2 - 1)^2 , \tag{6}$$

とする。右辺第二項が罰金項であり、$\lambda(> 0)$ は罰金項の強さを表すパラメータである。コスト関

数を m_i で微分することにより、

$$\tau\frac{dm_i}{dt} = -\frac{\partial V}{\partial m_i} = f_i(\beta, \boldsymbol{m}) - \lambda m_i(m_i^2 - 1), \tag{7}$$

$$f_i(\beta, \boldsymbol{m}) = \beta\sum_j \frac{\partial U^*(m_i, m_j)}{\partial m_i} \times [\Delta(R_i - R_j) - \langle\Delta(r_i - r_j)\rangle_\beta] \tag{8}$$

を得る。式 (7)-(8) を「設計方程式」と呼ぶ。ここで変数 t は模擬的な時間であり、定数 τ は時間スケールを決める。これは、コスト関数 (6) を最急降下法で解くことに外ならない。駆動力 f_i の $[\cdots]$ の中の第一項 $\Delta(R_i - R_j)$ はターゲットのコンタクト (目標) であり、第二項は状態 $\boldsymbol{m}$ によって実現されているコンタクト $\Delta(r_i - r_j)$ のカノニカル平均 $\langle\Delta(r_i - r_j)\rangle_\beta = \sum_{\boldsymbol{r}} \Delta(r_i - r_j)P_\beta(\boldsymbol{r}|\boldsymbol{m})$ である。設計方程式においては、両者の差が小さくなるように状態 $\boldsymbol{m}$ が時間変化し、両者が一致すれば正解に到達したことになる。コンタクトのカノニカル平均がターゲットのコンタクトに等しいということは、ほとんどの時間にわたって系がターゲット構造をとること (すなわちターゲット構造に折りたたむこと) に他ならないからである。このような駆動項の形式は、ボルツマンマシン学習とそれと等価である。設計方程式では最適化すべきコスト関数がターゲットのボルツマン重率で、ボルツマンマシン学習においては Kullback divergence になっている。つまり、ボルツマンマシンが、与えられた (スピン表示された) 神経発火パターンへと収束する相互作用を学習するのに対して、与えられた構造 (相互作用を決める) へと収束するような連続化されたアミノ酸配列を「学習」する (よって、本手法をヘテロポリマー設計に対する「学習方程式法」と呼ぶことも可能である)。時間 t を増大させつつコントロールパラメータ λ をゆっくりと無限大まで増大させていけば連続化されたアミノ酸配列変数 m_i は ± 1 へと収束する。このとき $\partial V/\partial m_i = 0$ であれば、上記の議論より与えられた構造 $\boldsymbol{R}$ が縮退のない基底状態 (天然構造という) となることは明らかである。ゆえに、十分低温では設計方程式の収束した先の $\boldsymbol{m}$ が元の HP モデルでの設計問題の適切な解 (の候補) となりうる。

具体的な計算過程は以下の通りである。

1. 初期化

   ランダムに $m_i(t=0)$ を初期化する。罰金項の強さ λ にも適当な値を与える。

2. カノニカル平均の計算

   $\langle\Delta(r_i - r_j)\rangle_\beta$ を厳密枚挙もしくはモンテカルロ法などにより計算する。厳密枚挙は格子モデルにおいてさえ非常に短い鎖 ($N < 27$) に対してのみ可能である。

3. 以下の離散化した設計方程式を繰り返す

$$f_i(t) = \beta\sum_j \frac{\partial U^*(m_i(t), m_j(t))}{\partial m_i(t)} \times [\Delta(R_i - R_j) - \langle\Delta(r_i - r_j)\rangle_\beta]$$

$$m_i(t+1) = m_i(t) + \delta_t\left\{f_i(t) - \lambda m_i(t)(m_i(t)^2 - 1)\right\}$$

4. はみ出した値を切り取る

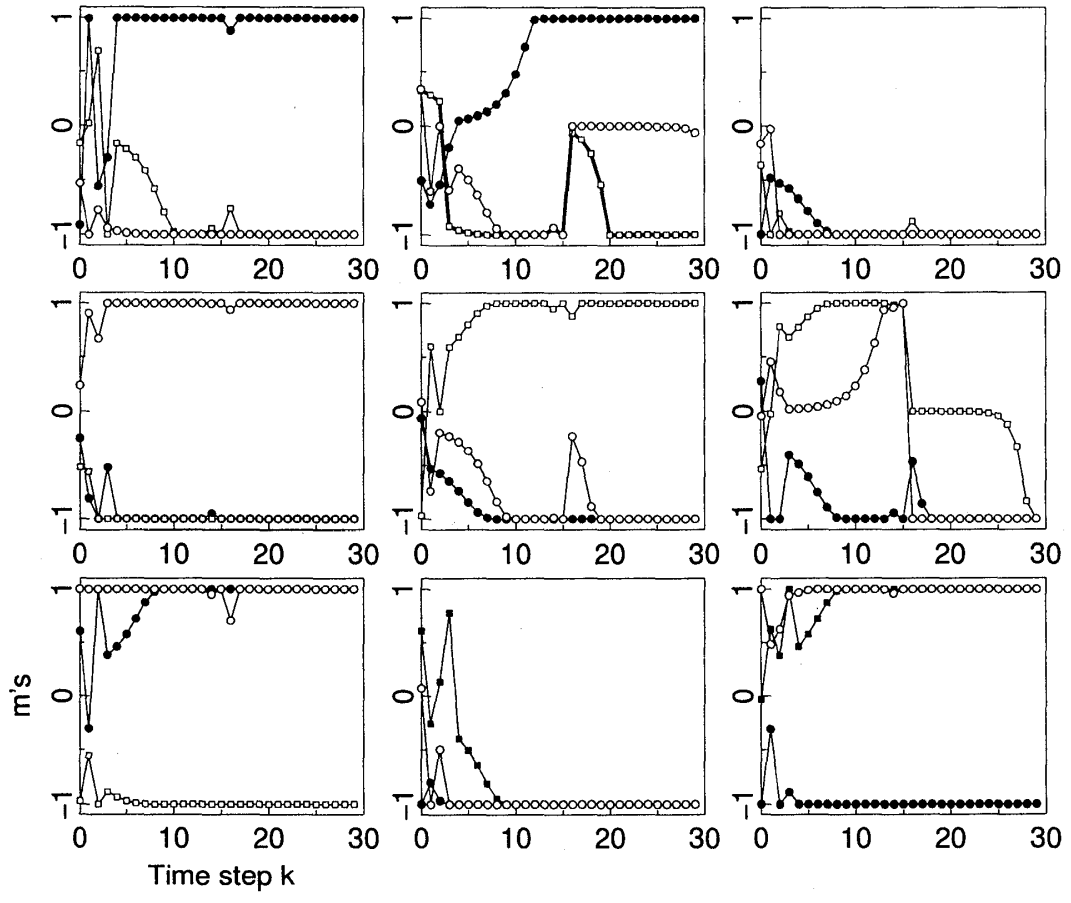

図 3: ランダムな初期状態から出発して正解に達した例。各ステップで強制的に二値化 (急冷) して「答え合わせ」をしてみると、２０ステップ目で正解のベーシンに入ったことがわかる。

もしも $m_i(t) > 1$ ならば $m_i(t) = 1$ とし、
$m_i(t) < -1$ ならば $m_i(t) = -1$ とする。

5. $\lambda(t+1) = \lambda(t) + a$(定数) とする。適当な収束条件を満たせば、終了。そうでなければステップ２へ。

離散時間の各ステップ (時間刻み $= \delta_t$) で、コンタクトのカノニカル平均は一度だけ計算され、それが全ての連続化アミノ酸変数 m_i の更新に使われることに注意しよう。カノニカル平均の計算コストは分配関数のそれとほぼ同じであるから、設計方程式法においては、律速となる計算１回ごとに、最適化すべき変数が並列的に時間発展する。一方文献 [15] のようなシミュレーテッドアニーリングにおいては、一つのアミノ酸変数 σ_i のトライアル更新ごとに分配関数の計算が必要になる。この設計方程式における並列的な時間発展は、高分子鎖がより長くなり、カノニカル平均や分配関数の計算コストがより高い状況ではより有利になることが予想される。

4 パフォーマンス比較

以下では、設計方程式を3次元立方格子上の一般化されたHPモデル($N=27$)に適用し、そのパフォーマンスをシミュレーテッドアニーリング(SA)と比較した結果について述べる。ここでは考える構造を$3\times3\times3$の立方格子上にぎっしりとつまった「最もコンパクトな自己回避構造(図2)」に限定する。以下で考えるようなエネルギーのパラメータにおいては、これらの構造は常に他のほどけた構造よりもエネルギーが低いのでそれらだけを考えればよい。$N=27$は蛋白質というにはあまりにも短すぎるが、純粋に配列の最適化部分だけを比較するために、「折りたたみ部分(=分配関数もしくはカノニカル平均の計算)」は可能な103346個の厳密数え上げで行い、実際に正解に達しているかを毎時間ステップで「答え合わせ」をする必要もあるので、現実的にはこの長さが限界である。表1はパフォーマンス比較の一例をまとめたものである。1回の実行に対する正解率はSAの方が常に高いが、正解に達するまでのステップ数(状態和などの律速となる計算回数)が10倍程度多く計算時間がかかっているので、結果的にはほとんどのパラメータ、例題に対して、設計方程式は、SAと同等かそれ以上の効率で正解を得ることがわかる。特に、表1の定義での「効率」Pは、比較的難しい例題(#2, #3, #4)に対して数倍、簡単な例題(#1)に対しては10倍以上という結果になった。また、表2の結果では、SAが正解に一度も到達しなかったのに対し、設計方程式は解を見つけた場合もあった(その逆もある)。なお、このパフォーマンスの差は、すでに述べたように、より長くて現実的な蛋白質に対しては、設計方程式に有利な形で現れることが予想される。

一方設計方程式では一個の初期状態からの実行では通常解は一つだけ得られる(図3に各連続化モノマー変数の時間発展を示す)。SAでは擬似温度が0に収束するまでに複数の正解に当たる傾向があり、特に解がたくさんあるような例題#1においては、同じ時間をかけた場合、明らかにSAの方がより多くの種類の解を得ることがわかった。つまり、#1のようなデザイナビリティの高い構造の解は、互いに比較的低いエネルギーバリアで隔てられていて、少数のスピンフリップで遷移することが可能である。言いかえれば、それぞれの解は、突然変異により遷移可能な相同蛋白質の配列になっているといえる。

以上の結果は、連続化相互作用U^1を用いた場合、特にエネルギーパラメータ$(\epsilon_1,\epsilon_2,\epsilon_3)=(-2.3,-1,0)$に対して成績が悪いことを示している。一方、$U^2$を用いた場合は、$(\epsilon_1,\epsilon_2,\epsilon_3)=(-1,0,0)$と$(-2.3,-1,0)$の両方に対して良好である。これらの差が起こる理由として考えられることは、U^2を用いた方が元のエネルギー関数のよりよい拡張になっているということである。(5)式においては、$|m_i|$が小さいときに、第一項m_im_jのオーダーと第二項$m_i|m_i|+m_j|m_j|$のオーダーが等しい。一方(4)式のU^1においては、$|m_i|$が小さいときには、第二項$\frac{\epsilon_k}{2}(m_i+m_j)$の効果がより大きく、本質的な第一項の効果を消してしまうからであると考えられる。ゆえに、計算の初期ステージで相互作用の非線型効果が適切に入るU^2を用いる方がよいのであろう。

Target	U^1			U^2			SA		
	R	S	P	R	S	P	R	S	P
#1	$0.82^{(*)}$	12.2	6.7	$0.56^{(\dagger)}$	16.4	3.4	1.0	162.0	0.62
#2	$0.215^{(*)}$	15.0	1.4	$0.16^{(*)}$	14.8	1.1	0.9	294.3	0.31
	$0.20^{(\dagger)}$	18.5	1.1	$0.25^{(\dagger)}$	14.0	1.8			
#3	$0.30^{(*)}$	16.0	1.9	$0.355^{(\dagger)}$	15.9	2.2	0.8	162.0	0.49
#4	$0.04^{(*)}$	17.6	0.23	$0.055^{(\dagger)}$	14.4	0.38	0.4	256.5	0.16
#5	$0.08^{(*)}$	12.1	0.66	$0.11^{(\dagger)}$	14.7	0.75	0.1	232.2	0.043

表 1: 5つの異なるターゲット構造に対する設計方程式法 (連続化相互作用 U^1 と U^2) とシミュレーテッドアニーリング (SA) の比較。エネルギーパラメータは $(\epsilon_1, \epsilon_2, \epsilon_3) = (-1, 0, 0)$。ターゲット構造#1は、図2に示すもので、文献 [23] の定義でもっともデザイナビリティが高く、最も「簡単な」例題の一つであるといえる。#2 ~ #5 はランダムに選んだもの。R は「正解率」で、設計方程式においては、各ターゲットに対して、異なるランダムな２００個の初期状態から時間発展させたときの正解に達した割合。SA においては１０個の異なる初期状態から正解へ到達した割合。S は正解に到達するまでの時間ステップ数 (平均コンタクトの平均計算回数) で、SA においてはスピンフリップ数 (状態和の計算回数)。$P = 100 \cdot R/S$ は、律速である状態和もしくはカノニカル平均一回当たりの正解率 (を１００倍したもの)。設計方程式においては、初期状態を $m_i \in [-w, w]$ の範囲の一様乱数で発生した。表の (*) は $w = 1.0$ の場合で、(†) は $w = 0.1$ の場合。温度はすべての場合で $1/\beta = 0.01$。時間刻みは $\delta_t = 0.5$、罰金項のパラメータは $\lambda(0) = 0, a = 0.5$。SA における徐冷スケジュールは初期擬似温度 $T_{SA} = 0$ とし、1アミノ酸あたり 1MCS 毎に $T_{SA} = 0.8 \times T_{SA}$ とした。

5 まとめ

従来手法とは異なる、蛋白質設計に対する力学的アプローチについて紹介した。シミュレーテッドアニーリングと比較した結果、よりよい効率で正解を得ることがわかった。設計方程式は、任意の折りたたみ部分 (=コンタクトのカノニカル平均の計算) [1, 2, 11, 12, 15, 17] と組み合わせることが可能である。最近提案された強力な折りたたみ手法である、"Multi-Self-Overlap Ensemble MC" 法 [29, 30] などを内側に実装することにより、より現実的な蛋白質設計への足がかりとすることができるのではないかと考えている。

謝辞

本稿は大阪大学サイバーメディアセンターの菊池誠氏と統計数理研究所の伊庭幸人氏との共同研究に基づくものであるが、文責は筆者にある。また、数値計算の一部は東京大学物性研究所のスーパーコンピュータによるものである。

参考文献

[1] Kurosky, T. and Deutsch, J. M. (1995), J. of Phy. A: Math. Gen., **27**, L387.

[2] Deutsch, J. M. and Kurosky, T. (1996), Phys. Rev. Lett., **76**, 323.

[3] Y. Iba, K. Tokita and M. Kikuchi, (1998), J. Phys. Soc. Jpn., **67**, 3985

Target	U^1			U^2			SA		
	R	S	P	R	S	P	R	S	P
#1	$1.0^{(*)}$	8.5	11.8	$0.75^{(\dagger)}$	10.0	7.5	1.0	105.3	0.95
#2	$0.005^{(*)}$	1.0	0.5	$0.005^{(*)}$	12.0	0.04	0.9	186.3	0.48
	$0.01^{(\dagger)}$	2.5	0.4	$0.1^{(\dagger)}$	10.8	0.93			
#3	$0.002^{(*)}$	5.0	0.04	$0.225^{(\dagger)}$	13.0	1.7	0.8	210.6	0.38
#4	$0.00^{(*)}$	—	—	$0.015^{(\dagger)}$	4.8	0.31	0.0	—	—
#5	$0.00^{(*)}$	—	—	$0.045^{(\dagger)}$	11.7	0.38	0.3	251.3	0.12

表 2: エネルギーパラメータ $(\epsilon_1, \epsilon_2, \epsilon_3) = (-2.3, -1, 0)$ 以外は表 1 と同じパラメータ。

[4] K. Tokita, M. Kikuchi and Y. Iba, (2000), Prog. Theor. Phys. Suppl., **138** (2000) 378

[5] B. Gutte, M. Däuminger and E. Wiischieber, (1979) Nature, **281**, 649

[6] W. F. DeGrado, Z. R. Wasserman and J. D. Lear, (1989) Science, **243**, 622

[7] M. H. Hecht, J. S. Richaredson, D. C. Richardson and R. C. Ogden, (1990) Science, **249**, 884

[8] C. P. Hill, D. H. Anderson, L. Wasson, W. F. DeGrado and D. Eisenberg, (1990), Science, **249**, 543

[9] K. W. Hahn, W. A. Klis and J. M. Stewart, (1990), Science, **248**, 1544

[10] G. S. Shaw, R. S. Hodges and B. D. Sykes, (1990) Science, **249**, 280

[11] Shakhnovich, E. I. and Gutin, A. M. (1993), Proc. Natl. Acad. Sci. USA, **90**, 7195.

[12] E. I. Shakhnovich and A. M. Gutin, (1993), Protein Engineering, **6**, 793

[13] H. Kono and J. Doi, (1994), Proteins, **19**, 244

[14] M. Sasai, (1995), Proc. Natl. Acad. Sci. USA, **92**, 8438

[15] Seno, F. , Vendruscolo, M. , Maritan, A. and Banavar, J. R. (1996), Phys. Rev. Lett. , **77**, 1901.

[16] Koehl, P. and Delarue, M. (1996), Curr. Opinion in Struc. Biol., **6**, 222.

[17] Morrissey, P. and Shakhnovich, E. I. (1996), Folding & Design, **1**, 391.

[18] M. Skorobogatiy, H. Guo and M. J. Zuckermann, (1997), Macromolecules, **30**, 3403

[19] F. Seno, C. Micheletti, A. Martian and J. R. Banavar, (1998), Phys. Rev. Lett., **81**, 2172

[20] A. Irbäck, C. Peterson, F. Potthast and E. Sandelin, (1998), Phys. Rev. E, **58**, R5249

[21] A. Irbäck, C. Peterson, F. Potthast and E. Sandelin, (1999), Structure with Protein & Design, **7**, 347

[22] K. F. Lau and K. A. Dill, (1989), Macromolecules, **22**, 3986

[23] Li, H. , Helling, R. , Tang, C. and Wingreen, N. (1996), Science, **273**, 666.

[24] E. I. Shaknovich and A. M. Gutin, (1990), J. Chem. Phys, **93**, 5967

[25] H. S. Chan and K. A. Dill, (1991), J. Chem. Phys, **95**, 3775

[26] J. J. Hopfield and D. W. Tank, (1985), Biological Cybernetics, **52**, 141

[27] T. Fukai and M. Shiino, (1990), Phys. Rev. A, **42**,7459

[28] T. Fukai and M. Shiino, (1992), J. Phys. A: Math. Gen., **25**, 2873

[29] Y. Iba, G. Chikenji and M. Kikuchi, (1998), J. Phys. Soc. Jpn., **67**, 3327

[30] G. Chikenji, M Kikuchi and Y. Iba, (1999), Phys. Rev. Lett., **83**, 1886

[31] K. Yue and K. A. Dill, (1995), Proc. Natl. Acad. Sci. USA, **92**, 146