Vol.2012-BIO-28 No.20
2012/3/29

Slice Chain Max-Sum アルゴリズムによるタンパク質のポテンシャルエネルギー最小化に関する研究

猪瀬直人[†1] 篠崎隆宏[†2] 杜世橋[†1]
古井貞熙[†1] 関嶋政和[†1]

タンパク質の立体構造予測は，複雑なポテンシャルエネルギー曲面での最小エネルギー探索問題に置き換えることができる．このようなアプローチとして一般的なのが，マルコフ連鎖モンテカルロ (MCMC) 法である．しかし，原子数の増加に応じて探索空間が増加し極小構造が増加してしまうために，分子サイズが大きくなるにつれて MCMC によるポテンシャルエネルギー最小化は困難になっていく．この問題を解決するために，Slice Chain Max-Sum(SCMS) アルゴリズムが提案された．この方法はポテンシャル関数を因子グラフとして表し，MCMC による局所的なサンプル生成と max-sum アルゴリズムによる最適サンプル選択を組み合わせることで，ポテンシャルエネルギーの最小化を行う手法である．これにより，MCMC に比べ効率的に探索が行えることが示された．しかし，従来の SCMS では，MCMC 同様にポテンシャルエネルギー曲面の極小状態を越えての最小状態探索は困難であるという問題があった．本研究では，SCMS でのサンプル生成において，準ニュートン法に基づく最適化を組み合わせた MCMC による探索効率の高い手法を応用し，またポテンシャル関数にこれまで考慮されていなかったファンデルワールス力を追加した SCMS2.0 を開発した．評価実験の結果，SCMS2.0 を用いることで，原子数の増加に伴い最適化込みの MCMC や従来の SCMS と比較して効率的な探索が可能であることが示された．

A Study on the potential energy minimization of proteins by Slice Chain Max-Sum algorithm

Naoto Inose,[†1] Takahiro Shinozaki,[†2] Shiqiao Du,[†1]
Sadaoki Furui[†1] and Masakazu Sekijima[†1]

Three-dimensional protein structure prediction can be modeled as a minimum energy search problem in a potential landscape. One of the most popular approaches with *ab initio* structure prediction is the Markov Chain Monte Carlo (MCMC) method. However, it doesn't perform well for large molecules such as proteins since the search space expands exponentially with the increase in the number of atoms. In order to solve this problem, Slice Chain Max-Sum (SCMS) algorithm has been proposed. This method represents the potential function by a factor graph and combines MCMC sampling and optimal sample sequence selection by the max-sum algorithm. Although it has been shown that SCMS is more efficient than MCMC, SCMS has a difficulty in finding the lowest potential energy state because it requires very long time before escaping from local minima. In this work, we have developed SCMS2.0 by improving the MCMC sampling process using the quasi-Newton method. The supported potential functions have also been extended by introducing van der Waals forces. It has been shown in experiments that SCMS2.0 is more efficient than MCMC and SCMS for large molecules.

1. はじめに

分子の立体構造は，その機能と深く関係しているため，生物学者・物理学者・化学者・計算機科学者やその他のさまざまな分野の研究者に注目されている．特に，タンパク質のような生体高分子の立体構造を知ることは，薬剤設計の上で重要なポケットと呼ばれる薬剤の結合する部位を明らかにしたり，分子の機能を理解することに繋がり，さらには生命の謎を明らかにすることが期待される．

タンパク質の立体構造予測にはさまざまなアプローチがある．既知の情報を用いずに予測を行うアブイニシオ (*ab initio*) 法の中には，マルコフ連鎖モンテカルロ (Markov Chain Monte Carlo method; MCMC)[1] 法を用いる手法がある．MCMC を用いたアプローチとは，確率過程によるサンプリングを行うことでエネルギー最小状態を求める方法である．カノニカルアンサンブル[2] において，ポテンシャルエネルギーの低い状態ほど存在確率が高くなることを利用している．MCMC は，現在の状態から乱数を用いて新しい候補構造の状態を生成し，現在の構造と候補構造のポテンシャルエネルギーの差を用いて，その候補構造を採択するかを判定する．採択された場合は候補構造を次の状態とし，棄却された場合には候補構造は破棄して現在の状態を次の状態として，新たな候補状態を生成する．

†1 東京工業大学
Tokyo Institute of Technology
†2 千葉大学
Chiba University

Vol.2012-BIO-28 No.20
2012/3/29

しかし，この MCMC を用いたアプローチには問題点がある．タンパク質のような自由度の高い系は複雑なポテンシャルエネルギー曲面を持つため，多くの極小状態がある．この極小状態の間には高いエネルギー障壁が存在するため，その極小状態に留まってしまい，探索が困難となってしまう．さらに，探索空間は原子の数に応じて指数関数的に増加するので，分子のサイズが大きくなるに従い，探索空間が大きくなってしまうという問題もある．このため分子が大きくなればなるほど，探索が困難となる．

この困難を解決するために，篠崎らにより Slice Chain Max-Sum (SCMS) アルゴリズム[3] が提案された．この手法はポテンシャル関数を因子グラフ[4] により表現し，MCMC によるサンプル生成と max-sum アルゴリズム[5] による最適なサンプル選択を組み合わせることで，効率的に高分子のポテンシャルエネルギーを最小化する．しかし，この手法には問題点がある．それは，MCMC によるサンプリング時の各原子の移動が単なる乱数によるものであるという点である．純粋な乱数による移動では，多数の原子からなるタンパク質を移動させた場合，原子同士の衝突が起こる可能性が高くなるという問題がある．また，ポテンシャル関数として結合相互作用しか考慮しておらず，多数の原子からなるタンパク質のポテンシャルエネルギーを表現するには不十分である．

本研究では，この SCMS の問題点の改良を行い SCMS2.0 を開発した．サンプリング時に準ニュートン法に基づく最適化込みの MCMC を用いることで効率的なサンプリングを可能にし，また，ポテンシャル関数として非結合相互作用であるファンデルワールス力を追加した．その結果，より実際的な分子モデルを対象としながら，より効率的な探索が可能になった．そして，評価実験により最適化込みの MCMC や従来の SCMS に比べ，原子数が多い状況において SCMS2.0 の有用性が示された．

2. 従来の Slice Chain Max-Sum アルゴリズム

2.1 SCMS の基本的な考え

SCMS は，因子グラフとして分子のポテンシャル関数を表現することに基づいている．しかし，ポテンシャル関数に因子グラフを適用するにあたって，表現されたグラフに閉路があることや原子の座標が連続していることは，扱いにくい特徴である．この問題を解決するために，SCMS では原子を範囲ごとに分けることで，元の因子グラフを**図 1** で示すような閉路のない線形構造のグラフに変換する．この変換された因子グラフを用いて，MCMC で生成したサンプルを各変数ノードの状態とし，これに対して max-sum アルゴリズムを適用することで最小ポテンシャルエネルギー構造が探索される．

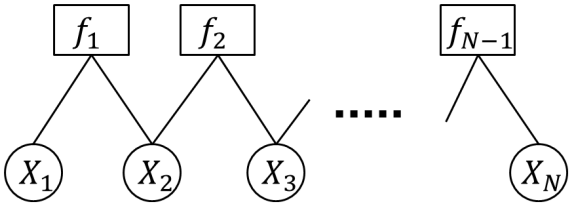

図 1 線形な構造を持つ因子グラフ
Fig. 1 A factor graph having a linear structure.

Step1: 分子を閉路のある因子グラフとして表現する．
Step2: 間隔 W 毎に分子をスライスに分割する．このとき，間隔 W の大きさは最大結合長の 3 倍とする．
Step3: スライス毎に，そのスライスに含まれる原子を集めて因子グラフの複合変数ノード S_m とする
Step4: S_m と S_{m+1} のみに依存する因子を集めて，1 つの複合因子ノード F_m とする．もし，元の因子が S_m のみに依存する場合は F_{m-1} か F_m のどちらかのノードに取り入れる．これにより，線形構造の因子グラフとすることができる．
Step5: S_m 毎に MCMC によるサンプリングを行う．このとき，他のスライスに属する原子の位置は固定する．生成したサンプルを変数ノードの状態とみなす．
Step6: 因子グラフに max-sum アルゴリズムを適用することで最小エネルギー構造を見つける．
Step7: 十分に反復した後に構造を出力，もしくは Step2 へ．

図 2 従来の SCMS のアルゴリズム
Fig. 2 Procedure of the conventional SCMS.

2.2 従来の SCMS のアルゴリズム

図 2 に従来の SCMS のアルゴリズムを記す． Step1 では分子の構造を因子グラフとして表現する．原子の座標を変数ノード，ポテンシャル関数 $V(\mathbf{r})$ の個々の因子を因子ノードとする．例えば，**図 3** に表すような 5 つの原子 a_1 から a_5 で構成された分子については結合長のポテンシャルを $b_1(a_1,a_2), b_2(a_2,a_3), b_3(a_3,a_4), b_4(a_4,a_5)$，結合角のポテンシャルを $c_1(a_1,a_2,a_3), c_2(a_2,a_3,a_4), c_3(a_3,a_4,a_5)$，二面角のポテンシャルを $d_1(a_1,a_2,a_3,a_4), d_1(a_2,a_3,a_4,a_5)$ と表現することができるので，これを因子グラフとして表現すると**図 4** のようになる．従来の手法ではポテンシャル関数 $V(\mathbf{r})$ として結合長，結合角，二面角のみを用いることにしているので，以下の式 (1) のように表される．このとき，この因子グラフは多くの閉路を含んでいる．

$$V(\mathbf{r}) = \sum_{b\in\mathcal{B}} k_b \left(d_b^{eq} - d_b\right)^2 + \sum_{a\in\mathcal{A}} k_a \left(\theta_a^{eq} - \theta_a\right)^2 + \sum_{d\in\mathcal{D}} k_d \left(1 + \cos\left[n\phi_d - \delta_d\right]\right) \qquad (1)$$

Vol.2012-BIO-28 No.20
2012/3/29

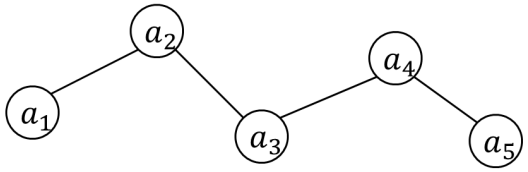

図 3 5つの原子からなる分子
Fig. 3 A molecule consisting of five atoms.

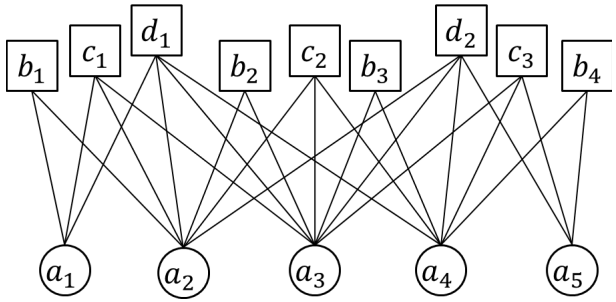

図 4 図 3 の分子のポテンシャルエネルギーの因子グラフによる表現
Fig. 4 A factor graph representing the potential energy of the molecule in Fig.3.

$\mathcal{B}, \mathcal{A}, \mathcal{D}$ はそれぞれ結合長，結合角，二面角に用いられる原子の組である．これらに関する項は，それぞれ結合長 d_b，結合角 θ_a，二面角 ϕ_d の関数であり，定数 k_b,d_b^{eq},k_a,θ_a^{eq},k_d, δ_d はそれらのパラメータである[6)]．

Step2 では分子を 3 次元の直交座標系において一定の間隔 W で平面による分割をしていく．これを表した模式図を**図 5** に示す．この間隔 W は式 (2) に示すように分子の最大結合長 d_{max} の 3 倍よりも大きな値とする．

$$W = 3d_{max} + \epsilon \tag{2}$$

ここで，ϵ は小さな正の値である．また，これにより分けられた区間の一つひとつをスライスと呼ぶことにする．分割する方向は最もスライスが多くなる方向に行うのが理想的であるが，簡単には x,y,z 軸の最長な方向に対して分割する．

分子を複数のスライスに分割することで，因子グラフの変数ノードもそれに応じてスライス毎にグループ分けされる．Step3 では，同じグループの変数ノードを集めることで複合変数ノード S_m とする．ここで，$m = 1, 2, \cdots, M$ はスライス番号，M はスライス数を表す．

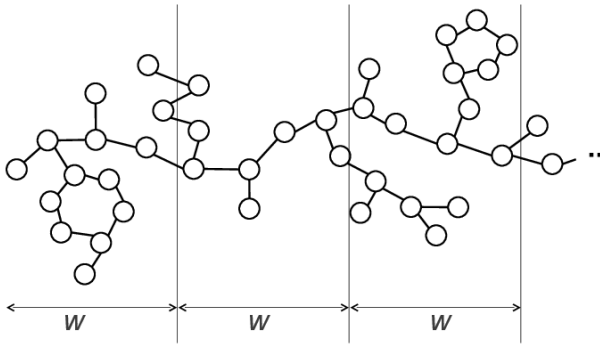

図 5 間隔 W の平面による分子のスライス分割
Fig. 5 Molecule slicing by parallel planes with an interval W.

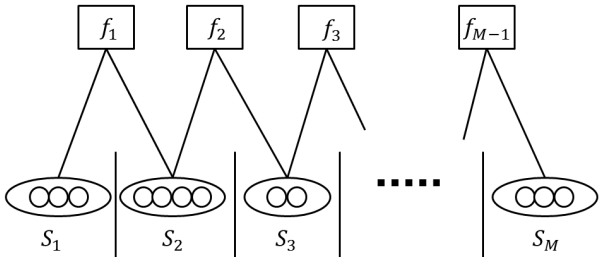

図 6 分子のポテンシャルを表した複合ノードによる線形構造の因子グラフ
Fig. 6 A linear structured factor graph having composite variable and factor nodes representing a potential of a molecule.

Step4 では，S_m と S_{m+1} のみに依存する因子を集めることで，1 つの複合因子ノード F_m とする．もし，因子が S_m にのみ依存する場合は，F_{m-1} か F_m のどちらかのノードに取り入れるが，F_{m-1} と F_m のどちらに取り入れるかは任意である．このとき，スライス幅 W の決め方から元の因子である $V(\mathbf{r})$ のそれぞれの因子は隣接した 2 つのスライスのみに及ぶことが保証される．なぜなら，元の因子は最大でも 4 つの連続した原子によるものであり，スライス幅である $3d_{max}$ を超えることはないためである．したがって，**図 6** のような線形構造の因子グラフが得られる．つまり，原子の位置情報を手掛かりとして閉路のある因子グラフが線形構造の因子グラフに変換される．

Step5 では，複合変数ノード S_m に対して，複合因子ノード F_{m-1} と F_m により表現さ

れるポテンシャル関数を用いて MCMC によるサンプリングを行う．このとき，S_m 以外のノードに対応する原子の座標は固定する．サンプルの生成により，原子座標の集合を得ることができ，これらの集合をそれぞれ複合変数ノード S_m の状態とみなすことができる．また，ここでの MCMC とは Metropolis アルゴリズム[5),7)] のことを指す．

サンプリングにより，各複合変数ノード S_m において複数の状態を得ることができる．Step6 では，得られた状態を用いて因子グラフに max-sum アルゴリズムを適用することで，すべてのスライス間のサンプルの組み合わせの中から最小エネルギー構造を見つける．また，max-sum アルゴリズムでのエネルギー計算は，サンプル間の接続を考慮するため MCMC で計算したものを用いずに再計算する．

max-sum アルゴリズムを適用することで，新しい分子の構造を得ることができる．Step7 では，以前の構造からのエネルギー減少量を調べる．エネルギーの減少量が小さくなったならば，構造が収束したものとして，現在の構造を出力して動作を終了する．そうでない場合は，新しく得た構造を初期状態として Step2 から繰り返す．また，この Step2 から Step7 を 1 エポックと呼ぶ．

3. SCMS2.0 のアルゴリズム

従来の SCMS の問題点として，以下の 2 点が挙げられる．第一の問題は，MCMC によるサンプリング時の各原子の移動が単なる乱数によるものであるという点である．純粋な乱数による移動では，多くの原子からなるタンパク質を移動させる場合，原子同士の衝突が起こり，エネルギー的に不利な状態になる可能性が高い．そのため，従来の SCMS では提案分布の分散を小さくする必要があり，構造が大きく変化するまでに非常に長い時間を要する．

第ニの問題は，ポテンシャル関数として結合相互作用である結合長，結合角，二面角しか考慮されていない点である．タンパク質は多数の原子から構成されているため，そのうちの結合部分のみの計算では複雑なタンパク質のポテンシャルエネルギーを表現するには不十分である．

本章では，本研究で開発した SCMS2.0 における，これら従来の問題点の解決方法を示し，効率な探索を行うための更なる改良について説明する．

3.1 準ニュートン法を組み合わせた MCMC の利用

MCMC における探索効率を向上させる手法として，提案分布からのサンプリングの後，そのサンプルを最近傍の極小状態へ移動させた上で採択判定を行う手法が提案されている[8)]．最適化を行うことで，乱数による移動で生じた原子同士の衝突によるエネルギー的不利を解消することが可能となる．したがって，従来の SCMS における第一の問題の解決のため，このような最適化込みの MCMC によるサンプリングを導入した．これにより，従来の SCMS では行うことができなかった大幅な原子の移動が可能となった．また，本研究では最適化手法として準ニュートン法に基づく L-BFGS 法[9)] を用いることとした．

3.2 ポテンシャル関数の見直し

従来の SCMS における第ニの問題の解決のため，ポテンシャル関数に非結合相互作用であるファンデルワールス力を追加した．これにより，周囲の原子との関係も考慮されるようになり，斥力項により原子同士が近づきすぎるのを抑えることができる．したがって，SCMS2.0 で扱うポテンシャル関数 $V(\mathbf{r})$ は以下のように表せる．

$$V(\mathbf{r}) = \sum_{b\in\mathcal{B}} k_b \left(d_b^{eq} - d_b\right)^2 + \sum_{a\in\mathcal{A}} k_a \left(\theta_a^{eq} - \theta_a\right)^2 + \sum_{d\in\mathcal{D}} k_d \left(1 + \cos\left[n\phi_d - \delta_d\right]\right) + \sum_{\mathcal{F}} \left\{ \left(\frac{A_{ij}}{r_{ij}^{12}}\right) + \left(\frac{B_{ij}}{r_{ij}^{6}}\right) \right\} \tag{3}$$

$\mathcal{F}$ はファンデルワールス力に用いる原子の組であり，この項において r_{ij} は原子ペアの距離，A_{ij},B_{ij} は原子ペアの種類に依存する定数である．また，ファンデルワールス力にカットオフ距離 R を設ける．これは，ファンデルワールス力において距離の離れた原子との相互作用は十分無視出来ると考えられるためである．

ポテンシャル関数の変更に伴い，SCMS のアルゴリズムにも改良が必要になってくる．2.2 の Step2 で決めたスライス幅 W のままでは，このファンデルワールス力の計算において，隣接するスライスを超えたスライス内の原子との相互作用も計算することになり，図 6 で表すような線形構造の因子グラフにより表現できなくなってしまう．したがって，このスライス幅 W を最大結合長の 3 倍である $3d_{max}$ とファンデルワールス力のカットオフ距離 R のどちらよりも大きくすることで対応した．これにより，$V(\mathbf{r})$ のそれぞれの因子の計算に用いる原子が隣接した 2 つのスライスのみに及ぶことが保証され，線形構造の因子グラフで表現できる．

3.3 サンプリング方法の改良

3.1 と 3.2 より，従来の SCMS における 2 つの問題点は解決したことになる．実際に，準ニュートン法による最適化を組み合わせた MCMC では，単純な MCMC と比べて大きな構造変化が得られる．しかし，SCMS においては両端のスライス，つまり変数ノード S_1 と S_M に対応するスライス以外では，ほとんど構造の変化を確認することができなかった．**図 7** はこの SCMS をアラニン 50 残基のポリペプチドに対して実行したときの最初の数エ

Vol.2012-BIO-28 No.20
2012/3/29

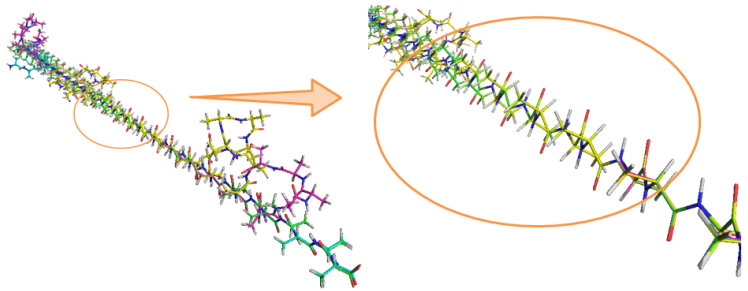

図 7 SCMS でのサンプリングにおいて，隣接スライスの原子座標を固定して最適化付き MCMC を実行した場合の結果例．分子の両端を除いて変化が少ない．

Fig. 7 A construction example after applying MCMC with optimization by fixing positions of atoms in the neighboring slices. Almost no structural change is observed except at both ends of the molecule.

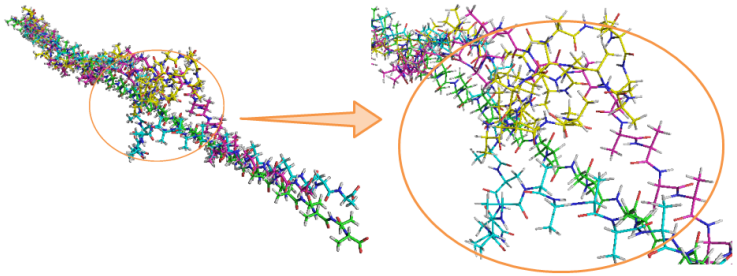

図 8 SCMS でのサンプリングにおいて，隣接スライスの原子座標を含めて最適化付き MCMC を実行した場合の結果例．分子の全体にわたって大きな変化が得られている．

Fig. 8 A construction example after applying MCMC with optimizing by including positions of atoms in neighboring slices. Large structural change is observed through the molecule.

ポックの構造の例である．緑が初期構造であり，青，赤，黄の順にそれぞれ 1 から 3 エポック後の構造を表している．両端は大きく変化しているが，中央部分はほとんど変化していないことがわかる．これは，隣接するスライスを固定していることが原因だと考えられる．現在，Step5 で述べているように複合変数ノード S_m におけるサンプリングを行う場合には他のスライスに属する原子は固定している．最適化込みの MCMC を用いることで大きく原子を動かすことが可能となったが，スライスの端にある原子が大きく移動した場合，隣接するスライスとの間でのポテンシャルエネルギーが大きくなってしまい，最適化の段階で元の位置に引き戻されてしまう．

このサンプリングにおける問題を解決する方法として，隣接するスライスも同時に動かすことが考えられる．つまり，複合変数ノード S_m に対してサンプルを生成する場合，そのノードに隣接する複合変数ノード S_{m-1} と S_{m+1} も含めて MCMC によるサンプリングを行い，得られたサンプルのうち注目しているスライスの状態のみを S_m の状態として保存する．これにより，スライス境界における原子の拘束が小さくなり，大きく移動させることが可能となる．**図 8** にこの改良を行った SCMS の最初の数エポックの構造の例を示す．図 7 と比較してわかるように，中央のスライスでも構造が大きく変化していることが分かる．

3.4 max-sum アルゴリズムへの準ニュートン法を用いた最適化の組み込み

3.3 により，すべてのスライスでの構造の変化が確認された．しかし，このままでは max-sum アルゴリズムによる探索時に隣接するスライス間での接続性が考慮されない．このため，真に最適な組み合わせを選択することが難しくなる．この問題はサンプル連鎖の評価の前に最適化を行うことで解決できるが，全てのサンプル連鎖を列挙してから最適化を行うのでは指数関数的な組み合わせを探索する max-sum アルゴリズムの利点が失われてしまう．そこで，max-sum アルゴリズムを適用する段階で最適化を行なっていくことを考える．すなわち，Step6 での max-sum アルゴリズム実行の際，S_m と S_{m+1} に依存する因子ノード F_m を計算する段階で S_m に属する原子座標の最適化を行うのである．この手法により，max-sum アルゴリズムの指数的な探索能力はそのままで最適化を行うことが可能となる．

この最適化を利用した改良版 max-sum アルゴリズムによる探索について詳しく説明する．まず，m 番目のスライスに対応する因子グラフの変数ノードを S_m，S_m の k 番目の状態を $x_{m,k}$ とし，因子ノード f_m は，S_m と S_{m+1} に依存して決まるものとする．また，$acc_{m,k}$ を S_m の状態 $x_{m,k}$ までの累積スコアとする．**図 9** はこの改良版 max-sum アルゴリズムを表したものである．なお，ここでの最適化は 3.2 で導入した L-BFGS 法を局所的に用いることで実現している．

従来の max-sum アルゴリズムでは，最初のステップとして S_2 の各状態 $x_{2,k}$ に対して $f_1(x_{1,j}, x_{2,k})$ を最小にする $x_{1,j}$ を求める．改良版 max-sum アルゴリズムでは，$f_1(x_{1,j}, x_{2,k})$ を最小にする状態 $x_{1,j}$ の選択の段階で，$x_{i,j}$ に対して最適化を行い状態 $x_{1,j,k}$ を得る．そして，$f_1(x_{1,j,k}, x_{2,k})$ を最小にする $x_{1,j,k}$ を求めるのである．この際，S_2 の異なる状態 $x_{2,k'}$ が S_1 の同じ状態 $x_{1,j}$ を選ぶ場合があるが，最適化による移動により $x_{1,j,k}$ と $x_{1,j,k'}$ は異なるものとなることに注意する．

Vol.2012-BIO-28 No.20
2012/3/29

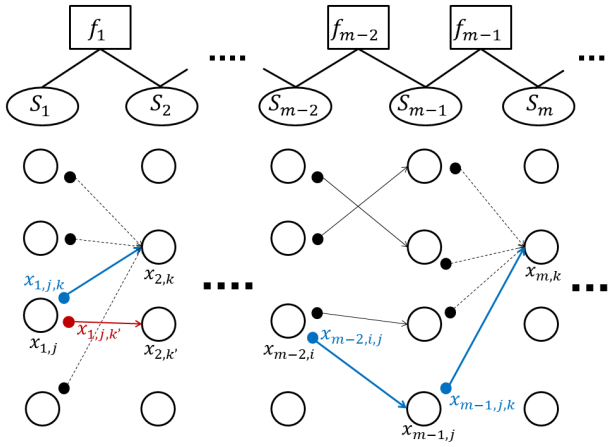

図 9 改良版 max-sum アルゴリズム
Fig. 9 Schematic diagram of the improved max-sum algorithm.

$S_m (m > 2)$ においては，従来の max-sum アルゴリズムでは S_m の各状態 $x_{m,k}$ に対して $acc_{m-1,j} + f_{m-1}(x_{m-1,j}, x_{m,k})$ を最小にする S_{m-1} の状態 $x_{m-1,j}$ を求める．改良版 max-sum アルゴリズムでは，この際に $x_{m-1,j}$ に対して局所的な最適化を行い，状態 $x_{m-1,j,k}$ とする．このとき，S_{m-2} の状態は $x_{m-1,j}$ に対応して選択された $x_{m-2,i}$ に対して最適化を行って得られた状態 $x_{m-2,i,j}$ とすることに注意をする．また，最適化を行うにあたり，累積スコア $acc_{m-1,j}$ についても $x_{m-1,j}$ の変位に伴い更新する必要がある．状態 $x_{m-1,j,k}$ までの累積スコアを $acc_{m-1,j,k}$ とすると，状態 $x_{m-2,i,j}$ までの累積スコア $acc_{m-2,i,j}$ を用いて，$acc_{m-1,j,k} = acc_{m-2,i,j} + f_{m-2}(x_{m-2,i,j}, x_{m-1,j,k})$ により求めることができる．したがって，改良版 max-sum アルゴリズムでは $acc_{m-1,j,k} + f_{m-1}(x_{m-1,j,k}, x_{m,k})$ を最小にする S_{m-1} の状態 $x_{m-1,j,k}$ を求めることになる．これを S_M まで左から右へ順に行うことで，スライス間の境界での接続性の問題を解決しながら最適な組み合わせの探索が可能となる．ここまで述べたことをまとめると，SCMS2.0 のアルゴリズムは**図 10** のように表される．

Step1: 分子を閉路のある因子グラフとして表現する．
Step2': 間隔 W 毎に分子をスライスに分割する．このとき，間隔 W は最大結合長の 3 倍とファンデルワールス力のカットオフ距離 R より大きくする．
Step3: スライス毎に，そのスライスに含まれる原子を集めて因子グラフの複合変数ノード S_m とする
Step4: S_m と S_{m+1} のみに依存する因子を集めて，1 つの複合因子ノード F_m とする．もし，元の因子が S_m のみに依存する場合は F_{m-1} か F_m のどちらかのノードに取り入れる．これにより，線形構造の因子グラフとすることができる．
Step5': S_m 毎に MCMC によるサンプリングを行う．このとき，隣接するスライスに対応する原子も同時に移動させ，それ以外のスライスに属する原子の位置は固定する．生成したサンプルを変数ノード S_m の状態とみなす．
Step6': 隣接したスライス間を接続するために，最適化を行いながら max-sum アルゴリズムを適用し，最小エネルギー構造を見つける．
Step7: 十分に反復した後に構造を出力，もしくは Step2' へ．

図 10 SCMS2.0 のアルゴリズム
Fig. 10 SCMS2.0 algorithm.

4. 実　　験

4.1 実験準備

実装にあたり，Tinker プログラムパッケージ[10] 内のサブルーチンを利用した．また，本研究で用いる実験対象は Tinker 内の protein プログラムを用いて生成された直線状の構造である．なお，本実験は TSUBAME 2.0[11] 上で実行した．

4.2 実験条件

4.2.1 実験方法

本研究の主な狙いは，従来の SCMS や最適化込みの MCMC と比較した場合に SCMS2.0 が効率的にポテンシャルエネルギーを減少させることができるかである．そのため，本研究では従来の SCMS，最適化込みの MCMC，SCMS2.0 を大きさの異なる分子を対象に実行した場合の CPU 時間とポテンシャルエネルギーの関係を比較した．今回は，実験の対象に 100 残基，150 残基および 200 残基のアラニンポリペプチドを用いる．また，3.2 で述べたポテンシャル関数についての改良のみ行った SCMS を従来法として用い，サンプリング手法についても改良を行った SCMS2.0 との比較を行う．MCMC，SCMS2.0 については各対象につき 6 度ずつ実行した．

4.2.2 パラメータについて

パラメータには，SCMS 特有のパラメータや従来の SCMS のみ異なる点がいくつかある．温度 T は MCMC，SCMS2.0 ともに $500K$ とし，従来の SCMS においては極低温 $(T = 1K)$ とした．$500K$ とする理由は，高温にすることでポテンシャルエネルギー曲面での高いエネルギー障壁を越えやすくするためである．しかし，従来の SCMS では高温で

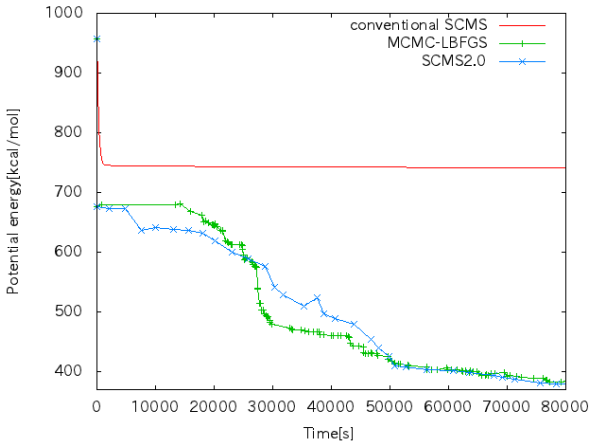

図 11 アラニン 100 残基のエネルギー変化例
Fig. 11 Example of energy change of 100-mer poly-alanine.

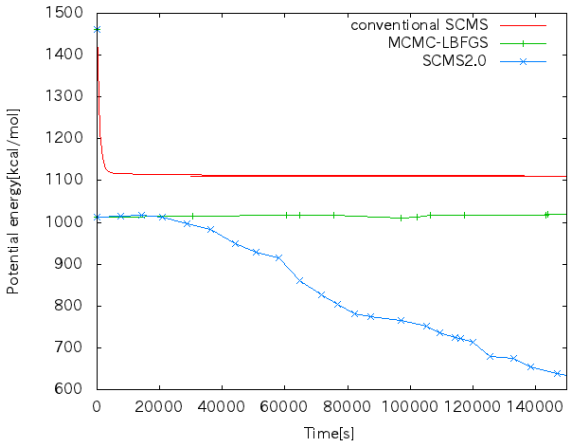

図 12 アラニン 150 残基のエネルギー変化例
Fig. 12 Example of energy change of 150-mer poly-alanine.

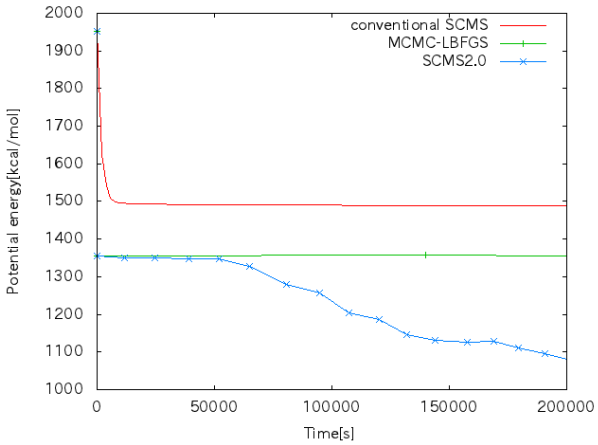

図 13 アラニン 200 残基のエネルギー変化例
Fig. 13 Example of energy change of 200-mer poly-alanine.

実行した場合にエネルギーを下げることができなかったため極低温とした．MCMC によるサンプリング時に原子に与える乱数の提案分布は，予備実験より従来の SCMS では標準偏差 0.0001Å である正規分布に基づくものとし，MCMC，SCMS2.0 では $[-2.0, 2.0]$ の一様分布に基づくものとする．提案分布の違いをみてわかるように，最適化込みの MCMC を用いることで原子を大きく移動させることができるようになっている．スライス間隔 W は，値が小さすぎると構造の変化があまり見られず，大きすぎると SCMS の利点を活かせなくなることから，予備実験より従来の SCMS，SCMS2.0 ともに 50Å とした．サンプル数 K は，従来の SCMS と SCMS2.0 で異なる値を取る．このサンプル数とは，各スライスでの MCMC によるサンプリングにより得られる状態数のことである．従来の SCMS では $K = 50$ とし，さらに 1 つのサンプルを取った後に次のサンプルを取るまでに間隔 $I = 400$ とした．これは，一回の乱数による移動が僅かであるため，連続したサンプル間の相関が大きいためである．つまり，各 S_m でのサンプリングを行うために，$K \times I$ 回の MCMC を行った．一方，SCMS2.0 では $K = 5$ として異なる 5 つの連続したサンプルを各 S_m の状態とする．これは，最適化を加えたことで一回の MCMC において大きく移動させられるので，相関が小さくなると考えられるためである．また，ファンデルワールス力のカットオフ距離 R は 9.0Å とする．

表 1 最適化込みの MCMC と SCMS2.0 での 6 回の試行で最終的に得られたエネルギー
Table 1 Results of six trials of MCMC with optimization and SCMS2.0.

ターゲット	手法	最終的に得られたエネルギー [kcal/mol]					
100 残基	MCMC	387.2	399.6	383.8	403.2	377.7	396.2
	SCMS2.0	414.1	399.8	379.6	379.6	403.1	389.6
150 残基	MCMC	1021.0	1016.3	1016.7	1016.7	633.7	1018.2
	SCMS2.0	625.3	699.6	674.6	674.6	654.9	656.5
200 残基	MCMC	1355.4	1358.8	1355.6	1355.6	1355.6	1354.5
	SCMS2.0	1149.1	1144.0	1125.5	1177.5	1074.4	1124.3

4.3 実験結果

図 11，**図 12**，**図 13** はそれぞれアラニン 100 残基，150 残基，200 残基のポリペプチドに対して従来の SCMS，最適化込みの MCMC，SCMS2.0 を実行したときのエネルギーの変化例を示す．横軸が CPU 時間 [s]，縦軸はポテンシャルエネルギー [kcal/mol] となっている．従来の SCMS では，どの結果を見ても最初にある程度エネルギーが減少した後に，ほとんどエネルギーの変化が見られなくなってしまう．これは，極小状態付近に達した後にその付近を彷徨っていることが原因であると考えられる．また，MCMC と SCMS2.0 では実行後すぐに大幅な減少をしているが，これは初期状態に最適化を行うことで，エネルギー

が大幅に下がっているためである．

次に，SCMS2.0 と MCMC の比較について考える．アラニン 100 残基においては，図 11 に示すように MCMC と SCMS2.0 で大きな差がなかった．アラニン 150 残基，200 残基では，図 12,13 に示すように MCMC の結果は，ほぼ水平でありエネルギーの減少が見られない．これはサンプル候補の生成を繰り返す中で何度かは採択されたが，エネルギーの低い構造は探索できていないことを示している．**表 1** は，最適化込みの MCMC と SCMS2.0 の 6 回の試行で得られた最終的なエネルギーを表したものである．この結果から，150 残基においては MCMC を用いた場合でも SCMS2.0 と同程度のエネルギー減少が一回だけ確認された．また，200 残基での結果では，初期構造への最適化により得られる 1355.6[kcal/mol] からほとんど変化していないことが確認できる．一方，SCMS2.0 を用いた場合には，6 回の試行すべての結果において安定してエネルギーを下げることができている．このように，原子数の増加に伴い MCMC での探索は困難になるが，SCMS2.0 を用いた場合には確実にエネルギーを下げることができる．これにより，大きな分子に対しての SCMS2.0 の有用性が示された．

5. ま と め

本研究では，従来の SCMS の問題点を明らかにし，これらの問題点の改良を行うことで，SCMS2.0 を開発した．最適化込みの MCMC の導入により，従来の SCMS では達成出来なかったポテンシャルエネルギー曲面でのエネルギー障壁を越えることが可能となった．また，MCMC では分子が大きくなるにつれて探索が困難になるのに対し，SCMS2.0 を用いることで効率的にポテンシャルエネルギーを減少させることができ，SCMS2.0 の有用性を示すことができた．

しかし，SCMS2.0 は今後の課題としていくつかの重要な改良の可能性がある．第一に，探索速度の向上である．探索速度の向上方法については，max-sum アルゴリズムによる最適な組み合わせ探索時の枝刈りの導入やスライス毎のサンプリング時の並列化が挙げられる．この 2 点が SCMS2.0 のアルゴリズムにおいて主に実行時間のかかる部分であるので，これらの改良を行うことで SCMS2.0 の 1 エポックあたりの実行時間を大幅に減らすことが期待でき，それに伴いより大きな分子に関しての実行も現実的となる．第ニに，サンプリング方法の改良が挙げられる．現在は max-sum に先行してスライス毎に隣接スライスを含めてサンプリングを行なっているが，この手順を工夫することでより効率的な探索が可能であると考えられる．第三に，溶媒環境下での実験が挙げられる．今回の実験では溶媒環境については考慮していないが，実際の分子は溶媒からの力を受けている．したがって，陰溶媒モデルの導入により溶媒の影響を計算により与えることで，より現実的な実験が可能となる．また，今回 SCMS2.0 と比較した対象は従来の SCMS と MCMC のみであったので，別の手法との比較も行い SCMS2.0 の性能評価を行っていく予定である．

謝辞 本研究は科学研究費補助金 挑戦的萌芽研究（研究課題番号 23650068）の支援を受けたものである．

参 考 文 献

1) Metroplis, N. and Ulam, S.: The monte carlo method, *Journal of the American Statostocal Association*, Vol.66, pp.335–341 (1949).
2) Landau, L.D. and Lifshitz, E.M.: *Statical*, Butterworth-Heinemann, Oxford, 3rd edition part 1 edition (1980).
3) Shinozaki, T., Iwaki, T., Du, S., Sekijima, M. and Furui, S.: Distance-based Factor Graph Linearization and Sampled Max-sum Algorithm for Efficient 3D Potential Decoding of Macromolecules, *IPSJ Transaction on Bioinformatics*, Vol.4, pp.34–44 (2011).
4) Frey, B.J.: *Graphical models for machine learning and digital communicaiton*, MIT Press, Cambridge, MA, USA, (1998).
5) Bishop, C. M.: *Pattern Recognition and Machine Learning(Information Science and Statistics)*, Springer-Verlag New York Inc. (2006).
6) 神谷成敏，肥後順一，福西快文，中村春木：タンパク質計算科学ー基礎と創薬への応用ー，共立出版 (2009).
7) Metroplis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E.: Equations of state calculatins by fast computing machines, *Journal of Chemical Physics*, Vol.21, pp.1087–1091 (1953).
8) Li, Z. and Scheraga, H.A.: Monte Carlo-Minimization Approach to the Multiple-Minima Problem in Protein Folding, *Proc. Natl. Acad. Sci. USA*, Vol.84, pp.6611–6615 (1987).
9) Liu, D.C. and Nocedal, J.: On the Limited Memory Method for Large Scale Optimization, *Mathematical Programming*, Vol.45, pp.503–528 (1989).
10) Ponder, J.W. and Richards, F.M.: An effcient Newton-like method for molecular mechanics energy minimization of large molecules, *Journal of Computational Chemistry*, Vol.8, pp.1016–1024 (1987).
11) Tokyo Institute of Technology: *Global Scientific Information and Computing Center 2011* (2011). http://www.gsic.titech.ac.jp/sites/default/files/gsic2011E.pdf.