科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　　６　日現在

機関番号：　１２６０８
研究種目：　挑戦的萌芽研究
研究期間：　2011～2013
課題番号：　２３６５００６８

研究課題名（和文）グラフィカルモデルを用いた高分子ポテンシャルデコーダの開発

研究課題名（英文）Macromolecular Potential Energy Decoder Based on Graphical Model

研究代表者
篠崎　隆宏（Shinozaki, Takahiro）

東京工業大学・総合理工学研究科（研究院）・准教授

研究者番号：８０４４７９０３

交付決定額（研究期間全体）：（直接経費）　　2,500,000円、（間接経費）　　750,000円

研究成果の概要（和文）：タンパク質の立体構造はその機能と深く関わるため、アミノ酸配列から立体構造を予測することは重要である。本プロジェクトでは効率的な立体構造予測の実現のため、分子のポテンシャルエネルギーにより定まるギブス分布の表現に因子グラフと呼ばれるグラフ構造を導入した上で、マルコフ連鎖モンテカルロ(MCMC) 法による局所的な探索とグローバルなグラフ探索手法であるmax-sumアルゴリズムを組み合わせたSCMS手法の提案と改良を行った。計算機実験により提案法が従来のMCMC 法や、MCMCに準ニュートン法を組み合わせた方法と比較して、少ない計算量でより低いエネルギーの分子形状を探索できることを示した。

研究成果の概要（英文）：Knowing tertiary structure is important to understand and predict protein function. However, it is an open question how to predict the tertiary structure of proteins from a sequence of amino acids. In this project, Slice Chain Max-Sum (SCMS) algorithm has been proposed. This method represents the potential function of a protein molecule as a factor graph, which is a kind of a graphical model. The factor graph is converted into a linearly structured one according to a slicing of the molecule in 3D space. Based on the converted graph, max-sum search is performed in combination with node-wise local MCMC sampling that approximates continuous variables by discrete ones. Experimental results show that SCMS is more efficient than conventional MCMC method. It is also shown that improved version of SCMS (i.e. SCMS2.0) outperforms MCMC method that is reinforced by the quasi-Newton method.

研究分野：　総合領域

科研費の分科・細目：　探索・論理・推論アルゴリズム

キーワード：　タンパク質　立体構造　因子グラフ　Max-Sumアルゴリズム　MCMC　ポテンシャルエネルギー

１．研究開始当初の背景

高分子の３次元立体構造をその構成原子の種類や結合関係の情報のみから推測する手法として、ニュートン力学を用いた分子動力学(MD)法があり、国内外で一般に用いられている。しかしMD法はフェムト(10-15)秒ステップの数値積分により原子運動を追跡するため、多数の原子から構成され熱平衡状態に至るまでに数ミリから数秒と長い時間がかかる高分子への適用には原理的な困難がある。

他方、熱統計力学の理論から、熱平衡状態における分子状態は確率分布として表現できることが知られている。すなわち粒子数(N)・体積(V)・温度(T)一定のカノニカルアンサンブルでは、粒子系の状態を指定する3N次元の運動量ベクトルqと3N次元の位置ベクトルrの分布は、ハミルトニアンがqとrに関して独立した和H(q,r)=K(q)+V(r)となることを利用すると、以下のように書ける。

$$p(q,r)=\frac{1}{Z}\exp\left(-\frac{H(q,r)}{k_bT}\right)=\frac{1}{Z}\exp\left(-\frac{K(q)}{k_bT}\right)\exp\left(-\frac{V(r)}{k_bT}\right)$$

（Zは分配関数、k_bはボルツマン定数）

この定式化のもとで、構造予測は構成原子の位置ベクトルrに関する、確率p(q,r)の最大化あるいはV(r)すなわちポテンシャルエネルギーの最小値探索問題ととらえることができる。最適解探索手法としてはモンテカルロ法を利用するものなどが提案されているが、局所最適解に捕まりやすい困難がある。これは高分子ではrの次元が大きく多数の局所解をもつため、従来の方法では適切なサンプリングが難しいためである。

これに対し、原子間の距離に応じてポテンシャルを近似することでV(r)をグラフィカルモデルとして複数要素に分解可能であり、分解後に自動音声認識分野で用いられている効率的な探索アルゴリズムを応用すればこれまでにない効果的な構造予測が可能となるのではないかと考えたのが、本研究の着想である。

２．研究の目的

本研究では熱統計力学の知識のもとに、高分子の熱平衡状態における３次元形状の予測問題をグラフィカルモデル上の最適解探索問題として定式化するアプローチを提案する。提案アプローチでは動的計画法を応用し直接的に熱平衡状態における最小ポテンシャル探索を行うため、時間方向の数値積分が不要であり、これまでにない高速な３次元立体構造予測が可能になると期待される。具体的な応用分野としては高分子材料の開発やタンパク質の機能解析などが挙げられる。

３．研究の方法

提案する構造予測アルゴリズムはサンプリングやグラフ構造の繰り返し最適化を行うものであるため、評価には実際のデータを用いた計算機実験が不可欠である。このためアルゴリズムとソフトウエアの開発を同時に進める。前述のポテンシャル関数*V(r)*は具体的には化学結合している原子間の結合長や結合角、離れた原子間に働くVan der Waals力等の総体から構成される。本研究では、分子力場についてはMDで用いられているAMBER力場等をそのまま使用する。アルゴリズムの核となるのは、高分子を初期原子配置における距離情報をもとに３次元空間内で区分化し、その区分化をもとに最適化に適した構造を持つ*V(r)*を表すグラフィカルモデルを得ることである。図１に本研究における高分子の区分化と原子レベルでの力場（分子力場）の関係を示す。グラフィカルモデルへの定式化や探索アルゴリズムは幾つかのバリエーションが考えられるが、まずは比較的シンプルな基本構成を一つ決定する。そしてアルゴリズムの拡張性を考慮しつつ、ソフトウエアの設計・実装を行う。その上で、アルゴリズムやソフトウエアの評価及び改良を行う。

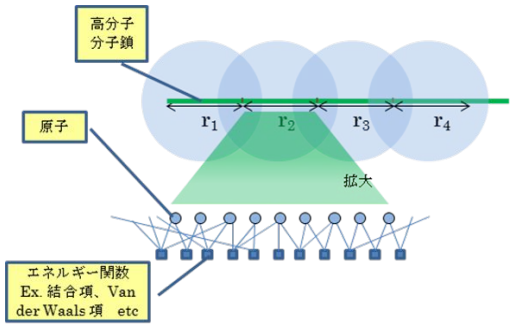

図　１　高分子構成原子の区分化と
原子レベルの依存関係の関係

４．研究成果

(1) Slice Chain Max-sum法(SCMS1.0)の提案

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。このため分子サイズが大きくなると、従来のMCMCでは探索が困難となる。

提案するSCMS法では、分子のポテンシャル関数をまずグラフィカルモデルの一種である因子グラフとして表現する。しかしポテンシャル関数を表す因子グラフは多数の閉路を含み、また原子座標を表す因子は連続変数であることから、そのままでは最適化が難しい。そこで提案法では原子の３次元空間内での初期配置の区分化を手掛かりとして、元の因子グラフを図２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、ノードごとの MCMC で生成したサンプルをそのノードの取り得る離散的な値とし、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を探索する。

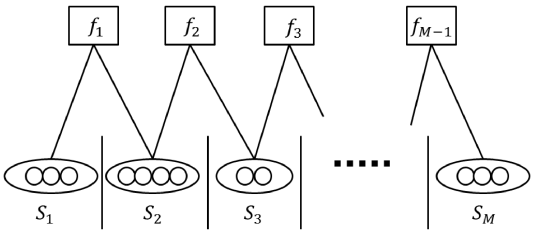

図 2　高分子の３次元空間内での区分化を手掛かりとした因子グラフ構造の線形化

SCMS のアルゴリズムの詳細は以下のとおりである。

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 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が得られる。
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。

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 (r) の個々の要素を因子ノードとする。このとき、この因子グラフは多くの閉路を含んでいる。V (r)としては結合エネルギーのみを仮定する（この点については後で拡張する）。Step2 では分子を３次元直交座標系において一定間隔 W で平面による仮想的に分割する。この間隔 W は分子の最大結合長 dmax の３倍よりも大きな値とする。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。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 (r) のそれぞれの要素は最大でも隣接した２つのスライスのみに依存することが保証される。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 3dmax を超えることはないためである。したがって、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とする。サンプルの生成により原子座標の集合が得られ、それをそれぞれの複合変数ノード S_m の有限個の状態とみなす。 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、新しい分子の構造が得られる。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一定以下であれ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 からの操作を繰り返す。この Step2 から Step7 までの操作を１エポックと数える。

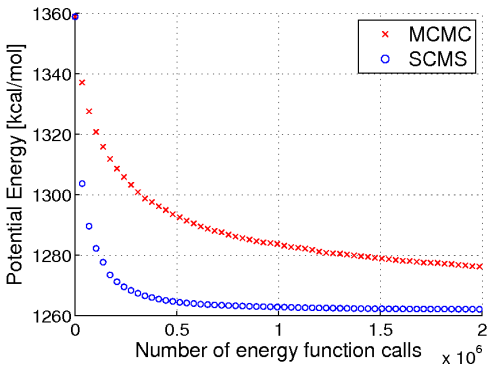

図 3　200 残基からなるポリアラニンの最適化における MCMC と SCMS の比較

図３に、200 残基からなるポリアラニンを対象として MCMC を適用した場合と SCMS を適用した場合のエネルギー変化を示す。ポリアラニンの初期原子配置は、アラニンを直線状に結合したものである。SCMS の方が MCMC よりも少ない計算量でより小さいエネルギーの原子配置を得ていることが分かる。

(2) 改良型 Slice Chain Max-sum 法(SCMS2.0)の提案

初期型の SCMS により MCMC と比較してより少ない計算量でより小さいエネルギーの原子配列が得られることを示した。しかし問題点として、以下の２点が挙げられる。第一に、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基づいているため、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確率が高い。そのため、サンプリング時に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。第ニに、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、二面角しか考慮していない。タンパク質は多数の原子から構成されているため、結合部分のみの計算ではポテンシャルエネルギーを表現するには不十分である。そこで、以下の４項目について改良を行った。

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。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。SCMS においてもこのような最適化込みの MCMC を利用することで探索効率の改善が期待されることから、ノードごとのサンプリングで使用している MCMC のステップを最適化込のものに置き換えた。

② ポテンシャル関数の見直し

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る。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小さくなることから、カットオフ距離 R を設ける。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。Step2 で決めたスライス幅 W のままでは、ファンデルワールス力によりスライスを超えた原子間の相互作用が発生し、線形構造の因子グラフで表現できなくなってしまう。そこでスライス幅 W を最大結合長の 3 倍である 3dmax 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い値に選ぶようにアルゴリズムを変更する。これにより再度 V(r) のそれぞれの要素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 サンプリング方法の改良

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。しかし予備実験を行ったところ、SCMS においては両端のスライス以外では、ほとんど構造の変化を確認することができなかった。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。すなわち Step5 において複合変数ノード 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} も含めて最適化付 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。

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

隣接するスライスを同時にサンプリングすることで、すべてのスライスにおいて構造がより大きく変化する改善が得られた。しかし今度は 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とによる問題が見られた。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適用する。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。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としながら最適化を行うことが可能となる。

図４に、最適化付 MCMC、初期バージョンの SCMS(SCMS1.0)、および改良型 SCMS(SCMS2.0)を 200 残基からなるポリアラニンのエネルギー最小化に適用した場合の結果を示す。SCMS2.0 が最適化付 MCMC と SCMS1.0 のどちらよりも少ない計算量でより小さいエネルギーの原子配置を得ていることが分かる。

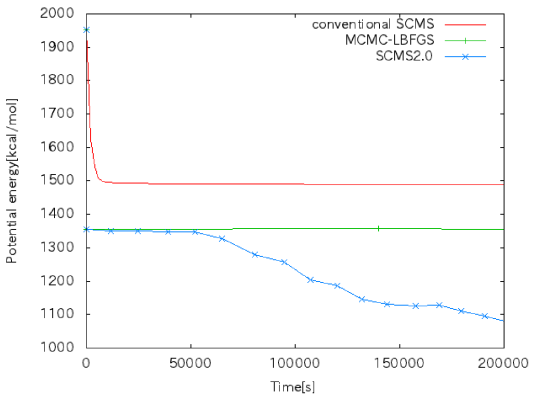

図 4　200 残基からなるポリアラニンの最適化における最適化付 MCMC、SCMS1.0 および SCMS2.0 の比較

また図５に、最適化付 MCMC と SCMS2.0 をヒトオキシヘモグロビンに適用した場合の結果を示す。初期状態として X 線結晶解析により決定された分解能 2.1Åの原子座標を用いた。こちらの条件においても、SCMS2.0 の方が最適化付 MCMC よりも小さいエネルギーを与える原子配置を得ていることが分かる。

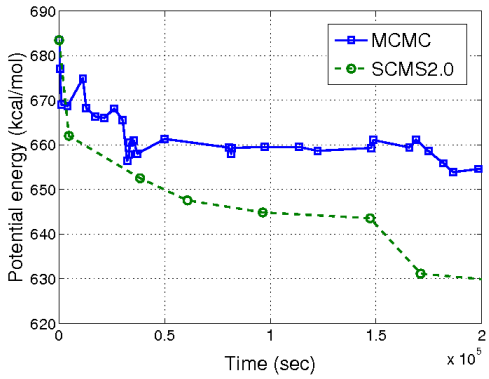

図 5　ヒトオキシヘモグロビンの最適化における最適化付 MCMC と SCMS2.0 の比較

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1　件）

① Takahiro Shinozaki, Toshinao Iwaki, Shiqiao Du, Masakazu Sekijima and Sadaoki Furui, "Distance-based Factor Graph Linearization and Sampled Max-sum Algorithm for Efficient 3D Potential Decoding of Macromolecules," IPSJ Transaction on Bioinformatics, 査読あり, Vol.4, pp.34-44, 2011

〔学会発表〕（計　5　件）

① 篠崎 隆宏, 関嶋 政和, "SCMS2.0 によるタンパク質ポテンシャルエネルギー最小化の諸条件における評価, " 情報処理学会バイオ情報学研究会(SIG BIO)第 37 回研究会, 2014.3.5, 九州工業大学
② Takahiro Shinozaki, Naoto Inose, Shiqiao Du, Sadaoki Furuiy and Masakazu Sekijima, "Macromolecular Potential Energy Minimization Based on Slice-Wise Sampling and Max-Sum Algorithm," The 7th IAPR International Conference on Pattern Recognition in Bioinformatics, 2012.11.10, Tokyo Institute of Technology
③ Naoto Inose, Takahiro Shinozaki, Shiqiao Du, Sadaoki Furui and Masakazu Sekijima "Protein Potential Energy Minimization Using Slice Chain Max-Sum Algorithm," The 26th Annual Symposium of The Protein Society, 2012.8.6, San Diego
④ 猪瀬 直人, 篠崎 隆宏, 杜 世橋, 古井 貞熙, 関嶋 政和, "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, " 情報処理学会バイオ情報学研究会(SIG BIO)第 28 回研究会, 2012.3.29, 東北大学
⑤ 岩木 聡直, 篠崎 隆宏, 古井 貞熙, "Sampled Max-Sum Algorithm For 3D Structure Prediction of Proteins, " 第 11 回日本蛋白質科学会年会 2011.6.7, ホテル阪急エキスポパーク

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://www.ts.ip.titech.ac.jp

６．研究組織
(1)研究代表者
篠崎　隆宏　（SHINOZAKI, Takahiro）
東京工業大学・大学院総合理工学研究科・准教授
研究者番号：８０４４７９０３

(2)研究分担者
篠田　浩一　（SHINODA, Koichi）
東京工業大学・大学院情報理工学研究科・教授
研究者番号：１０３４３０９７

(3)連携研究者
関嶋　政和　（SEKIJIMA, Masakazu）
東京工業大学・学術国際情報センター・准教授
研究者番号：８０３７１０５３