様式 C-19

科学研究費補助金研究成果報告書

平成　２１年　５月１５日現在

研究種目：基盤研究（C）
研究期間： 2006～2008
課題番号：１８５００１７２
研究課題名（和文）　確率的空間埋込法とガウス型大域的最適化の統合による
タンパク質立体構造推定
研究課題名（英文）　Determination of protein structure by a unified method of stochastic embedding and Gaussian global optimization

研究代表者
土居　伸二（DOI SHINJI）
大阪大学・大学院工学研究科・准教授
研究者番号：50217600

研究成果の概要：
タンパク質の立体構造を，構成する原子間距離から効率的に推定する方法を考案した．この方法は，最適化問題（誤差を最小にする問題）を極めて単純な確率的手法で解くものであり，大規模なタンパク質の立体構造推定にも応用可能な効率の良いアルゴリズムである．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	900,000	0	900,000
２００７年度	900,000	270,000	1,170,000
２００８年度	700,000	210,000	910,000
年度			
年度			
総　計	2,500,000	480,000	2,980,000

研究分野：非線形システム工学
科研費の分科・細目：情報学・感性情報学・ソフトコンピューティング
キーワード：距離幾何問題，大域的最適化，確率アルゴリズム，タンパク質立体構造

１．研究開始当初の背景

われわれの生命活動を担うタンパク質の機能を明らかにすることは，生物学や医学的見地からだけでなく，創薬などの産業応用上も極めて重要な問題である．タンパク質の機能を解明するためには，まずその立体構造を決定することが重要である．

蛋白質の構造を決定する方法には，X 線結晶解析法と NMR 解析法がある．どちらの方法もメリット・デメリットがあり，相補的な方法である． NMR 解析法では，部分的な原子間の，しかも正確な値ではなく距離制限（距離の上限と下限）データしか得られない．このデータから原子座標を決めるには，距離制限を満足する距離をランダムに選び，力ずくで試行錯誤を繰り返す方法か，ある種の最適化問題を解く方法のどちらかが用いられる．後者の方法も，多数の局所解の存在のため，多くの試行錯誤を必要とする．局所解を避け大域的最適化を行う方法としてガウス変換が知られているが，従来用いられてきた目的関数では，そのガウス変換が近似的にしか得られず，しかも複雑な計算を要する．本研究で用いるガウス型の目的関数に対してはガウス変換が解析的に得られるので，従来の方法に比べて圧倒的に計算効率が良く，巨大な蛋

白質にも適用可能である．この目的関数は，あまりにも単純であるが，これまで誰も用いたことがなく，類似研究はない．

確率的空間埋込法は，もともとはデータマイニングのために提案された手法であり，一般 Distance Geometry 問題に適用された例はなく，蛋白質立体構造推定問題に適用可能なように拡張・改良を行い，シミュレーション実験を行ったところ，上記の最適化手法に比べて 2 桁から 3 桁も計算速度が速いという驚くべき結果が得られた．ただ，この結果は好条件下（得られる距離データ量が多い）での結果であるので，悪条件下では，様々な問題が生じる可能性がある．そこで，ガウス型大域的最適化による方法と確率的空間埋込法を統合することにより，効率的な立体構造推定ができるのではないかと，本研究提案に至った．

２．研究の目的

本研究では，「一般」Distance Geometry 問題を考察する．Distance Geometry 問題とは，N 個の対象（原子）の間に距離が与えられたときに，その距離関係を満たすように原子を M 次元空間に埋め込む（N 個の原子座標を決定する）問題である．この問題は古くから知られており，全ての原子間の距離が正確に与えられている場合は，単純に行列の固有値問題として解が得られる．しかし，距離データが部分的な原子対にしか与えられなかったり，また正確な距離データでなく，ある幅を持ったデータ（上下限データ）の場合には単純な解法はなく，最適化手法も含めて様々な方法が研究されている．

本研究で提案する方法は，ガウス型目的関数を用いた大域的最適化手法と確率的なアルゴリズムである確率的空間埋込法を統合したハイブリッド手法である．それぞれの方法については，既に予備研究を行っており，前者は探索空間を狭めるという大域的最適化の点において，後者は，圧倒的な計算速度の点において優れていることが予想される．両方法とも我々の提案による独創的な方法であるが，まだ研究の端緒段階であるので，本研究では，これまでの研究を発展させ，両手法の優れた面を統合した一般 Distance Geometry 問題の解法を提案する．本研究では，具体的対象として蛋白質の立体構造推定問題へ適用する．

３．研究の方法

【本研究で考察する問題：一般 Distance Geometry 問題】

本研究では，距離制限（原子間距離の上限と下限）データから，分子（主に蛋白質）の立体構造を決める．つまり，以下の問題を考える：

> N: 原子の個数,
> x_i： i 番目の原子の座標
> $u_{i,j}$： i 番目と j 番目の原子間距離の上限
> $l_{i,j}$： i 番目と j 番目の原子間距離の下限
> **問題：**
> $u_{i,j}$ と $l_{i,j}$ が（実験から）与えられたとき，
> $l_{i,j} \leqq \|x_i - x_j\| \leqq u_{i,j}$ を満足するように，
> 原子座標 x_i（$i=1,...,N$）を求める．

この問題に対して，以下の 2 種類の異なる方法を用いる．

【本研究で用いる方法１：ガウス型大域的最適化法】

以下の目的関数 $F(x_1, x_2,..., x_N)$ を最大にする x_i（$i=1,...,N$）を求める：

$$F(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N) = \sum_{i<j} \omega_{i,j} f_{i,j}\left(||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right)$$

$$f_{i,j}\left(||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right) = \frac{1}{\sqrt{2\pi}\delta_{i,j}} \exp\left(-\frac{\left(||\boldsymbol{x}_i - \boldsymbol{x}_j||^2 - m_{i,j}^2\right)^2}{2\delta_{i,j}^2}\right)$$

ただし，$\omega_{i,j}$ は適当な重み係数であり，通常 $\omega_{i,j}=1$ とする．この目的関数は，極めて単純なガウス関数（の重ね合わせ）で，原子間距離の上限と下限の平均（$m_{i,j}=(u_{i,j} + l_{i,j})/2$）をガウス関数のピークとし，上限下限の差を標準偏差（$\delta_{i,j} = (u_{i,j} - l_{i,j})/2$）とするものである．

一般に，上のような目的関数は，多くの局所的な極大値を持ち，必ずしも大域的な最大値を求めることができない．そこで，本研究では，大域的最適化を達成するためにガウス変換を用いる．ガウス変換は，目的関数 F とガウス関数との畳み込み積分である．ガウス関数の分散を大きく取ることで，ガウス変換により目的関数が（極大・極小の少ない）滑らかな関数に変換される．

【本研究で用いる手法２：確率的空間埋込法】

上と同じ問題に対し，次のアルゴリズムを適用する：N 個の原子の中から，ランダムに 2 個の原子を選び，その番号をそれぞれ k, l とする．この二つの原子の座標を

$$\boldsymbol{x}_k \longleftarrow \boldsymbol{x}_k + \lambda\frac{1}{2}\frac{m_{k,l} - ||\boldsymbol{x}_k - \boldsymbol{x}_l||}{||\boldsymbol{x}_k - \boldsymbol{x}_l||}(\boldsymbol{x}_k - \boldsymbol{x}_l)$$

$$\boldsymbol{x}_l \longleftarrow \boldsymbol{x}_l + \lambda\frac{1}{2}\frac{m_{k,l} - ||\boldsymbol{x}_k - \boldsymbol{x}_l||}{||\boldsymbol{x}_k - \boldsymbol{x}_l||}(\boldsymbol{x}_l - \boldsymbol{x}_k)$$

と更新し，この過程を繰り返す．ただし λ は座標更新過程を調節するパラメータであり，$\lambda=1$ のときは，x_k と x_l の座標を，その原子間距離が $m_{k,l}$ に一致するように更新する．このアルゴリズムの確率的性質は，２つの原子

をランダムに選ぶという点だけである．

以上の方法を用いて，具体的には，以下の課題を遂行する．

(1) ガウス型大域的最適化による方法，確率的空間埋込法のそれぞれについて，悪条件下（原子間距離 データが全ての原子対に対して数パーセントしか与えられないような場合）で立体構造推定実験を行い，収束速度，構造推定成功率，構造推定が失敗する場合の特徴等を洗い出す．

(2) Protein Data Bank (PDB)に収められている蛋白質の中で，さらに大きなものに対して実験を行う．また，蛋白質の特徴に応じた立体構造推定プロセスの特徴を明らかにする．

(3) 蛋白質の部分構造（αヘリックスやβシート等の２次構造）が既知の場合に，それらの事前知識を立体構造推定に活用できるように，アルゴリズムを工夫する．

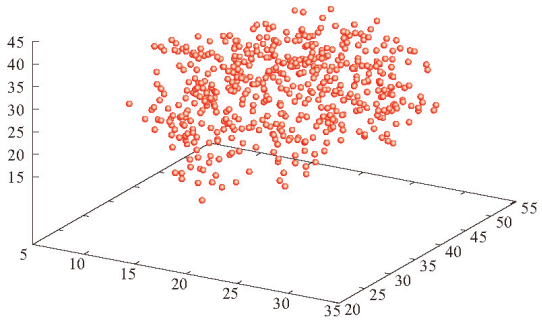
(4) 以上の研究を，ガウス型大域的最適化法と確率的空間埋込法の両方に対して行い，それぞれの方法の特徴・優れている点を明らかにする．

４．研究成果

以上の研究に対し，以下の成果を得た：

(1) 確率的近接データ埋込法（SPE: Stochastic Proximity Embedding）を一般 Distance Geometry 問題に適用可能なように拡張を行った．特に，上下限値の取り扱いや確率的近接データ埋込法における近傍半径値の取り扱い方法を工夫した．

(2) 実際に，既知のタンパク質であるインスリンデータ（原子の分布を以下の図に示した）を用いて計算機実験を行い，提案手法が正しく動くことを確認した．



(3) より正確な分子構造決定を行うために，大幅な拡張方法を提案した：２点の座標ではなく，３点の座標を同時に更新する学習アルゴリズムを提案した．このアルゴリズムの部分修正版を３種類提案し（三角形の長辺と重心を保存する方法，放射状に３点を動かす方法，重心を保存かつ２乗平均誤差を最小にする方法），同様の計算機実験により詳細な性能比較を行うことで提案手法の有効性を確認した．

(4) 上下限データが確率的に分布している場合の検討を正規分布と一様分布に対して行った．このような条件下では，分子構造決定は全体としては困難になるが，提案手法はこの悪条件下でも正しく機能することを確認した．

(5) ２次関数型とガウス型目的関数を用いた大域的最適化方法（ガウス変換を用いる）と確率的近接データ埋込法に基づく我々の提案手法の性能比較を行い，分子構造決定の正確さ及び計算効率の両方において，提案手法が優れていることを確認した．

(6) 距離データに即して，高次元データを低次元空間にうまく埋め込む手法の開発を行った．特に，与えられたデータ間の補間データを自動生成する方法を考案した．

(7) 本研究の距離幾何問題は，データマイニング問題と密接な関連を持つが，データマイニングで一番問題になる「与えられた高次元データ間の距離（類似度）をどのように定義するか」という問題の検討を行った．与えられた高次元データ間の距離構造を自動学習するアルゴリズムを提案し，高次元データを用いて数値実験を行った．その結果，与えられたデータに即した距離構造をうまく学習することができ，より適した低次元埋込みが実現できることが分かった．

(8) 確率的近接データ埋込法 SPE は，与えられたデータ対をランダムに選択するところに確率性が挿入されているアルゴリズムであるが，埋込みを行う（低次元空間での座標を更新する）アルゴリズムにも確率性を導入し，一つの座標値ではなく座標の分布を推定する方法を考案した．インスリンやグルカゴンデータを用いて，埋込み（分子立体構造推定）に関する数値実験を行い，その性能に関して従来法との詳細な比較・検討を行った．

(9) 埋め込み時の学習パラメータを「自動学習する」アルゴリズムを考案した．これにより，データ毎にいちいち学習パラメータを手動で調整する必要がなくなり，提案手法の適用範囲が格段に向上した．

(10) 本研究の距離幾何問題は，データマイニング問題と密接な関連を持つが，データマイニングで一番問題になる「与えられた高次元データ間の距離（類似度）をどのように定義するか」という問題の更なる検討を行った．与えられた高次元データに対して，ユークリッド距離の重み係数を自動学習することで，与えられたデータに即した距離構造をうまく学習する

ことができ，より適した低次元埋込みが実現できることが分かった．

以上のように，研究目的は十分に達成された（Gauss 型目的関数を用いた最適化の性能が悪いことが判明し，主に手法２を用いたことは多少予想外ではあったが，結果的に極めて効率の良い埋め込みアルゴリズムの開発に成功した）．

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７件）

(5) N.Nishikawa, S.Doi, Optimization of Distances for a Stochastic Embedding and Clustering of High-Dimensional Data, Proc. The 23rd International Technical Conference on Circuits/Systems, Computers and Communications, pp. 1125-1128 (2008) 査読有.

(6) 西川尚斗，土居伸二，熊谷貞俊，確率的埋め込み法によるデータクラスタリング，電子情報通信学会技術研究報告 NLP, pp.39-44 (2008) 査読無.

(7) N.Nishikawa, S.Doi, S.Kumagai, Embedding of high-dimensional data in low-dimensional space by a simple stochastic method, Proc. 15th IEEE International Workshop on Nonlinear Dynamics of Electronic Systems, pp.325-328 (2007) 査読有.

(8) 西川尚斗，土居伸二，熊谷貞俊，確率的手法による高次元データの低次元埋め込み，電子情報通信学会技術研究報告 NLP, pp.35-40 (2007) 査読無.

(9) 西川尚斗，加嶋浩之，土居伸二，熊谷貞俊，確率的手法を用いた高次元データの低次元埋め込み地図の作成，電子情報通信学会総合大会講演論文集，p.64 (2007) 査読有.

(10) 加嶋浩之，土居伸二，熊谷貞俊，距離幾何学問題における確率的近接データ埋め込み法の拡張，電子情報通信学会技術研究報告 NLP, pp.39-44 (2006) 査読無.

(11) H.Kashima, S.Doi, S.Kumagai, Application of Stochastic Proximity Embedding to Distance Geometry Problems, Proc. SICE-ICASE International Joint Conference 2006, pp.4451-4456 (2006) 査読有.

〔学会発表〕（計７件）

(1) N.Nishikawa, Optimization of Distances for a Stochastic Embedding and Clustering of High-Dimensional Data, The 23rd International Technical Conference on Circuits/Systems, Computers and Communications, Yamaguchi, Japan, July 7, 2008.

(2) 西川尚斗，確率的埋め込み法によるデータクラスタリング，電子情報通信学会非線形問題研究会，北海道大学学術交流会館，2008 年 2 月 1 日．

(3) N.Nishikawa, Embedding of high-dimensional data in low-dimensional space by a simple stochastic method, The 15th IEEE International Workshop on Nonlinear Dynamics of Electronic Systems, Tokushima Japan, July 26, 2007.

(4) 西川尚斗，確率的手法による高次元データの低次元埋め込み，電子情報通信学会非線形問題研究会，広島工業大学広島校舎, 2007 年 6 月 8 日．

(5) 西川尚斗，確率的手法を用いた高次元データの低次元埋め込み地図の作成，電子情報通信学会総合大会，名城大学天白キャンパス，2007 年 3 月 21 日．

(6) 加嶋浩之，距離幾何学問題における確率的近接データ埋め込み法の拡張，電子情報通信学会非線形問題研究会，北海道湯の川温泉 KKR はこだて，2006 年 12 月 13 日．

(7) H.Kashima, Application of Stochastic Proximity Embedding to Distance Geometry Problems, SICE-ICASE International Joint Conference, Bexco, Busan, Korea, Oct. 19, 2006.

６．研究組織

(1)研究代表者

土居　伸二（DOI SHINJI）
大阪大学・大学院工学研究科・准教授
研究者番号：50217600

(2)研究分担者

なし．

(3)連携研究者

なし．