MEOを用いた蛋白質立体構造アラインメントとその性能評価

Protein Structure Alignment using MEO and Its Performace Evaluation

中田 章宏
Akihiro Nakada
広島市立大学大学院
情報科学研究科
E-mail: mw67025@edu.ipc.hiroshima-cu.ac.jp

田村 慶一
Keiichi Tamura
広島市立大学大学院
情報科学研究科
E-mail: ktamura@hiroshima-cu.ac.jp

北上 始
Hajime Kitakami
広島市立大学大学院
情報科学研究科
E-mail: kitakami@hiroshima-cu.ac.jp

***Abstract*—Proteins are important biochemical compounds that have biogenic functions for biological activities. The three-dimensional structures of proteins are closely related to its biological functions, and therefore, techniques for comparing them have been studied. Many of these techniques for comparing protein structures are based on protein structure alignment, which is one of the most effective methods. CMO (Contact Map Overlap) is formulated as combinatorial optimization to find the optimal structure alignments. We have proposed a novel heuristic using Modified Extremal Optimization (MEO) for CMO. Our MEO-based heuristic is characterized by three features. First, the proposed heuristic uses MEO for alternation generations. Second, an initial solution is created by dynamic programming (DP). Third, state transition is executed using the best admissible move strategy. In this paper, we discuss the performace evaluation of the EO-based heuristic.**

I. はじめに

蛋白質は酵素，抗体やホルモンなど，我々の生命活動を支える生体機能を持つ重要な物質のひとつである．蛋白質は，立体構造がその蛋白質が持つ生体機能を決定すると言われているがその関係性は十分に解明されていない．ただし，アミノ酸配列が似ていなくとも立体構造が類似する蛋白質同士はその生体機能がお互いに類似していると言われており，蛋白質の立体構造を比較する研究 [1] が盛んに行われている．

蛋白質の立体構造を比較するときに必要とされている機能が類似構造の抽出であり，類似構造を抽出するために広く利用されているのが蛋白質立体構造アラインメント [2] である．蛋白質立体構造アラインメントは蛋白質を構成するアミノ酸の数が増えるとともにその組合せが膨大となり，最適なアラインメントを求めることは，バイオインフォマティクスにおいて最も難しい問題のひとつとして知られている [3].

蛋白質立体構造アラインメントを組合せ最適化問題として定式化したのが CMO（Contact Map Overlap）問題 [4] である．CMO 問題では，蛋白質の残基を頂点とし，近接する残基同士を辺で結んだコンタクトマップと呼ばれるグラフを作成する．CMO 問題は，コンタクトマップ間のアラインメントにより保存される共通コンタクトと呼ばれるオーバラップ構造の数を最大化する問題として定義される．

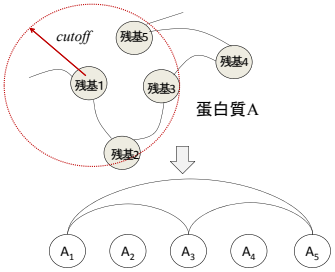

図 1.　コンタクトマップの例

CMO 問題は，NP 困難な問題のひとつであることが知られており，分枝限定法や線形計画法にラグランジュ緩和を組み合わせた手法がその解法として提案されている．また，厳密解を求めるには非常に多くの計算量が必要なため，実用的な観点から，進化計算やEO（Extremal Optimization）[5] などの発見的解法を用いた手法の研究 [6] も行われている．

我々は，MEO（Modifined EO）を用いた CMO 問題の発見的解法 [7] を提案している．提案手法は，(1）世代交代に MEO を用いる，(2）初期個体は動的計画法を用いて作成する，(3）MEO における状態遷移に即時移動戦略ではなく，最良移動戦略を用いる，手法となっており，その有効性が示されている．しかしながら，CMO 問題で求めた最適構造アラインメントと結果と実際の蛋白質立体構造間の空間的な類似性の評価が課題として残っている．本論文では，MEO を用いた CMO 問題の発見的解法を示すとともに，CMO 問題で求めた最適構造アラインメントに対する RMSD(root-mean-square deviation) による性能評価結果を報告する．

本論文の構成は次の通りである．第 2 章では，CMO 問題について紹介する．第 3 章では，MEO による発見的解法を用いた CMO 問題の解法について述べる．第 4 章で性能評価の実験結果を報告し，第 5 章で本論文のまとめを行う．

II. CMO 問題

コンタクトマップでは，蛋白質 v の i 番目の残基と，j 番目の残基は，それぞれ，i 番目の頂点 v_i，j 番目の頂点 v_j として表現される．頂点 v_i と頂点 v_j とが辺で

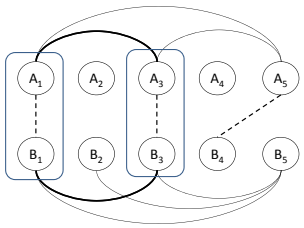

図 2. アラインメントとオーバラップの例

結ばれている場合，残基 i と残基 j 間の距離が，与えられたカットオフ距離 $cutoff$ 未満であることを示す．図 1 にコンタクトマップの例を示す．

蛋白質 v のコンタクトマップ CM_v を，$CM_v = (RV_v, CE_k)$ と表現する．ただし，RV_v= $\{v_1, v_2, \cdots, v_n\}$ は頂点集合であり，

$$CE_v = \{(v_i, v_j) | \quad v_i \in RV_v, v_j \in RV_v, \\ i < j, \quad dist(v_i, v_j) < cutoff\} \qquad (1)$$

はコンタクトエッジの集合を表す．ただし，関数 $dist$ は残基 i と残基 j の中心座標間の距離を返す関数である．例えば，図 1 の蛋白質 A のコンタクトマップは，$CM_{\mathrm{A}} = (RV_{\mathrm{A}}, CE_{\mathrm{A}}$ と表し，このとき，$RV_{\mathrm{A}} = \{A_1, A_2, A_3, A_4, A_5\}$，$CE_{\mathrm{A}} = \{(A_1, A_3), (A_1, A_5), (A_3, A_5)\}$ である．

蛋白質 v と蛋白質 w とをそれぞれ表現するコンタクトマップ CM_v と CM_w の部分頂点集合 ($RV_v^+ \subseteq RV_v$,$RV_w^+ \subseteq RV_w$) 間を一対一に対応付けることをアラインメントという．また，アラインメントされた頂点のペアをアラインメントペアと呼ぶ．ここで，このアラインメントを全単射として，

$$\phi : {RV_v}^+ \to {RV_w}^+, v_i \mapsto w_{\phi(i)}, \qquad (2)$$

と定義すると，アラインメントペア集合 AL^ϕ は，

$$AL^\phi = \{(v_i, w_{\phi(i)}) \mid v_i \in {RV_v}^+, w_{\phi(i)} \in {RV_w}^+\}, \quad (3)$$

と表現することができる．

図 2 はふたつの蛋白質 A と蛋白質 B のコンタクトマップ間に作成されたアラインメントの例を示している．点線で結ばれた頂点同士がアラインメントされた残基を示している．この例では，3 つのアラインメントペア (A_1, B_1)，(A_3, B_3)，(A_5, B_4) が存在する．よって，$AL^\phi = \{(A_1, B_1), (A_3, B_3), (A_5, B_4)\}$ となる．

ここで，アラインメントペア $(v_i, w_{\phi(i)})$ と $(v_j, w_{\phi(j)})$ とについて，頂点 v_i と頂点 v_j 間と，頂点 $w_{\phi(i)}$ と頂点 $w_{\phi(j)}$ 間とにコンタクトエッジが存在する場合，つまり，$(v_i, v_j) \in CE_v$ かつ $(w_{\phi(i)}, w_{\phi(j)}) \in CE_w$ が成り立つ場合，コンタクトマップがオーバラップするといい，このオーバラップのことを共通コンタクトと呼ぶ．

図 2 において，(A_1, B_1) と (A_3, B_3) の 2 つのアラインメントペアに着目する．ここで，頂点 A_1 と頂点 A_3 の間，また，頂点 B_1 と頂点 B_3 の間にコンタクトエッジ（図中の太線）が存在するため，(A_1, B_1) と (A_3, B_3) の 2 つのアラインメントペアに共通コンタクトがひとつ存在する．

CMO 問題ではこの共通コンタクト数を最大化するアラインメントペア集合を求める問題である．具体的には，以下のコスト関数 f を最大化する問題として定義される．

$$f(AL^\phi) \quad = \quad \sum g(v_i, w_{\phi(i)}, v_j, w_{\phi(j)}) \qquad (4)$$

ただし，$(v_i, w_{\phi(i)}) \in AL^\phi, (v_j, w_{\phi(j)}) \in AL^\phi$ である．また，関数 g は，$(v_i, v_j) \in CE_v$ でかつ，$(w_{\phi(i)}, w_{\phi(j)}) \in CE_w$ のときに 1 を返し，それ以外は 0 を返す．

III. MEO による発見的解法

本章では，MEO を用いた CMO 問題の発見的解法の詳細内容を示す．

A. 個体と構成要素の定義

蛋白質 v と蛋白質 w を表現するコンタクトマップ $CM_v = (RV_v, CE_v)$ と $CM_w = (RV_w, CE_w)$ とすると，本研究では，アラインメントペア集合 AL^ϕ をそのまま個体 I として定義する．また，アラインメントペアを構成する頂点ひとつひとつを構成要素 $O_i (\in RV_v \cup RV_w)$ とする．

B. 適応度の定義

蛋白質 v と蛋白質 w を表現するコンタクトマップを $CM_v = (RV_v, CE_v)$ と $CM_w = (RV_w, CE_w)$ とすると，個体 I の適応度である大域的適応度は第 3 章で示したコスト関数 f を用い，

$$global\_fitness(I) = \frac{f(I)}{min(|CE_v|, |CE_w|)}, \qquad (5)$$

と定義する．

ここで，頂点 v_k と頂点 $w_{\phi(k)}$ に接続しているコンタクトエッジの中で共通コンタクトであるコンタクトエッジの数をそれぞれ $com(v_k)$，$com(w_{\phi(k)})$ とする．

構成要素 O_i の適応度である局所的適応度は，構成要素 O_i に対応する頂点の次数で頂点が持つ共通コンタクト数を次数で割った値とする．

$$local\_fitness(O_i) = \begin{cases} \frac{com(v_k)}{dig(v_k)} & if\ O_i = v_k \in CV_v \\ \frac{com(w_{\phi(k)})}{dig(w_{\phi(k)})} & if\ O_i = w_{\phi(k)} \in CV_w. \end{cases} \qquad (6)$$

上記の式で，頂点の次数を $dig(v_k)$，$dig(w_{\phi(k)})$ とする．ただし，次数が 0 である頂点は常に局所的適応度は 0 とする．

C. 初期個体

比較するふたつの蛋白質の残基間の構造的な類似度をスコア関数（スコア行列）により求め，動的計画法を用いて，スコア行列 D の各要素 $D_{i,j}$ を計算する．スコア行列が算出できたら，スコア $D_{n,m}$ から最大値を算出する経路をトレースバックしていく．つまり，スコア $D_{n,m}$ を算出するのに，上，左上，左のどちらの要素の数値が採用されたかトレースバックする．詳しい式は文献 [7] を参照されたい．

Algorithm 1: 改良版 MEO による CMO 問題の解法

input : 蛋白質 A と蛋白質 B の座標配列データ，カットオフ値 $cutoff$，最大世代数 $gmax$，近傍個体生成数 $nmax$
output: 最良個体 I_{best} が持つアラインメントペア集合 AL^{ϕ}

```
1  コンタクトマップ CM_A と CM_B を作成し，類似度行列 S を生成する．
2  類似度行列 S を用いて動的計画法により初期アラインメントを生成し，初期アラインメントを I とする．
3  I_best = I
4  g = 0
5  while g < gmax do
6      I の全構成要素 O_i について，局所的適応度 local_fitness(O_i) を算出する．
7      NI = make_neighbor_individuals(I,nmax)
8      I =best(NI)
9      if global_fitness(I) > global_fitness(I_best) then
10         I_best = I
11     g++
12 return 最良個体 I_best が持つアラインメントペア集合 AL^φ
```

D. アルゴリズム

提案手法のアルゴリズムを Algorithm 1 に示す．最初に，入力した蛋白質の座標配列データからコンタクトマップと類似度行列 S を作成する．次に，動的計画法を用いて初期アラインメントを求める．そして，初期アラインメントを初期個体，また現時点の最良解として設定する．続いて，ユーザが指定した世代数まで MEO を用いて，状態遷移を繰り返す．最初に，構成要素についてその適応度 $local\_fitness(O_i)$ を求める．次に，関数 **make_neighbor_indivisuals** を呼び出し，個体の近傍個体となる複数の個体（近傍個体集合 NI とする）を生成する．近傍個体集合 NI から個体の適応度が最良の個体をひとつ選択し，次世代の個体とする．もし，次世代の個体が最良個体よりも評価の高い個体ならば最良個体としてその個体のコピーを保存する．

関数 **make_neighbor_individuals** は近傍個体を生成する関数である．最初に，個体 I のコピーを作成し，$I_{neighbor}$ に保存する．$I_{neighbor}$ の構成要素をその局所的適応度を用いて，ルーレット選択でひとつ選択する．次に，選択した構成要素を状態遷移する．状態遷移の方法については，次節に示す．状態遷移を行った $I_{neighbor}$ を近傍個体集合 NI に保存する．この一連の処理を $nmax$ 回繰り返すことで，$nmax$ 個の近傍個体を作成し，作成した近傍個体集合 NI を返す．

E. 状態遷移

構成要素の状態遷移はアラインメントペアの組み換えにより行う．例えば，構成要素 O_k が状態遷移の候補として選択され，$O_k = v_k$ と仮定する．アラインメントペア $(v_k, w_{\phi(k)})$ について，v_k を蛋白質 v の他の頂点に変更する．逆に，$O_k = w_{\phi(k)}$ と仮定すると，アラインメントペア $(v_k, w_{\phi(k)})$ について，$w_{\phi(k)}$ を蛋白質 w の他の頂点に変更する．アラインメントの組み換えでは，ランダムに最初に選んだ他の頂点を選択する即時移動戦略と，組み換え可能なすべての頂点の候補の中から個体の大域的適応度の高くなる頂点を選択する最良移動戦略の 2 種類が考えられる．本研究では，最良移動戦略を用い，組み換え可能なすべての頂点の中から個体の大域的適応度が最も大きくなる頂点を選び，アラインメントの組み換えをする．

表 I
SOKOL テストデータセット

PDB ID	残基数	コンタクトエッジ数
1bpi	58	195
1knt	55	192
2knt	58	200
5pti	58	190
1vii	36	120
1cph	21	65
3ebx	73	275
6ebx	62	205
1era	62	208

表 II
XU らのデータセット

PDB ID	残基数	コンタクトエッジ数
1b00A	122	488
1b00B	122	423
1dbwA	125	474
1qmpA	125	454
1qmpC	125	452
4tmyA	118	473
1byoA	99	355
1dpsB	154	586
1dpsC	154	585
1nat	119	435
1amk	250	1086
2pcy	99	357
8timA	247	930
1aw2B	254	1043
1b9bA	252	953

IV. 評価実験

評価実験では，文献 [6] で用いられている蛋白質立体構造データ 24 件を評価用の蛋白質立体構造データとして用いた．蛋白質立体構造データ 24 件の内訳は，9 件が Sokol テストデータセット（表 I）で，残り 15 件は，オリジナルの Skolnick データセット 40 件から Xu らが抜き出して評価実験に用いた 15 件の蛋白質立体構造データ（表 II）である．オリジナルの Skolnick データセットと区別するために本論文では Xu らのデータセットと呼ぶ．

実験では，提案手法と EO を用いた発見的解法とで最良個体の共通コンタクト数と RMSD 値を比較する．ただし，実験で使用する EO を用いた発見的解法は，文献 [6] に示された手法とは初期解として動的計画法を用いている点が異なる．EO を用いた発見的解法では，100 秒間，世代交代を繰り返す．提案手法では，近傍個体生成数を 100 と設定し，同じく，100 秒間，世代交代を繰り返す．また，それぞれ 3 回ずつ実行し，得られた最良個体の共通コンタクト数の平均と RMSD 値を求める．

RMSD は平均二乗偏差（Root Mean Square Deviation）のことで，蛋白質立体構造の非類似性や誤りの指標として用いらている．例えば，蛋白質 A と蛋白質 B のアライメントされた残基の座標を $A = \{\mathbf{a}_1, \mathbf{a}_2, .., \mathbf{a}_n\}$, $B = \{\mathbf{b}_1, \mathbf{b}_2, .., \mathbf{b}_n\}$ とすると，

表 III
SOKOL テストデータセットの実験結果

PDB 番号		共通コンタクト数		RMSD	
蛋白質 A	蛋白質 B	提案手法	EO	提案手法	EO
1bpi	1knt	164	**168**	3.977029	**2.894972**
1bpi	2knt	171	**175**	**1.33919**	2.171393
1bpi	5pti	185	**188**	1.258139	**1.226277**
1knt	2knt	188	**191**	3.345494	**1.758625**
1knt	5pti	163	**169**	3.96366	**2.903412**
1vii	1cph	49	**56**	12.1549	**5.825044**
2knt	5pti	166	**175**	3.923252	**2.045472**
3ebx	1era	146	**174**	**3.294866**	4.670349
3ebx	6ebx	164	**197**	4.375532	**2.091195**
6ebx	1era	177	**185**	2.789644	**1.313177**

表 IV
XU らのデータセットにおけるフラボドキシンに似た形状を持つ蛋白質データの組合せの実験結果

PDB 番号		共通コンタクト数		RMSD	
蛋白質 A	蛋白質 B	提案手法	EO	提案手法	EO
1b00A	1dbwA	**291**	268	**6.460011**	6.938559
1b00A	1nat	**290**	288	7.826805	**6.770336**
1b00A	1qmpC	**317**	289	8.401165	**8.011432**
1nat	1b00B	279	**283**	**12.85953**	13.39517
1nat	1dbwA	365	**370**	9.133495	**4.545009**
1nat	4tmyA	317	**325**	14.00019	**6.676143**
1qmpC	1b00B	**294**	283	**13.46852**	14.09271
1qmpC	4tmyA	**324**	318	12.52931	**7.981899**
4tmyA	1b00B	235	**242**	16.82389	**14.05792**

$$\mathrm{RMSD} = \min_{T} \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(||T(\mathbf{a}_i) - \mathbf{b}_i||\right)^2}$$

ただし，写像 T は，回転ベクトル（行列）を R，平行移動ベクトルを $\mathbf{p}$ とする以下の式で定義される．

$$T(\mathbf{a}_i) = R\mathbf{a}_i + \mathbf{p}$$

Sokol テストデータセットにおける組合せの実験結果を表 III に示す．Sokol のテストデータセットにおける 10 組の組合せについては，全ての組み合わせにおいて共通コンタクト，RMSD 共に EO で良い結果が得られた．ただし，若干の差はあるものの，両者ほぼ同じ結果が得られ，共通コンタクト数は近差である．これは，Sokol テストデータセットに含まれる蛋白質データは残基数が少なく，大部分が類似する構造であるため，EO による発見的解法でも十分に良い最良個体が得られるためである．

次に，Xu らのデータセットにおけるフラボドキシンに似た形状を持つ蛋白質のデータセットの組合せ（表 IV）と，Xu らのデータセットにおける異なる形状を持つ蛋白質データの組合せ（表 V）においては，提案手法の方が EO を用いた発見的解法よりも良い結果が得られた．RMSD 値においても，ほぼその結果に比例する形となった．ただいくつか例外もあり，共通コンタクトが優れていても RMSD 値で負けているものも若干あり，必ずしも共通コンタクト数が多ければ良いというわけではないことがわかった．

V. まとめ

本論文では，MEO を用いた CMO 問題の発見的解法を示すとともに，CMO 問題で求めた最適構造アラインメントに対する RMSD(root-mean-square deviation) による性能評価結果を報告した．その結果として，Sokol テストセットでは EO が，Xu でのデータセットでは MEO がそれぞれ優れた結果を出した．RMSD もその結果に準ずる形をとったが，いつくか例外もあり，コンタクト数が優れているにもかかわらず，RMSD にも劣るデータが存在した．その点に関して，視野にいれ，今後の研究に活かしていく必要がある．

表 V
XU らのデータセットにおける異なる形状を持つ蛋白質データの組合せの実験結果

PDB 番号		共通コンタクト数		RMSD	
蛋白質 A	蛋白質 B	提案手法	EO	提案手法	EO
1b00A	1bawA	**157**	**157**	13.74165	**13.27508**
1b00A	1byoA	**156**	151	15.14218	**13.66219**
1b00A	1dpsB	**252**	237	**16.78592**	17.77635
1nat	1amk	**247**	185	**13.68861**	15.23867
1nat	1dpsB	**254**	246	**18.3215**	18.41292
1qmpC	2pcy	157	**162**	14.49383	**14.48558**
1qmpA	8timA	**253**	203	**13.58972**	14.75359
4tmyA	1bawA	145	**158**	15.96979	**14.53201**
4tmyA	1amk	**194**	171	19.11666	**15.20745**
4tmyA	1dpsC	**240**	221	20.10043	**18.18274**
1bawA	1aw2B	**176**	140	**16.61772**	18.94407
1bawA	1b9bA	**179**	145	**16.01051**	17.18831
1bawA	1dpsB	**184**	174	**17.01611**	18.33449

謝辞

本研究の一部は，広島市立大学・特定研究費（一般研究，研究課題名「時空間文書ストリーム上におけるバースト領域の抽出手法」） の支援により行われた．

参考文献

[1] C. I. Branden and J. Tooze, *Introduction to Protein Structure*. Garland Publishing, 1999.

[2] T. WR and O. CA, “Protein structure alignment,” *Journal of Molecular Biology*, vol. 208, no. 1, pp. 1–22, 1989.

[3] M. J. Sippl and M. Wiederstein, “A note on difficult structure alignment problems,” *Bioinformatics*, vol. 24, no. 3, pp. 426–427, 2008.

[4] R. Andonov, N. Malod-Dognin, and N. Yanev, “Maximum contact map overlap revisited,” *Journal of Computational Biology*, vol. 18, no. 1, pp. 27–41, 2011.

[5] S. Boettcher and A. Percus, “Nature’s way of optimizing,” *Artificial Intelligence*, vol. 119, no. 1-2, pp. 275–286, 2000.

[6] H. Lu, G. Yang, and L. F. Yeung, “Extremal optimization for the protein structure alignment,” in *Proceedings of the 2009 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 15–19, 2009.

[7] A. Nakada, K. Tamura, and H. Kitakami, “Optimal protein structure alignment using modified extremal optimization,” in *Proceedings of SMC2012*, pp. 697–702, 2012.

問い合わせ先
〒 731-3194
広島市安佐南区大塚東 3-4-1
広島市立大学大学院情報科学研究科
中田 章宏