Vol.2013-MPS-92 No.13
2013/2/27

CMO問題に対する改良版EOを用いた発見的解法

中田 章宏[†1,a)]　田村 慶一[†1,b)]　北上 始[†1,c)]　高橋 誉文[†1]

概要：蛋白質の類似構造を抽出する手法として蛋白質立体構造アラインメントがある．蛋白質立体構造アラインメントは構造的な類似性を使い，比較する蛋白質同士の残基間の対応関係を求める．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）問題である．CMO問題はコンタクトマップと呼ばれるグラフについて，頂点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構造の数を最大化する問題である．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，進化計算を用いた手法などの研究が行われている．本論文では，CMO問題に対する改良版EOを用いた発見的解法を提案する．提案手法の特徴は，(1) 世代交代に改良版EOを用いること，(2) 動的計画法を用いて初期個体を作成すること，(3) 改良版EO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ことである．提案手法を実際に実装し，評価実験を行った結果，EOによる発見的解法よりも評価の高い最良解が得られた．

キーワード：CMO問題, Extremal Optimization，発見的解法

1. 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，DNA塩基配列から翻訳されたアミノ酸配列から作られ，特有のかたち（三次元の立体構造）を形成する．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], [2] 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 [3] である．蛋白質立体構造アラインメントは構造的な類似性を使い，比較する蛋白質同士の残基間の対応関係を求める．蛋白質はアミノ酸が鎖状にペプチド結合した高分子であり，元々のアミノ酸部分を残基と呼ぶ．残基間の対応関係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 [4].

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 [5], [6] 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．CMO問題を解くことで，共通コンタクトの数が類似度，また，アラインメントが示す残基間の対応関係が類似構造として抽出される．

CMO問題はNP困難な問題であることが示されているため [7]，線形計画問題に置き換えて定式化するアプローチ [8]，線形計画問題に対してラグランジュ緩和を導入した手法 [9]，緩和問題を作成し，分枝限定法を用いた手法 [10] などが提案されている．また，CMO問題を最大クリーク問題に置き換える手法 [11] や，CMO問題をMCS問題に置き換え，動的計画法を用いてMCS問題を解く手法 [12] が提案されている．さらに，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）[13] などの発見的解法を用いた手法の研究も行われている．その中でもEOを用いたCMO問題の発見的解法 [14] は他の手法と比較して，より良い最良解が得られることが示されている．

[†1] 現在，広島市立大学大学院情報科学研究科
Presently with Graduate School of Information Sciences, Hiroshima City University
[a)] mw67025@edu.ipc.hiroshima-cu.ac.jp
[b)] ktamura@hiroshima-cu.ac.jp
[c)] kitakami@hiroshima-cu.ac.jp

本論文では，改良版 EO[15] を用いた CMO 問題の発見的解法を提案する．提案手法は，

（1）世代交代に改良版 EO を用いる，
（2）初期個体は動的計画法を用いて作成する，
（3）改良版 EO 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，

という 3 つの特徴を持っている．

改良版 EO では，複数の近傍個体の中から一番良い個体を次世代の個体として選択する．複数の近傍個体を生成し，状態遷移を繰り返すため，EO と比較して局所解に陥りにくい手法であることが示されている．また，動的計画法を用いて初期アラインメントを作成し，その初期アラインメントを初期個体とする．動的計画法で求めたアラインメント結果を初期個体として用いることで，ある程度最適な構造からスタートできる．

提案手法を実際に実装し，PDBj（Protein Data Bank Japan）から取得した 24 件の蛋白質立体構造データを用い，33 個の組み合わせにおいて，提案手法の評価を行った．評価実験の結果，提案手法は 24 組の蛋白質の組み合わせにおいて，EO を用いた発見的解法と同等，または，評価の高い最良解を求めることができ，CMO 問題に対する新しい解法として有効であることを示すことができた．

本論文の構成は以下の通りである．第 2 章では，関連研究について述べる．第 3 章では，CMO について，その問題定義を示す．第 4 章では改良版 EO について説明し，第 5 章では提案手法を示す．第 6 章で評価実験の結果を示し，第 7 章において，本論文をまとめる．

2. CMO 問題

蛋白質の残基を頂点，近接する残基間をコンタクトエッジと呼ばれる辺で結んだグラフをコンタクトマップと呼ぶ．コンタクトマップの各頂点は残基の中心座標と結び付けられる．残基の中心座標として，本研究では他の CMO 問題を対象とした研究と同様に Cα 原子の座標を用いる．また，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．ただし，隣接する残基同士は除く．

図 1 にコンタクトマップの例を示す．図 1 の上部に蛋白質 A の残基の空間的な構造を示し，図 1 の下部に蛋白質 A のコンタクトマップを示す．ここで，蛋白質 A の残基 1 に着目すると，残基 1 の中心座標から $cutoff$ 以内に残基 3 と残基 5 が存在する．よって，コンタクトマップにおいて，頂点 A_1 と頂点 A_3 間にコンタクトエッジ (A_1, A_3)，頂点 A_1 と頂点 A_5 間にコンタクトエッジ (A_1, A_5) が作成されている．同様に，残基 3 と残基 5 間の距離も $cutoff$ 以内であるため，コンタクトエッジ (A_3, A_5) が作成されている．

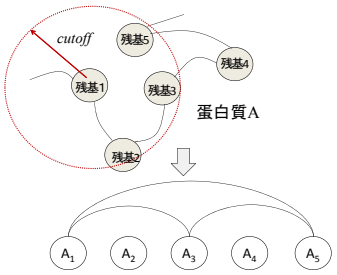

図 1 コンタクトマップの例
Fig. 1 Example of Contact Map

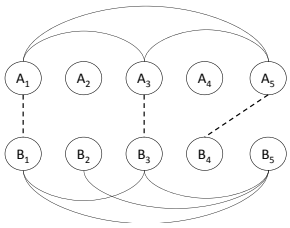

図 2 アラインメントの例
Fig. 2 Example of Alignment

ここで，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，

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

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とする．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 $(RV_v^+ \in RV_v, RV_w^+ \in RV_w)$ 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単射として，

$$\phi : RV_v{}^+ \rightarrow RV_w{}^+, v_i \rightarrow w_{i^\phi}, \quad (2)$$

と定義すると，アラインメントペア集合 AL^ϕ は，

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

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．

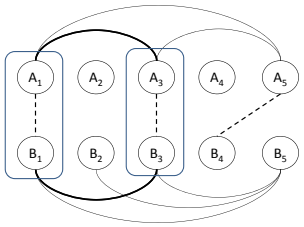

図 3 共通コンタクトの例
Fig. 3 Example of Common Contact

ただし，アラインメントペアは次の条件を満たす必要がある．

$$i<j \mapsto i^{\phi}<j^{\phi}. \tag{4}$$

この制約は，アラインメントペア間に交差が生じないことを示している．

ここで，アラインメントペア (v_i, w_{i^ϕ}) と (v_j, w_{j^ϕ}) とについて，頂点 v_i と頂点 v_j 間と，頂点 w_{i^ϕ} と頂点 w_{j^ϕ} 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 $(w_{i^\phi}, w_{j^\phi}) \in CE_w$ 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．

例えば，**図 3** 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．

$$f(AL^{\phi}) = \sum_{(v_i, w_{i^\phi}) \in AL^{\phi}, (v_j, w_{j^\phi}) \in AL^{\phi}} g(v_i, w_{i^\phi}, v_j, w_{j^\phi})$$

$$g(v_i, w_{i^\phi}, v_j, w_{j^\phi}) = \begin{cases} 1 & if\ (v_i, v_j) \in CE_v \\ & and\ (w_{i^\phi}, w_{j^\phi}) \in CE_w \\ 0 & otherwise. \end{cases} \tag{5}$$

図 1 の例では，共通コンタクトは 1 つ存在する．よって，このコスト関数 f は値として 1 を返す．

3. 改良版 EO

EO は個体の中で適応度が悪い構成要素を選択し，その構成要素を状態遷移することで，個体の適応度を向上させていく発見的解法である．構造物において強度が弱い部分を補強することでその構造物の強度が向上する考えに基づいている．改良版 EO では，複数の近傍個体を生成し，近傍個体の中で最良の個体を次世代の個体として選択する．個体のコピーを複数個用意し，コピーして作成した各個体についてルーレット選択を用いて構成要素を選択する．そして，選択した構成要素を状態遷移させる．最後に，近傍固体中で一番適応度の高い個体を次世代の個体として選択する．通常の EO ならば，もし選択された結果が改悪となるにしても，その状態遷移を行うしかないが，改良版 EO ならば近傍個体を複数生成するため改良となる個体が生成される可能性も高まり，結果，改悪へと進む可能性が減る．

4. 提案手法

本章では，提案手法である改良版 EO を用いた CMO 問題の発見的解法の詳細内容を示す．

4.1 個体と構成要素の定義

改良版 EO を CMO 問題に適応するにあたり，はじめに個体とその構成要素を定義する必要がある．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．例えば，図 2 では，$I = \{(A_1, B_1), (A_3, B_3), (A_5, B_4)\}$ であり，個体は $O_1 = A_1$，$O_2 = B_1$，$O_3 = A_3$，$O_4 = B_3$，$O_5 = A_5$，$O_6 = B_4$ の 6 つの構成要素から構成される．

4.2 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 2 章で示したコスト関数 f を用い，

$$global\_fitness(I) = \frac{f(I)}{min(|CE_v|, |CE_w|)}, \tag{6}$$

と定義する．例えば，図 2 の例では，$|CE_{\mathrm{A}}| = 3$，$|CE_{\mathrm{B}}| = 4$ で，$f(I) = 1$ であるため，$global\_fitness(I) = 1/3$ となる．

ここで，頂点 v_k と頂点 w_{k^ϕ} 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 $com(v_k)$，$com(w_{k^\phi})$ とする．

$$com(v_k) = com(w_{k^\phi}) = \sum_{(v_j, w_{j^\phi}) \in AL^{\phi}} g(v_k, w_{k^\phi}, v_j, w_{j,j^\phi}). \tag{7}$$

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．

$$local\_fitness(O_i) = \begin{cases} \frac{com(v_k)}{dig(v_k)} & if\ O_i = v_k \in CV_v \\ \frac{com(w_{k^\phi})}{dig(w_{k^\phi})} & if\ O_i = w_{k^\phi} \in CV_w. \end{cases} \tag{8}$$

上記の式で，頂点の次数を $dig(v_k)$，$dig(w_{k^\phi})$ 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．

4.3 初期個体

初期個体は残基間の構造的な類似度を使い，動的計画法を用いて作成する．最初に，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）として定義する．残基間の類似度 $s_{i,j}$ は以下の定義式で求める．

$$s_{i,j} = \alpha \times \left(\frac{min(dig(v_i), dig(w_j))}{max(dig(v_i), dig(w_j))} + \frac{min(sd(v_i), sd(w_j))}{max(sd(v_i), sd(w_j))} \right), \quad (9)$$

ここで，$dig(v_i)$ と $dig(w_i)$ は頂点の次数であり，$sd(v_i)$ と $sd(w_i)$ はコンタクトエッジで接続している他の頂点が示す残基との空間的な距離の総和である．また，α は係数である．

次に，残基の並びを配列要素として，最適アラインメントを求める．はじめに，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．

$$D_{i,0} \leftarrow i \times g \quad i = 0, \cdots, n$$
$$D_{0,j} \leftarrow j \times g \quad j = 0, \cdots, m$$

$$D_{i,j} \leftarrow min \begin{cases} D_{i-1,j} + g \\ D_{i,j-1} + g \\ D_{i-1,j-1} + \frac{max(S) - s_{i,j}}{max(S)} \end{cases} \quad (10)$$

ここで，式中の $max(S)$ は類似行列 S の最大値，$s_{i,j}$ は類似行列 S の第 (i,j) 要素の値である．また，g はペナルティスコアであり，次の値を用いる．

$$g = \frac{\sum_{k=0}^{n} \sum_{l=0}^{m} \frac{max(S) - s_{k,l}}{max(S)}}{n \times m}. \quad (11)$$

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．

4.4 アルゴリズム

提案手法のアルゴリズムを Algorithm1 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改良版 EO 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．

Algorithm2 に関数 **make_neighbor_individuals** の内容を示す．最初に，個体のコピー $I_{neighbor}$ を作成する．$I_{neighbor}$ 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．そして，状態遷移を繰り返し，複数の近傍個体を作成し，作成した近傍個体集合 NI を返す．

Algorithm 1: 提案手法

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}

```
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．
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．
I_best = I
g = 0
while g < gmax do
    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．
    NI = make_neighbor_individuals(I,nmax)
    I =best(NI)
    if global_fitness(I) > global_fitness(I_best) then
        I_best = I
    g++
return 最良個体 I_best が持つアラインメントペア集合 AL^φ
```

Algorithm 2: make_neighbor_individuals

input : 個体 I，最大近傍個体生成数 $nmax$
output: 近傍個体集合 NI

```
n = 0
NI = φ
while n < nmax do
    I_neighbor = I
    局所的適応度 local_fitness(O_i) のルーレット選択により構成要素 O_k を選択する．
    選択した構成要素 O_k が示す頂点について，当該頂点を移動することで作成可能なアラインメントペアをすべて調べ，個体の大域的適応度が最も大きくなるアラインメントペアを選択し，置き換える．
    NI = NI ∪ I_neighbor
    n++
return NI
```

4.5 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 $(v_k, w_{k^{\phi}})$ について，v_k を蛋白質 v の他の頂点に変更する．逆に，

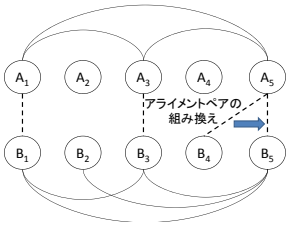

図 4 状態遷移の例
Fig. 4 Example of Change State

表 1 Sokol テストデータセット
Table 1 Sokol Test Data Set

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 2 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
Table 2 Data Set of Proteins with Flavodoxin-like Fold

PDB ID	残基数	コンタクトエッジ数
1b00a	122	488
1dbwa	125	474
1nat	119	435
1qmpc	125	452
1b00b	122	423
4tmya	118	473

$O_k = w_{k^\phi}$ と仮定すると，アラインメントペア (v_k, w_{k^ϕ}) について，w_{k^ϕ} を蛋白質 w の他の頂点に変更する．ただし，状態遷移することで，アラインメントペアで交差が生じないように制約を満たす必要がある．

図 4 に状態遷移におけるアラインメントペアの組み換えの例を示す．図 4 の例では，B_4 が選択された構成要素で，アライメントペア (A_5, B_4) をアライメントペア (A_5, B_5) に組み換えた例を示している．このアライメントペアの組み換えで，組み換え前と比較して共通コンタクトの数が増加し，組み換え前と比較して評価の高いアライメントとなっていることが分かる．

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．

5. 評価実験

提案手法を評価するために，PDBj(Protein Data Bank Japan) に登録がある蛋白質立体構造データ 24 件（**表 1**, **表 2** と **表 3**）を用い，評価実験を行った．表 2 と表 3 とで重複があるのは，両者に含まれるデータがあるためである．各コンタクトマップは $cutoff = 6.75$Å として作成した．表 1，表 2 と表 3 にそれぞれの蛋白質の残基数とコンタクトエッジ数を示す．

表 3 クプレドキシンの形状を持つ蛋白質のデータセット
Table 3 Data Set of Proteins with Cupredoxins Fold

PDB ID	残基数	コンタクトエッジ数
1b00a	122	488
1bawa	105	387
1byoa	99	355
1dpsb	154	586
1nat	119	435
1amk	250	1086
1qmpc	125	452
2pcy	99	357
1qmpa	125	454
8tima	247	930
4tmya	118	473
1dpsc	154	585
1aw2b	254	1043
1b9ba	252	953

評価実験では，最良個体の共通コンタクト数の平均について，提案手法と EO による発見的解法との比較を行った．ただし，実験で使用する EO を用いた発見的解法は，文献 [14] 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，10 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，10 秒間，世代交代を繰り返す．また，それぞれ 30 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を求める．

表 4 に実験結果を示す．蛋白質 33 組の組合せについて評価を行った．表 4 から，33 組中 24 組で提案手法が良いか，EO による発見的解法と同じ共通コンタクト数が得られた．Sokol テストデータセット（表 1）の組合せでは，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，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（表 2）の組合せでは，EO による発見的解法の方が提案手法と比較して良い最良個体が得られている．このデータセットは残基数が多く，また，6 割~ 8 割の連続構造が似ているため，局所解が少ない．よって，解の探索が早く進む EO による発見的解法の方が有利であるため，EO による発見的解法の方が良い最良個体を得られている．

最後に，クプレドキシンの形状を持つ蛋白質のデータセット（表 3）は，提案手法の方が EO による発見的解法と比較して大幅に良い最良個体が得られている．このデータ

表 4　最良個体の共通コンタクト数の平均
Table 4　Average of Number of Common Contact of Best Solution

Protein A	Protein B	提案手法	EO
1bpi	1knt	**175**	**175**
1bpi	2knt	**180**	177
1bpi	5pti	**184**	180
1knt	1bpi	**175**	**175**
1knt	2knt	187	**188**
1knt	5pti	**175**	**175**
1vii	1cph	**57**	53
2knt	5pti	**179**	174
3ebx	1era	**185**	**185**
3ebx	6ebx	**199**	**199**
6ebx	1era	**170**	164
1b00a	1dbwa	328	**354**
1b00a	1nat	369	**372**
1b00a	1dbwa	338	**348**
1nat	1b00b	344	**352**
1nat	1dbwa	328	**364**
1nat	4tmya	340	**343**
1qmpc	1b00b	**342**	340
1qmpc	4tmya	**368**	363
4tmya	1b00b	314	**319**
1b00a	1bawa	199	**200**
1b00a	1byoa	**183**	179
1b00a	1dpsb	**293**	279
1nat	1amk	**281**	216
1nat	1dpsb	**300**	290
1qmpc	2pcy	**197**	177
1qmpa	8tima	**287**	207
4tmya	1bawa	**193**	**193**
4tmya	1amk	**234**	205
4tmya	1dpsc	**283**	261
1bawa	1aw2b	**193**	138
1bawa	1b9ba	**209**	148
1bawa	1dpsb	**218**	198

セットも残基数が多く，また共通コンタクト数も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（表 2）と比較して多い．ただし，部分的な構造が似ているため，局所解に陥りやすい問題となっている．

6. まとめ

本論文では改良版 EO を用いた CMO 問題の発見的解法を提案した．提案手法は，(1) 世代交代に改良版 EO を用いる，(2) 初期個体は動的計画法を用いて作成する，(3) 改良版 EO 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，という 3 つの特徴を持っている．提案手法を実際に実装し，PDBj（Protein Data Bank Japan）から取得した 25 件の蛋白質立体構造データを用い，33 個の組み合わせにおいて，提案手法の評価を行った．評価実験の結果，提案手法は 24 組の蛋白質の組み合わせにおいて，EO を用いた発見的解法と同等，また，評価の高い最良解を求めることができ，CMO 問題に対する新しい解法として有効であることを示すことができた．これからの課題として，初期個体の生成方法の工夫，また，突然変異によるアラインメント数の動的な増減などがあげられる．また，多様性という観点では個体数がひとつであるのは限界があるので，複数個体を用いた手法などの検討なども必要である．

謝辞　本研究の一部は，文部科学省・科学研究費補助金（若手研究 (B)，課題番号：23700124），日本学術振興会・科学研究費補助金 (基盤研究 (C)，課題番号：20500137) の支援により行われた．

参考文献

[1] Branden, C. I. and Tooze, J.: *Introduction to Protein Structure*, Garland Publishing (1999).
[2] Lancia, G. and Istrail, S.: Protein Structure Comparison: Algorithms and Applications, *Mathematical Methods for Protein Structure Analysis and Design*, pp. 1–33 (2003).
[3] WR, T. and CA, O.: Protein Structure Alignment, *Journal of Molecular Biology*, Vol. 208, No. 1, pp. 1–22 (1989).
[4] Sippl, M. J. and Wiederstein, M.: A note on difficult structure alignment problems, *Bioinformatics*, Vol. 24, No. 3, pp. 426–427 (2008).
[5] Godzik, A. and Skolnick, J.: Flexible algorithm for direct multiple alignment of protein structures and sequences, *Computer Applications in the Biosciences*, Vol. 10, No. 6, pp. 587–596 (1994).
[6] Andonov, R., Malod-Dognin, N. and Yanev, N.: Maximum Contact Map Overlap Revisited, *Journal of Computational Biology*, Vol. 18, No. 1, pp. 27–41 (2011).
[7] Goldman, D., Papadimitriou, C. H. and Istrail, S.: Algorithmic Aspects of Protein Structure Similarity, *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pp. 512–522 (1999).
[8] Carr, R. D., Lancia, G., Istrail, S. and Genomics, C.: Branch-and-Cut Algorithms for Independent Set Problems: Integrality Gap and an Application to Protein Structure Alignment, *SAND Report SAND2000-2171, Sandia National Laboratories, 2000* (2000).
[9] Lancia, G., Carr, R., Walenz, B. and Istrail, S.: 101 optimal PDB structure alignments: a branch-and-cut algorithm for the maximum contact map overlap problem, *Proceedings of the fifth annual international conference on Computational biology*, pp. 193–202 (2001).
[10] Xie, W. and Sahinidis, N. V.: A Reduction-Based Exact Algorithm for the Contact Map Overlap Problem, *Journal of Computational Biology*, Vol. 14, No. 5, pp. 637–654 (2007).
[11] Balaji, S., Swaminathan, V. and Kannan, K.: A Simple Algorithm for Maximum Clique and Matching Protein Structures, *International Journal of Combinatorial Optimization Problems and Informatics*, Vol. 1, No. 2, pp. 2–11 (2010).
[12] Jain, B. J. and Lappe, M.: Joining softassign and dynamic programming for the contact map overlap problem, *Proceedings of the 1st international conference on Bioinformatics research and development*, pp. 410–423 (2007).
[13] Boettcher, S. and Percus, A.: Nature's way of optimizing, *Artificial Intelligence*, Vol. 119, No. 1-2, pp. 275–286 (2000).
[14] Lu, H., Yang, G. and Yeung, L. F.: Extremal Optimization for the Protein Structure Alignment, *Proceedings of the 2009 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 15–19 (2009).
[15] 田村慶一，森　康真，北上　始：Extremal Optimization による調停グラフの交差数減少，情報処理学会論文誌. 数理モデル化と応用，Vol. 49, No. 4, pp. 105–116 (2008).