改良版 EO を用いた蛋白質立体構造アラインメント
Modified Extremal Optimization for Optimal Protein Structure Alignment

中田　章宏*　田村　慶一**　北上　始**　三村　賢太*　高橋　誉文*
Akihiro Nakada　Keiichi Tamura　Hajime Kitakami　Kenta Mimura　Shigehumi Takahasi
広島市立大学大学院情報科学研究科
Email: *{mw67025, mu67028, dt65002}@edu.ipc.hiroshima-cu.ac.jp,**{ktamura,kitakami}@hiroshima-cu.ac.jp

Abstract—**Extracting similar three-dimentional structures between proteins is one of the most important challenges for bioinformatics. The protein structure alignment is one of the most effective methods to extract similar three-dimentional structures of two proteins, and CMO(Contact Map Overlap) is formulated as a combinational optimization for the protein structure alignment. In this paper, we propose a novel heuristic for CMO problem using Modified-EO.**

I.　はじめに

蛋白質立体構造間の類似構造抽出はバイオインフォマティクスにおいて重要な研究課題のひとつである．そこで，類似構造抽出に関する研究[1]，[2]が盛んに行われており，その手法のひとつに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，文字列のアラインメントと同様に，構造的な類似性を使い，比較するふたつの蛋白質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は残基間の最適な対応関係を求め，類似度を計算するため，他の類似性を求める手法と異なり予め対応関係を求める必要がないという特徴がある．

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問題が CMO（Contact Map Overlap）問題[3]である．CMO 問題では，頂点を蛋白質の残基，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構造を作成する．CMO 問題は，コンタクトマップ間の頂点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．

本研究では，改良版 EO（Extremal Optimization）[4]を用いた CMO 問題の解法を提案する．提案手法の特徴は次の通りである．

(1) 改良版 EO を用いて世代交代を繰り返すため，局所最適解におちいりにくい．EO[5]では個体の構成要素一つを状態遷移させることで，その個体の適応度の向上を図る．改良版 EO は，確率的に複数の構成要素を選択し，選択した複数の構成要素を状態遷移した近傍解を生成する．そして，近傍解の中で最も評価の高い解を次世代の個体として選択する．

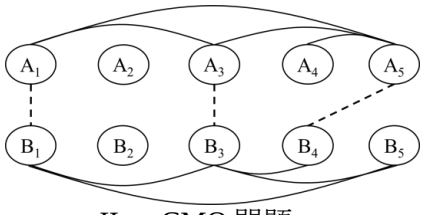
(2) 初期解を動的計画法により作成しているため，最適な構造から解の探索をスタートできる．残基間の構造的な類似度を使い，動的計画法により準最適な構造アラインメントを求める．

(3) 改良版 EO を単純に適用すると，初期解で得られたアラインメント数に得られる解が依存する．そこで，アラインメントの増減を行う突然変異操作を世代交代時に確率的に実施する．

提案手法を実際に実装し，Hengyuu Lu らの実験データを用い EO による解法[6]と比較実験を行った．比較の結果，アラインメント 33 組全てで提案手法の方が EO による解法と比較して良い最良解が得られた．

本論文の構成は以下の通りである．第 2 章では，CMO 問題の定義を示す．第 3 章で，提案手法を説明する．第 4 章で評価実験の結果を示し，第 5 章で本論文のまとめを行う．

図 1. Contact Map の例



II.　CMO 問題

蛋白質の残基を頂点，近接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 Cα 原子の座標を用いる．また，近接残基とは，2 つの連続性のない残基iと残基jの間の空間距離が，与えられたカットオフ距離 *cutoff* 以内であることを示す．

図 1 に，2 つの蛋白質 A，B のコンタクトマップの例を示す．各残基を順番に頂点A_i，B_jと表現する．図 1 中の実線はコンタクトエッジを示している．例えば，A_1とA_3の間にはコンタクトエッジ $(\mathrm{A}_1, \mathrm{A}_3)$ が存在しているため，残基A_1と残基A_3間の距離はカットオフ距離 *cutoff* 以内であることを示している．

本論文では，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，$RV_v = \{v_1, v_2, \ldots, v_n\}$は頂点集合であり，

$$CE_v = \{ (v_i, v_j) \mid v_i \in RV_v, v_j \in RV_v, i+1 < j, \mathbf{dist}(v_i, v_j) < cutoff \}$$

はコンタクトエッジの集合を表す.

蛋白質vと蛋白質wを表現するコンタクトマップCM_v, CM_w間の部分頂点集合$(RV_v^+ \in RV_v, RV_w^+ \in RV_w)$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，

$$\emptyset = RV_w^+ \to RV_w^+, v_i \to w_{i^\emptyset},$$

と定義すると，アラインメントペア集合 AL は，

$$\mathrm{AL}^\emptyset = \{(v_i, w_{i^\emptyset}) | v_i \in RV_v^+, w_{i^\emptyset} \in RV_w^+\},$$

と表現することができる．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．

$$i < j \to i^\emptyset < j^\emptyset$$

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

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.

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

$$g\left(v_i, w_{i^\emptyset}, v_j, w_{j^\emptyset}\right) = \begin{cases} 1 & if (v_i, v_j) \in CE_v \text{ and } (w_{i^\emptyset}, w_{j^\emptyset}) \in CE_w \\ 0 & otherwise \end{cases}$$

III. 提案手法

本章では，提案手法の詳細な内容を説明する.

A. 個体の定義

比較する蛋白質vと蛋白質wを表現するコンタクトマップ$CM_v = (RV_v, CE_k)$, $CM_w = (RV_v, CE_k)$とすると，本研究では，アラインメントペア集合$AL^\emptyset$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 1 では，$I = \{(\mathrm{A}_1, \mathrm{B}_1), (\mathrm{A}_3, \mathrm{B}_3), (\mathrm{A}_5, \mathrm{B}_4)\}$であり，個体は$O_1 = \mathrm{A}_1, O_2 = \mathrm{B}_1, O_3 = \mathrm{A}_3, O_4 = \mathrm{B}_3, O_5 = \mathrm{A}_5, O_6 = \mathrm{B}_4$の 6 つの構成要素から構成される.

B. 適応度の定義

個体 I が示すアラインメント集合を$AL^\emptyset$，アラインメント集合に含まれるあるアラインメントペアを$(v_k, w_{k^\emptyset})$とする．個体の適応度である大域的適応度は II で示したコスト関数fを用い，

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

と定義する.

それぞれの頂点の次数を$dig(v_k)$，$dig(w_{k^\emptyset})$とする．また，頂点v_kと頂点$w_{k^\emptyset}$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$com(v_k)$，$com(w_{k^\emptyset})$とする.

$$com(v_k) = com\left(w_{k^\emptyset}\right) = \sum_{(v_i, w_j) \in AL^\emptyset} g(v_k, w_{k^\emptyset}, v_j, w_{j^\emptyset})$$

構成要素適応度である局所的適応度は，構成要素に対応する頂点の次数と頂点が持つ共通コンタクトの数の差を次数で割った値とする.

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

C. 初期個体生成

初期個体は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．そして，残基の並びを配列要素として扱い，最適アラインメントを求め，求めたアラインメント結果を初期解とする．最初に動的計画法を用いて，最適アラインメントのスコア$D_{i,j}$を計算する．そしてスコア$D_{n,m}$から最大値を算出する経路をトレースバックすることで，アラインメント結果を求めることができる.

$$D_{i,0} \leftarrow i \times g \quad i = 0, \cdots, n$$
$$D_{0,j} \leftarrow j \times g \quad j = 0, \cdots, m$$
$$D_{i,j} \leftarrow max \begin{cases} D_{i-1,j} + g \\ D_{i,j-1} + g \\ D_{i-1,j-1} + \frac{max(S) - s_{i,j}}{\max(S)} \end{cases}$$
$$\mathrm{g} = \sum_{k=0}^{n} \sum_{l=0}^{m} \frac{max(S) - s_{k,l}}{max(s)}.$$

残基間の類似度$\mathrm{s}_{i,j}$は以下の定義式で求める.

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

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間の距離の総和である．また，α は係数である．

D. アルゴリズム

提案手法の処理手順を Algorithm1 に示す．最初に，類似度行列を作成する．次に，動的計画法で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，または突然変異を繰り返す．

状態遷移を行う場合，個体について構成要素についてその適応度$\lambda_i(= local_fitness(O_i))$を求める．次に，関数make_neighbor_indivisualを呼び出し，個体の近傍解となる複数の個体（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．突然変異の場合は，確率的な選択により，局所適応度の最も悪い要素の削除または，局所適応度が最も良くなるように要素の追加を行う．

Algorithm 2 に関数make_neighbor_individualの内容を示す．近傍解となる個体は状態遷移により作成する．最初に個体の構成要素をその局所的適応度を用いて，ルーレット選択により選択する．そして，選択した構成要素を状態遷移の方法については，次節に示す．

Algorithm 1 提案手法

入力: 蛋白質 A と蛋白質 B の座標配列データ
最大世代数 gmax
近傍個体生成数 nmax
突然変異率 mrate
出力: 最良個体

1. 蛋白質 A と蛋白質 B の座標配列データを用い，コンタクトマップと類似度行列 *S* を作成
2. 類似度行列 *S* を用い動的計画法により初期アラインメントを求め，初期個体 I に設定
3. $I_{best} = I$ /* I_{best}は最良個体 */
4. $g = 0$
5. **while** $g < gmax$ **do**
6. *I* の全ての構成要素 O_i について適応度 λ_i を算出
7. *NI* = ***make_neighbor_individual***(*I*, *nmax*)
8. *I* = ***best***(*NI*)
9. **if** ***global_fitness***（*I*） > ***global_fitness***（I_{best}） **then** $I_{best} = I$
10. 乱数値 *RAN* に 0～1 の乱数を発生
11. **if** (*RAN* < mrate)
12. アラインメント追加または削除
13. **end if**
14. *g*++
15. **end while**
16. **return** *I*

Algorithm 2 **make_neighbor_individual**

入力： 個体 *I*，近傍個体生成数 *nmax*
出力： 近傍個体集合

1. *n*=0
2. *NI* = NULL
3. **while** *n* < *nmax* **do**
4. $I_{neighbor} = I$
5. 個体の構成要素 O_k を適応度 λ_i のルーレット選択に よりひとつ選択
6. 構成要素 O_k を対象として個体 $I_{neighbor}$ を状態遷移
7. $NI = NI \cup I_{neighbor}$
8. *n*++
9. **end while**
10. **return** *NI*

E. 状態遷移

アラインメントペアの組み換えにより構成要素の状態を変異させる．例えば，構成要素O_kが状態遷移の候補として選択され，$O_k = v_k$と仮定すると，アラインメントペア$(v_k, w_{k\emptyset})$について，v_kを蛋白質vの他の頂点に変更する．逆に，$O_k = w_{k\emptyset}$と仮定すると，アラインメントペア$(v_k, w_{k\emptyset})$について，$w_{k\emptyset}$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．

オリジナルの EO や改良版 EO では状態遷移は即時移動戦略で行われるが，本研究では個体の状態遷移は最良移動戦略で行う．即時移動戦略では，アラインメントペアに交差が生じないように選択した構成要素が示す頂点をランダムに他の頂点に変更する．最良移動戦略では，変更の候補となる頂点についてその頂点に変更することで個体の大域的適応度が最良となる頂点を選択する．

IV. 評価実験

提案手法を評価するために，PDBj(Protein Data Bank Japan)に登録がある実際の蛋白質構造データ 27 件を用い評価実験を行った．性能評価では，Hengyuu Lu らの実験と同じ蛋白質データの組合せを用い，提案手法，EO，改良版 EO＋突然変異の手法を比較する．カットオフ距離，EO の世代数，改良版 EO の世代数， 個体数はそれぞれ 6.75, 100000, 1000, 100 として実行した．また，突然変異率は 5%とした．

表 1～表 3 に今回使用した蛋白質データを記す．そしてそのデータを用いて得られた結果を表 4 に示す．表 4 では，通常の EO と改良版 EO, 改良版 EO に突然変異を加えた手法それぞれの最良コンタクト数を示す． 表 4 から改良版 EO が，明らかに値が上回っているので性能が良くなっているといえる．しかし突然変異は，あまり結果は良くおらず，単純な確率だけでなく，最適な突然変異操作をさらに開発する必要がある．

表 1　データセット(Sokol テストセット)

PDB ID	Number of Residue	Number of Contact Edge
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　データセット(FLAVODOXIN-LIKE FOLD)

PDB ID	Number of Residue	Number of Contact Edge
1b00a	122	488
1dbwa	125	474
1nat	119	435
1qmpc	125	452
1b00b	122	423
4tmya	118	473

表 3　データセット(CUPREDOXINS FOLD)

PDB ID	Number of Residue	Number of Contact Edge
1b00a	122	488
1bawa	105	387
1byoa	99	355
1dpsb	154	586
1nat	119	435
1amk	250	1086
1qmpc	125	452
8tima	247	930
4tmya	118	473
1dpsc	154	585
1aw2b	254	1043
1b9ba	252	953

表 4　最良解の共通コンタクト数

蛋白質 A	蛋白質 B	改良版 EO	改良版 EO +突然変異	EO
1bpi	1knt	**175**	173	95
1bpi	2knt	180	**182**	97
1bpi	5pti	184	**185**	99
1knt	1bpi	**175**	173	97
1knt	2knt	**187**	**187**	96
1knt	5pti	**175**	173	100
1vii	1cph	**57**	54	51
2knt	5pti	179	**180**	98
3ebx	1era	**185**	180	90
3ebx	6ebx	**199**	193	93
6ebx	1era	170	**184**	82
1b00a	1dbwa	326	**333**	267
1b00a	1nat	**368**	364	267
1b00a	1qmpc	**338**	337	244
1nat	1b00b	**346**	311	251
1nat	1dbwa	339	**345**	253
1nat	4tmya	**338**	324	214
1qmpc	ab00b	**342**	313	366
1qmpc	4tmya	**366**	344	223
4tmya	1b00b	**310**	277	252
1b00a	1bawa	199	**200**	190
1b00a	1byoa	184	**185**	174
1b00a	1dpsb	**295**	289	273
1nat	1amk	285	**288**	226
1nat	1dpsb	**301**	298	285
1qmpc	2pcy	**197**	191	177
1qmpa	8tima	**304**	302	225
4tmya	1bawa	**193**	190	186
4tmya	1amk	**245**	238	210
4tmya	1dpsc	**284**	277	257
1bawa	1aw2b	**193**	189	153
1bawa	1b9ba	**209**	203	164
1bawa	1dpsb	**218**	212	191

V.　おわりに

本論文では，改良版 EO を用いた蛋白質立体構造アラインメントに対する CMO 問題の解法を提案した．提案手法の特徴は，（1）改良版 EO を用いていること，（2）初期解を動的計画法で作成，（3）アラインメントの増減を行うために突然変異を行なっていることである．提案手法を 33 組の蛋白質の組合せで評価したところ，通常の EO に比べ，33 組全ての組合せで最良解を求めることができた．また突然変異では，全体的にあまりよくなっているとは言えないが，アラインメント増減を単純な確率のみで行うのではなく，さらに工夫して突然変異を行えば改善できる可能性がある．これからの課題として，突然変異操作の工夫，多くのデータセットを使った性能評価や初期解の生成方法の工夫があげられる．

謝辞

本研究の一部は，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)，文部科学省・科学研究費補助金（若手研究(B)，課題番号：23700124）の支援により行われた．

参考文献

[1] Giuseppe Lancia and Sorin Istrail, “Protein structure comparison: Algorithms and applications,” In Mathematical Methods for Protein Structure Analysis and Design, pp. 1–33, 2003.

[2] Branden C and Tooze J, “Introduction to protein structure,” Garland Publishing, Inc., New York, USA,1999.

[3] Adam Godzik and Jeffrey Skolnick, “Flexible algorithm for direct multiple alignment of protein structures and sequences,” Computer Applications in the Biosciences, Vol. 10, No. 6, pp. 587–596, 1994.

[4] Natsumi Hara, Keiichi Tamura, and Hajime Kitakami, “Modified eo-based evolutionary algorithm for reducing crossovers of reconciliation graph,” In Proceedings of NaBIC ’10, pp. 169–176, 2010.

[5] Stefan Boettcher and Allon G. Percus, “Extremal optimization: Methods derived from co-evolution,” In Proceedings of GECCO ’99, pp. 825–832, 1999.

[6] Hengyun Lu, Genke Yang, and Lam Fat Yeung, “Extremal optimization for the protein structure alignment,” In Proceedings of BIBM'09, pp. 15–19, 2009.