Vol.2009-BIO-18 No.10
2009/9/18

# 遺伝的アルゴリズムを用いた局所的構造アラインメント

石田秀徳† 鈴木智典†† 宮崎智†
†東京理科大学大学院 薬学研究科
††東京理科大学 薬学部

タンパク質は，アミノ酸配列により立体構造が決定され，立体構造によって機能が決定されるといわれている．すなわち，同じような機能をもったタンパク質同士は類似した局所構造を共有している．したがって，類似局所構造を検出することは，タンパク質の機能を予測するために有用であるといえる．本研究では，クエリ局所構造をターゲットのタンパク質全体構造に重ね合わせ，その類似部分を検出するアラインメント法の開発を試みた．

# Local structure alignment using genetic algorithm

Hidenori Ishida† Tomonori Suzuki†† Satoru Miyazaki†
†Graduate School of Pharmaceutical Sciences, Tokyo University of Science
††Faculty of Pharmaceutical Sciences, Tokyo University of Science

Protein function is related to 3D structure, especially related to local structure of function domain. Proteins with same function share similar local structure. Thus, finding of similar local structures is a useful way to predict protein function. In this study, we attempt to develop the alignment method to search similar local structure, superimposing a query local structure on the target protein.

## 1. はじめに

遺伝子発現制御のメカニズムを解明するために，当研究室では DNA Binding Protein（DBP） の結合部位について研究を行ってきた．現在の DBP の機能分類は，配列モチーフ基づいたものが多く，DNA 結合部位の立体構造にも特化した分類は少ない．そこで, DNA 結合部位の局所立体構造に基づいて機能部位を特徴付けることを目的としたツール「FCANAL （Fast Calculable protein function ANAlyzer）」[1][2]を応用し，DNA 結合部位におけるアミノ酸の空間的出現確率と機能との関連性を調べ，機能と機能部位との関係性を明らかにした[3]．しかしながら，FCANAL は計算量を抑えるため，立体構造情報を Key と呼ばれる中心残基からの距離へ変換して計算を行っている．そのため，実際の機能部位立体構造と機能との間の直接的な関係は明らかではない．そこで，FCANAL で得られた結果について，立体構造情報（3 次元座標情報）を用いて構造を重ね合わせるといった直接的な構造比較により，機能部位構造特徴の解明を試みた．

現在，タンパク質の構造比較の方法として構造アラインメントというものがある．構造アラインメントとは，立体構造データを最適に重ね合わせるための，アミノ酸残基間の対応を決めることをいう．構造アラインメントについても配列アラインメントと同様に，大域的アラインメントと局所的アラインメントを考えることができるが，現在存在する構造アラインメントは，構造全体を重ね合わせる大域的アラインメントがほとんどで，局所的アラインメントは少ない．そこで本研究では，局所構造を対象とする局所的構造アラインメント法の開発を行った．

## 2. 材料と方法

### 2.1 構造類似性指標

立体構造の類似性の指標としてよく用いられているものが RMSD （Root Mean Square Deviation）であり，タンパク質 A と B の $i$ 番目に対応付けられた原子位置をそれぞれ $\boldsymbol{r}_i^{\boldsymbol{A}}, \boldsymbol{r}_i^{\boldsymbol{B}}$ とすると，

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} (\boldsymbol{r}_i^{\boldsymbol{A}} - \boldsymbol{R}(\boldsymbol{r}_i^{\boldsymbol{B}} + \boldsymbol{v}))^2}{\boldsymbol{N}}}$$

で表される値である．この値が小さいほど，2 つのタンパク質がよく重なることを示している．しかし，最もよい重ね合わせ（最小 RMSD）を得るには，回転行列 $\boldsymbol{R}$ と並進ベクトル $\boldsymbol{v}$ を計算したうえで，RMSD を求める必要がある．これらは計算量がかか

Vol.2009-BIO-18 No.10
2009/9/18

るため，本研究では RMSD と似た指標である dRMSD（distance RMSD）も構造類似性指標として用いた．dRMSD は，タンパク質 A, B の対応する原子ペアの距離をそれぞれ $\boldsymbol{d}_{ij}^{A}, \boldsymbol{d}_{ij}^{B}$ とすると，

$$\mathrm{dRMSD} = \sqrt{\frac{2\sum\sum_{i<j}(\boldsymbol{d}_{ij}^{A} - \boldsymbol{d}_{ij}^{B})^2}{N(N-1)}}$$

で与えられる値である[4]．dRMSD も RMSD と同様に値が小さいほど 2 つの立体構造がよく重なることを意味しているが，最適な回転と並進を計算する必要がないという特徴があり，計算が容易である．

### 2.2 構造アラインメント

構造アラインメントとは，タンパク質間で構造的に対応するアミノ酸残基対を見つけることをいう．配列アラインメントと同様に，構造アラインメントにも，大域的アラインメントと局所的アラインメントが存在する．大域的アラインメントでは，タンパク質構造全体を対象に，局所的アラインメントでは，よく似た部分のみを対象にアラインメントを行う．先に述べたように本研究では，機能部位の局所的構造について解析を行うため，局所的構造アラインメントについて考える．

一般的に構造的に対応する残基対を見つけるためには，考えられるパターン全てについて重ね合わせを行い，構造類似性を評価する必要があると考えられる．しかしながら，図 1 のように全パターンについて残基の対応を考えるとすると，計算量が膨大になるため，現実的には不可能であるという問題がある．局所的アラインメントでは大域的アラインメントに比べ，アラインメントする局所構造を選択しなければならないため，さらに問題が難しくなる．

そこで，対応する残基対（最適アラインメント）を実現可能時間内に検索する手法として，本研究では遺伝的アルゴリズムを採用し構造アラインメントを行った．

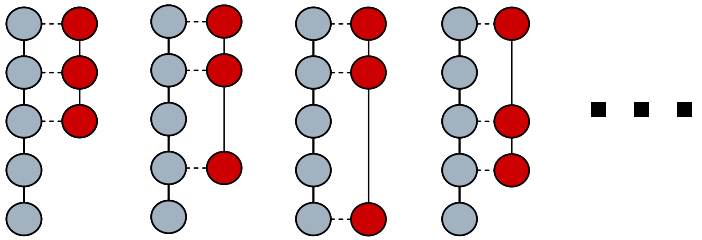

図 1　全パターンを検索するイメージ図
各アラインメントの左側（青）は，対象タンパク質の各 $C_\alpha$ 原子を表しており，右側（赤）は，クエリ構造の $C_\alpha$ 原子を表している．

### 2.3 遺伝的アルゴリズム

遺伝的アルゴリズムとは，生物集団が遺伝的変化により環境に適応していくことを模倣して，最適な解を探索していく手法である．解を個体と考え，生物における突然変異や交叉，環境適応度に応じた選択（淘汰）をシミュレーションし，最適な個体（解）を探索する．このアルゴリズムは，前述したような組み合わせが膨大になるような問題に対して近似解を高速に求めることが出来る手法の一つである．本研究ではアラインメントを個体と考え，遺伝的アルゴリズムを用いて最適なアラインメントを探索する．遺伝的アルゴリズムで解探索のためによく使われる操作には以下のようなものがある．

**(1) 選択**
適応度の高い個体が個体群に広がるように優れた個体を残し，劣った個体を淘汰する．

**(2) 突然変異**
個体をランダムに変化させる．新しい特徴をもつ個体を作成することが出来る．

**(3) 交叉**
親個体から子個体を生成する．優れた個体同士のよい部分をもつ子個体の作成が期待できる．

### 2.4 遺伝的アルゴリズムを用いた局所的構造アラインメント法

本研究で行った局所的構造アラインメントの流れは以下の通りである．

#### 2.4.1 クエリ構造と被検索構造の定義

まず，FCANAL で得られた機能部位局所構造の立体構造をクエリ構造として定義し

た．FCANAL では活性中心残基の $C_\alpha$ 原子から一定距離内の $C_\alpha$ 原子を機能部位局所構造と定義しているため，クエリ構造も同様のものとなる．次に，このクエリ構造と類似の構造を探す対象タンパク質(被検索構造)を決めた．構造比較には FCANAL 同様，計算を簡略化するために各残基の $C_\alpha$ 原子のみを利用した．

#### 2.4.2 初期集団の作成

はじめに，被検索構造からランダムに局所構造を抽出・作成した．局所構造の抽出・作成方法は以下の通りである．まず，$C_\alpha$ 原子を一個ランダムに選択し，その近隣に存在する $C_\alpha$ 原子をクエリ構造の $C_\alpha$ 原子数と同数選択し，それを局所構造とする．これをあらかじめ決めた回数（個体数）繰り返し，初期集団（第一世代）とした．

#### 2.4.3 個体の評価・選択

こうしてできた第一世代の集団（個体群）に対して，それぞれ評価値を求めた．評価値には，クエリ構造との構造類似性指標（RMSD または dRMSD）を用いた．

各個体（局所構造）の構造類似性指標を求め，上位 50%の個体を選択し，残り 50%を淘汰した．その後，1 世代あたりの個体数を維持するために選択した個体を 2 倍に増やした．

#### 2.4.4 次世代の作成

遺伝的アルゴリズムでは交叉と変異（突然変異）という 2 つの操作を行い，次世代の集団を作成する．しかし，今回の局所立体構造を個体とする遺伝的アルゴリズムでは，交叉を行うことが困難であるため変異のみのアルゴリズムを採用した．変異はアラインメントから 2 残基選択し，対応関係を入れ替えるという方法で行った．

#### 2.4.5 最終的なアラインメント

2.3.3～2.3.4 の操作をあらかじめ決めた回数（終了世代数）繰り返し，最終世代において最も高い評価の個体を最終的な解（類似局所構造）とした．

### 2.5 アラインメント法の検証

前述した構造アラインメント法を用いてアラインメントを行うプログラムを作成し，プリンリプレッサー（PDBID: 1BDH, 1BDI[5]）チェーン A の立体構造情報を用いて検証した．この 2 つのタンパク質はアミノ酸配列が 1 残基異なるだけで立体構造がほぼ同じであり，クエリ構造に相当する局所構造があることが分かっている．これらの構造データはタンパク質構造データバンク PDBj（Protein Data Bank Japan）[6]より取得した．

まず，クエリ構造を 1BDH の 17Thr から半径 9Å以内にある $C_\alpha$ 原子と定義した．これは FCANAL を用いた解析によって推定された機能部位である．被検索構造には，1BDI のチェーン A を用いた．

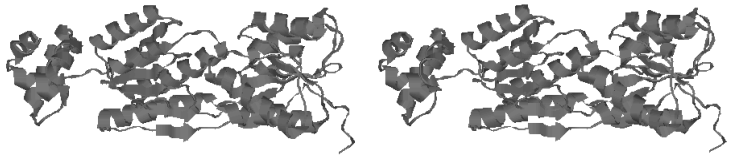

図 2　使用したタンパク質の立体構造
（左）PDBID: 1BDH　（右）PDBID: 1BDI

## 3. 結果と考察

### 3.1 検証結果と考察

表 1 のパラメータを用いて，アラインメントを実行した．アラインメントは類似性指標に RMSD および dRMSD を用いた 2 通りで行った．その結果を以下に示す．各世代において作成されたアラインメントの RMSD および dRMSD は，それぞれ図 3 および図 4 のように推移した．どちらの指標を用いた場合でも，世代数を重ねるにつれて，類似性指標が小さいアラインメントが得られることが確認できた．これにより，遺伝的アルゴリズムを用いて局所類似構造のアラインメントができることがわかった．

Vol.2009-BIO-18 No.10
2009/9/18

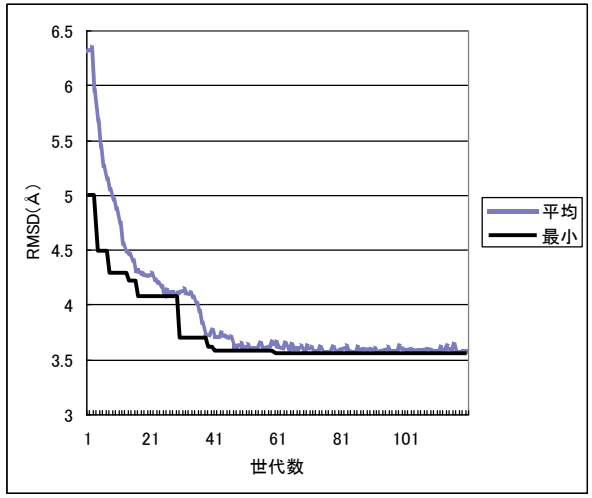


図 3　類似性指標に RMSD を用いた場合の RMSD の推移
平均は世代内での平均 RMSD を，最小は最小 RMSD を表している．

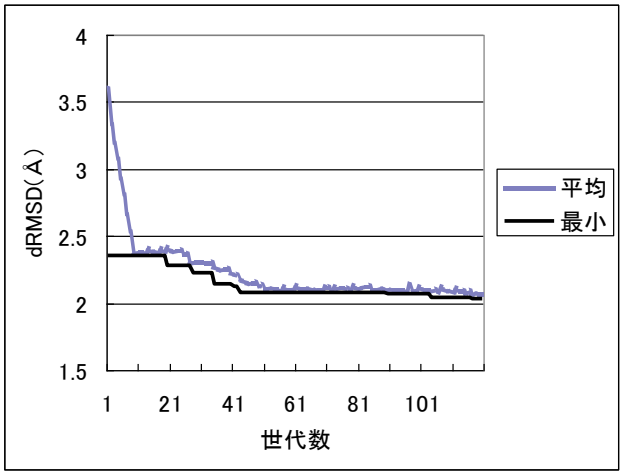


図 4　類似性指標に dRMSD を用いた場合の dRMSD の推移

表 1　本研究で用いた遺伝的アルゴリズムパラメータの値

| パラメータ | 値 |
|---|---|
| 突然変異率 | 5% |
| 世代あたりの個体数 | 200 個体 |
| 終了世代数 | 120 世代 |

また，構造類似性指標として，dRMSD を用いた場合，通常の RMSD を構造類似性指標とした場合に比べて，計算時間を 1/30 以下に短縮することができた．

今回は，類似した局所構造をもつタンパク質に対して，構造アラインメントを実行したが，本手法を応用することで，クエリ局所構造に類似する構造を機能未知のタンパク質中から検索することが期待される．しかしながら，クエリ局所構造として多くのタンパク質が共通に持つ構造（$\alpha$ ヘリックスや $\beta$ シート構造など）を用いると，被検索構造中に類似構造が多く存在することになり，検索が困難になるという問題点があり，今後改善していきたい．

## 4.　参考文献

1 T.Asaoka, T.Ando, T.Meguro and I.Yamato, CBIJ, 2, 96-113 (2003)
2 A.Suzuki, T.Ando, I.Yamato and S.Miyazaki, CBIJ, 3, 39-55 (2005)
3 Y.Sakatsuji, R.Okihara, I,Yamato and S.Miyazaki, 情報処理学会研究報告, Vol.2007, No.89, 41-48 (2007)
4 日本バイオインフォマティクス学会編，バイオインフォマティクス事典，共立出版 (2006)
5 M.A.Schumacher, K.Y.Choi, H.Zalkin and R.G.Brennan, Science, 266(5186), 763-770 (1994)
6 PDBj, http;//www.pdbj.org/