実数値GAによるタンパク質立体構造の2層比較

朴 聖俊[†] 高田彰二[††] 山村雅幸[†]

爆発的に増加するタンパク質立体構造を比較することは構造—機能相関の解析にきわめて重要である．既存の立体構造比較手法はタンパク質全体を剛体として扱う．しかし，進化的に新しい機能を獲得する際にタンパク質構造は部分的特異的に変形を受けるため，剛体としての取扱いには限界がある．本論文では機能進化過程において，構造変形を受けにくいビルディングブロックと構造変形が顕著なループ部分が存在することを考慮に入れた立体構造比較手法を開発する．提案手法は部分構造比較と全体構造比較を2層で並列探索し，遺伝的アルゴリズムの集団探索性能を活用してタンパク質の機能進化における構造変形の柔軟性を可視化する．2層比較の基本的なアイデアと実装について説明したうえで探索アルゴリズムと評価関数の特徴と性能について述べ，構造—機能相関の解析ツールとしての有効性を示す．

Two-layered Comparison of Protein Structures by Real-coded GA

Sung-Joon Park,[†] Shoji Takada[††] and Masayuki Yamamura[†]

Comparing protein tertiary structures that are explosively increasing is indispensable to the investigation into protein structure-function relationship. Methods hitherto published, however, treat proteins as rigid-bodies and thus are not able to capture the function-related structural variability of proteins, which appear through evolution. In this paper, we develop a protein structure comparison tool that emphasizes physico-chemical rearrangement of local fragments as building blocks of global structures. The proposed tool optimizes local fragment alignment and global superposition simultaneously. Using the population search ability of Genetic Algorithm, this tool shows the protein flexibility. We describe first the approach and the implementation. To address the large-scaled analysis of protein structure-function relationship, we show the effectiveness of the global search ability and fitness functions.

1. はじめに

タンパク質は20種類のアミノ酸配列（一次構造）がペプチド結合によってつながった高分子である．3次元空間で折り畳まれた立体構造（三次構造）は一次構造の変異に対してロバストであり[1),2)]，生化学的機能発現と密接に関係する．ポストゲノム時代において，立体構造と機能関係の曖昧な Twilight Zone[3)~5)] を明らかにし，タンパク質の構造—機能相関を議論することは自然界を理解するうえできわめて重要である．

本論文で議論するタンパク質立体構造比較（Protein Structure Alignment，PSA）は，一次構造比較では不可能な構造—機能相関の手掛かりを発見するための出発点である．また，構造モデリング，構造予測，創薬などのタンパク質構造関連研究に欠かせない基本的な手法である[6)]．現在，多様なアイデアに基づく PSA 手法が提案されている（総説は文献 7)〜9) を参照）．

既存手法[10)~17)] はタンパク質を剛体として比較する．しかし，新しい機能獲得などに際してタンパク質は進化的に部分構造の変形を受けており[18)~20)]，それを考慮して部分構造と全体構造との保存・変異関係[4),21),22)] を発見することは困難である．近年，部分構造の進化的多様性を考慮する PSA 手法[23),24)] が構造—機能相関解析の新しい方法として注目されているが，部分構造の並びと全体構造の重ね合わせとの関係が観察できない問題点がある．

本研究では，ビルディングブロックとしての「硬い」タンパク質の部分構造と，それらをつなぐ柔らかいループの部分構造が，どう保存され，どう再構築されているかという疑問を解決する PSA 法を設計する．

提案手法の基本的なアイデア（以下，「2層比較」と呼ぶ）は，部分構造比較（Local Fragment Alignment，

† 東京工業大学大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
†† 神戸大学理学部化学科
Department of Chemistry, Faculty of Science, Kobe University

LFA）と全体構造重ね合わせ（Global Superposition Alignment，GSA）を 2 層で並列に探索すると同時に，GSA の探索に LFA 類似度情報を参照することである．2 層比較は生物の進化様子を模倣する遺伝的アルゴリズム（Genetic Algorithm，GA）によって実装される．GA は集団探索性能を活用する近似最適化手法であり[25]，複数の評価関数を扱うことができる[26]．ここでは，LFA 類似度評価関数と GSA 類似度評価関数を最適化し，GA の集団探索挙動を観察することで部分構造と全体構造との保存・変異関係を可視化する．

本論文は以下のように構成される．2 章で PSA 法と GA，既存手法とその問題点について説明を行い，3 章で GA による 2 層比較を設計する．4 章では提案手法の性質確認と性能比較のための実験と考察を行う．5 章でまとめを行う．

2. 従 来 研 究

2.1 タンパク質立体構造比較（PSA）

タンパク質立体構造比較（Protein Structure Alignment，PSA）は，部分構造の局所的類似度に注目する部分構造アラインメント（Local Fragment Alignment，LFA）と中心炭素原子 Cα で表現される全体構造骨格の大域的類似度に注目する全体構造重ね合わせ（Global Superposition Alignment，GSA）に分類される．

図 1 に示すように LFA は 2 種類に区別される．ここでは，対応部分構造の出現順が一次構造の N 末から C 末への方向に従う場合を「順序 LFA」（図 1 (a)），従わない場合を「非順序 LFA」（図 1 (b)）と呼ぶことにする．GSA の対応原子の出現順は必ず一次構造の並び方に従う（図 1 (c)）．以下，GSA は「順序 GSA」と同意である．

PSA の目標は問合せ構造（$\mathcal{Q}$）と参照構造（$\mathcal{R}$）における等価原子ペア数の最大化である．ここで，Cα 原子数の少ない構造を $\mathcal{Q}$，他方を $\mathcal{R}$ とする．

$\mathcal{Q}$ と $\mathcal{R}$ の重ね合わせは最小二乗法（Least Square Fitting，LSQ-fitting）[27] によって行われる．LSQ-fitting は，$\mathcal{Q}$ と $\mathcal{R}$ における等価原子ペアの集合 $\mathcal{P}$ の最適な合同変換 $\mathcal{P}^{\mathrm{new}} = \mathcal{P}^{\mathrm{current}} \times \mathcal{M} + \mathcal{T}$ を行って，原子ペアの平均二乗距離誤差（Root Mean Squared Deviation，RMSD）を最小化する．ここで $\mathcal{M}$ と $\mathcal{T}$ は，それぞれ回転行列，並行移動ベクトルである．$\mathcal{P}$ は様々なスコア関数を組み込んだ動的計画法（Dynamic Programming，DP）[28),29)] などによって決定される．

2.2 遺伝的アルゴリズム（GA）

生物の進化過程を真似る遺伝的アルゴリズム（Genetic Algorithm，GA）は，様々な最適化問題に効果的な確率・近似・非決定性探索手法である．GA は，最適化すべき問題の変数（表現型）をビットストリングなどの遺伝子型へコーディングし，潜在的な解候補（集団）に対して選択・交叉・突然変異の操作を繰り返し行う．

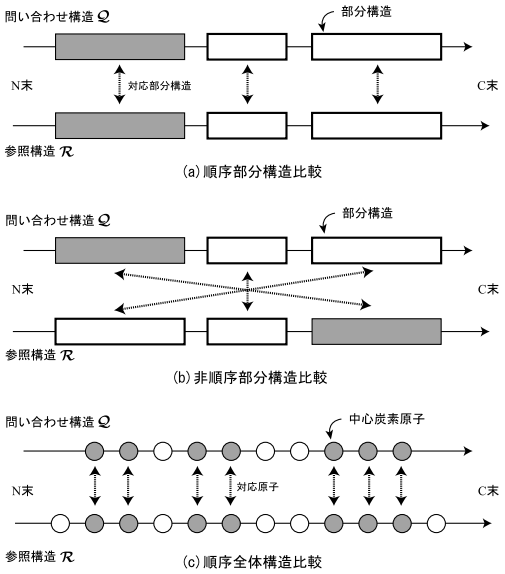

図 1 タンパク質立体構造比較の分類：一次構造の N 末から C 末への方向に従う比較は順序，従わない比較は非順序である．Cα 原子数の少ない構造を問合せ構造（$\mathcal{Q}$），他方を参照構造（$\mathcal{R}$）とする

Fig. 1 Schematic representation of alignment patterns of protein tertiary structures.

有効な GA 設計は，致死遺伝子を生成しないコード化方法，親集団の統計量を継承する交叉方法，個体群を最適解へシフトさせる世代交代モデル設計が重要である[30]．単純 GA [25] はランダム性に基づく交叉方法や無条件の世代交代などのため，大域探索が行えず局所解に陥りやすいことが知られている[31]．

遺伝子型として実数値ベクトルを用いる実数値 GA は，ビットストリング型に比べて関数最適化問題に効果的である[32]．実数値 GA における単峰性正規分布交叉（Unimodal Normal Distribution Crossover，UNDX）[33] は多峰性高次元関数の最適化に優れている[34]．

2.3 既存手法とその問題点

タンパク質における部分構造は全体構造を作り上げるビルディングブロックであり[18),35)]，部分構造の挿入・欠失は新しい機能のタンパク質を生み出す[4),19),20)]．このようなタンパク質の柔軟性[36] は LFA と GSA の関係を観察することによって発見することができる．

表 1 既存タンパク質立体構造比較ツール
Table 1 Difference of algorithmic properties in alignment methods.

ツール	対象構造比較種類 類似度評価 最適化手法
SARF2 [15]	順序・非順序 LFA Cα 距離，部分構造の結合角 反復改善法
KENOBI [17]	順序・非順序 LFA 二次構造 組合せ最適化 GA
DALI [14]	順序 LFA →GSA Cα 距離 距離行列 → モンテカルロ法
CE [12]	順序 LFA →GSA Cα 距離 組合せ拡張法 → 反復改善法
FlexProt [23]	順序 LFA →GSA Cα 距離 非循環有向グラフ
SAP [13]	GSA Cα 距離，残基の埋もれ度，二次構造 二重動的計画法
GA_FIT [10]	GSA Cα 距離 単純 GA
FROG （提案手法）	Cα 距離，Cβ 距離，残基の埋もれ度 順序・非順序 LFA，順序 LFA→GSA 実数値 GA

しかしながら，**表 1** のような既存手法では LFA—GSA 関係を考察することがきわめて困難である．

SARF2 [15] は順序・非順序 LFA が行えるが，CE [12] のように GSA への拡張ができない．SAP [13] は GSA 類似度の低いタンパク質ペアにおける順序・非順序 LFA が発見できない．GA を採用する手法[10),17),37)] は単点探索に基づく既存手法の初期対応原子への依存性[37)] を解決している．しかし，類似部分構造のパターン[17)] と回転角度[10)] をビットストリングで表現するなど，古典的 GA の枠組みにとどまっている．

既存手法が注目するタンパク質の特性（たとえば，局所的・大域的骨格の幾何学的類似度，物理化学的性質も含めた類似度，二次構造の保存性など）は互いに異なる．さらに，PSA は NP 困難な問題であるため[38)]，多様な発見的手法が用いられる．したがって，PSA 手法における絶対的な評価関数と比較結果というものは存在しない[37),39),40)]．

このような状況で，既存手法からの LFA 類似度と GSA 類似度を用いてタンパク質の柔軟性を考察することは非常に困難であり，新しい PSA 手法の開発が望まれる．

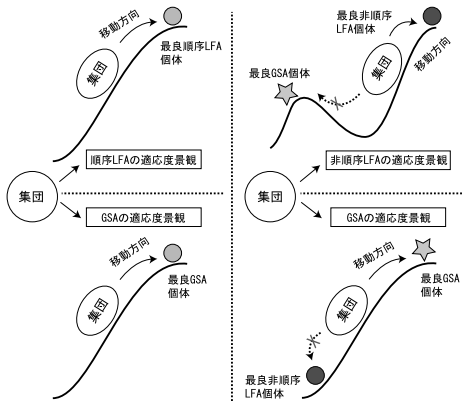

図 2 GA による 2 層比較の基本的な考え方：GA の集団を LFA 類似度評価関数と GSA 類似度評価関数の適応度景観へマッピングし，最良 LFA 個体が順序の場合 (a) と非順序の場合 (b) における集団の移動方向を観察する
Fig. 2 Concept of two-layered comparison of protein structures by GA.

3. 提 案 手 法

3.1 基本的な考え方

2 つのタンパク質における順序 LFA の情報を利用して GSA を最適化する．最適 LFA 結果が非順序の場合においても GSA 最適化は順序 LFA の情報を参照する．したがって，最適 GSA における LFA 結果と最適 LFA 結果を観察することによって部分構造と全体構造の保存・変異関係が明らかになる．

このような 2 層比較を実装するには，効果的な LFA 手法と GSA 手法による並列探索と多点探索が必要となる．そこで，GA を用いて**図 2** のような最適化を行う「FROG」（Fitted Rotation and Orientation of protein structures by real-coded Ga）を設計する☆．

GA の集団を LFA 類似度評価関数と GSA 類似度評価関数の適応度景観へマッピングする．ある探索ステップにおける適応度の最も良い LFA 個体が図 2 (a) のような順序 LFA の場合，2 つの適応度景観における集団の移動方向は同じになる．一方，最良 LFA 個体が非順序 LFA の場合の GSA 適応度は図 2 (b) のように低くなる．したがって，集団の移動方向は LFA 層と GSA 層において大きく異なる．

3.2 アルゴリズムの流れ

3.2.1 コ ー ド 化

個体の遺伝子型は合同変換における $\mathcal{M}$ と $\mathcal{T}$ の実

☆ 実験データ，カラー図版，プログラムなどの補足資料は文献 41) からダウンロード．

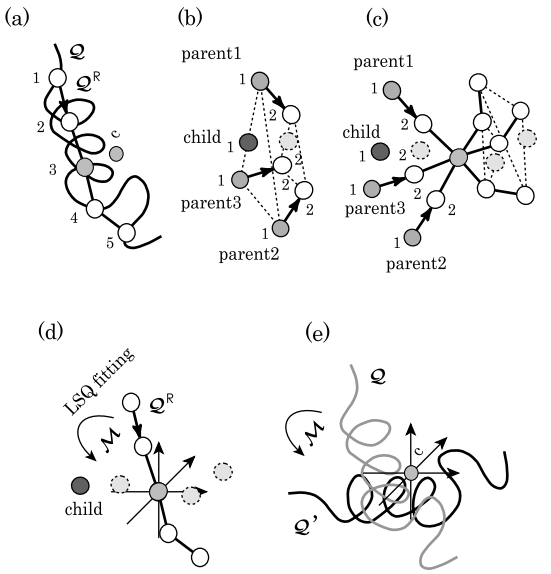

図 3 UNDX と LSQ-fitting を併用する提案手法の回転交叉：$\mathcal{Q}$ は問合せ構造，$\mathcal{Q}^{\mathrm{R}}$ は $\mathcal{Q}$ の代表構造，c は $\mathcal{Q}$ の重心であり，$\mathcal{M}$ は回転行列である．親 3 個体の初期原子座標を用いた UNDX によって子個体の初期原子座標が決定される (b)．親と子の初期原子座標の距離比例関係から子の $\mathcal{Q}^{\mathrm{R}}$ における残りの原子座標が決定される (c)．初期 $\mathcal{Q}^{\mathrm{R}}$ (a) が子の $\mathcal{Q}^{\mathrm{R}}$ へ回転する $\mathcal{M}$ は LSQ-fitting によって求まる (d)

Fig. 3 Novel crossover method for structure rotation using UNDX and LSQ-fitting.

数値ベクトルでコーディングされる．FROG は LFA 層において対応部分構造数を最大化する $\mathcal{M}$ を，GSA 層において対応原子数を最大化する合同変換を探索する．

3.2.2 初 期 化

比較される 2 つのタンパク質の座標系を揃え，制限された空間上に初期集団を生成する．

Step 1. まず，$\mathcal{Q}$ の重心 c と $\mathcal{R}$ の重心 c' を Cartesian 軸の原点へ平行移動させる．

Step 2. $\mathcal{Q}$ における n^{rep} 個の Cα を選択して代表構造 $\mathcal{Q}^{\mathrm{R}}$ を作る（**図 3** (a)，便宜上連続な $n^{\mathrm{rep}} = 5$ とした）．$\mathcal{Q}^{\mathrm{R}}$ は回転交叉の計算時間を短縮するために用いる $\mathcal{Q}$ の粗視化構造である．

Step 3. $\mathcal{Q}$ のランダムな合同変換を表す n^{pop} 個の個体群を生成する（初期集団）．ここで，初期集団の $\mathcal{T}$ は $\mathcal{R}$ の回転半径球の内側に生成した c の位置ベクトルである．

3.2.3 最適化：選択と交叉方法

子集団の $\mathcal{T}$ は UNDX によって，$\mathcal{M}$ は UNDX と LSQ-fitting を併用する「回転交叉」によって生成される．

式 (1) によって与えられる UNDX は親 3 個体のガウシアン空間の近傍に n^{undx} 回の交叉で $n^{\mathrm{undx}} \times 2$ 個体からなる子集団を生成する．子集団と親 3 個体からなる集団を家族と呼ぶ．

$$
\begin{aligned}
\overrightarrow{C}_1 &= \vec{m} + u_1\vec{e}_1 + \sum_{k=2}^{n} u_k\vec{e}_k \\
\overrightarrow{C}_2 &= \vec{m} - u_1\vec{e}_1 - \sum_{k=2}^{n} u_k\vec{e}_k \\
\vec{m} &= (\overrightarrow{P}_1 + \overrightarrow{P}_2)/2 \\
u_1 &\simeq N(0, \sigma_1^2) \\
u_k &\simeq N(0, \sigma_2^2) \quad (k = 2, \ldots, n) \\
\sigma_1 &= \alpha^{\mathrm{undx}} d_1, \quad \sigma_2 = (\beta^{\mathrm{undx}} d_2)/\sqrt{n} \\
\vec{e}_1 &= (\overrightarrow{P}_2 - \overrightarrow{P}_1)/|\overrightarrow{P}_2 - \overrightarrow{P}_1| \\
\vec{e}_i &\perp \vec{e}_j \ \ (i, j = 1, \ldots, n)(i \neq j).
\end{aligned} \tag{1}
$$

ここで n は次元数，$\overrightarrow{P}_i$ と $\overrightarrow{C}_i$ は第 i 番目の親と子のベクトル，u_i は正規分布に従う乱数，d_1 は $\overrightarrow{P}_1$ と $\overrightarrow{P}_2$ 間の距離，d_2 は $\overrightarrow{P}_1$ と $\overrightarrow{P}_2$ を結ぶ軸と $\overrightarrow{P}_3$ との距離，α^{undx} と β^{undx} は定数である（理論的考察は文献 34) を参照）．

親 3 個体の $\mathcal{T}$ を式 (1) の $\overrightarrow{P}$ とした UNDX によって子の $\mathcal{T}$ が生成される．次のステップに従う回転交叉には $\mathcal{Q}^{\mathrm{R}}$ が用いられる．

Step 1. 親の第 1 番目の Cα 原子座標を式 (1) の $\overrightarrow{P}$ とした UNDX を実行し，子の第 1 番目の Cα 原子座標（図 3 (b) の child$_1$）を決める．

Step 2. 生成された子の第 1 番目の Cα 原子と親の第 1 番目の Cα 原子との距離比例関係を計算する．

Step 3. 図 3 (c) のように，距離比例関係を用いて子の $\mathcal{Q}^{\mathrm{R}}$ における残りの原子座標を決める．

Step 4. 初期 $\mathcal{Q}^{\mathrm{R}}$ が子の $\mathcal{Q}^{\mathrm{R}}$ へ回転する $\mathcal{M}$ を LSQ-fitting を用いて計算する（図 3 (d)）．全体構造 $\mathcal{Q}$ は $\mathcal{M}$ によって回転され，評価される（図 3 (e)）．

3.2.4 最適化：世代交代モデル

GSA の最適探索が LFA 類似度情報を参照することは，家族における最良 LFA 個体の $\mathcal{M}$ を集団へ戻すことによって実現される．

Step 1. 3.3 節で定義する 2 つの関数を用いて家族の適応度を評価し，最良 LFA 個体と最良 GSA 個体を決める．

Step 2. GSA 類似度に比例するルーレット選択[25] を行ってランダムな 1 個体を選択する．

Step 3. 最良 GSA 個体，ランダム選択による個体と親 2 個体を入れ替える．

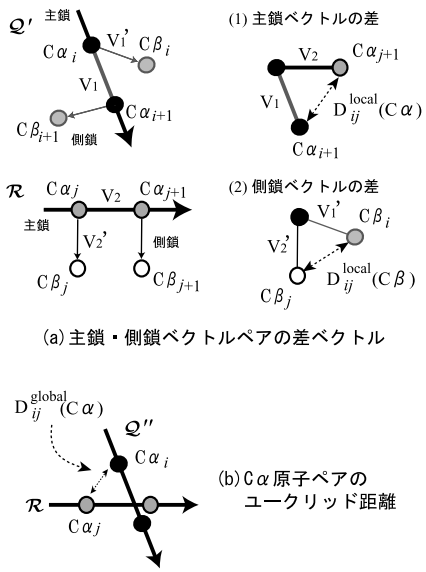

図 4 主鎖と側鎖，Cα 原子における対応ペア距離類似度の計算方法：主鎖ペアと側鎖ペアは長さ 1 の方向ベクトルの差ベクトルを用いる (a)．回転した $\mathcal{Q}'$ と $\mathcal{R}$ における V_1 と V_2 の差ベクトルの大きさは主鎖の類似度評価に，V'_1 と V'_2 の差ベクトルの大きさは側鎖の類似度評価に用いられる．原子ペアはユークリッド距離を類似度として用いる (b)．平行移動した $\mathcal{Q}''$ が $\mathcal{R}$ へ重なったときの原子間の距離を評価する

Fig. 4 Measurements of backbone distance, Cβ-Cβ distance, and Cα-Cα distance.

Step 4. 最良 LFA 個体の $\mathcal{M}$ だけを親 1 個体の $\mathcal{M}$ と置き換える．

3.3 評価関数

GSA 類似度に合同変換による全体構造の重ね合わせ，残基の溶媒接触可能性[42]，ギャップペナルティが含まれる．LFA 類似度には主鎖（$\mathrm{C}\alpha_i \to \mathrm{C}\alpha_{i+1}$）と側鎖（$\mathrm{C}\alpha_i \to \mathrm{C}\beta_i$）☆の方向ベクトル，残基の溶媒接触可能性が反映される．

以下で説明する方法を使って $\mathcal{Q}$ と $\mathcal{R}$ における全組合せの主鎖ペア，側鎖ペア，原子ペアの距離類似度を計算する．閾値判断による距離類似度の 2 値化行列を構築して対応部分構造と対応原子を決定し，評価する．

3.3.1 対応ペアの類似度計算方法

ある個体 Indi の $\mathcal{M}'$ によって回転した $\mathcal{Q}$ を $\mathcal{Q}'$ とする．$\mathcal{Q}'$ の $\mathrm{C}\alpha_i \to \mathrm{C}\alpha_{i+1}$ と $\mathrm{C}\alpha_i \to \mathrm{C}\beta_i$ を表す長さ 1 の方向ベクトルを，それぞれ $\mathrm{V}(\mathcal{Q}'_{\mathrm{C}\alpha_{i,i+1}})$，$\mathrm{V}(\mathcal{Q}'_{\mathrm{C}\beta_i})$ とする（**図 4** (a) における V_1 と V'_1）．同様に，$\mathcal{R}$ に対して $\mathrm{V}(\mathcal{R}_{\mathrm{C}\alpha_{j,j+1}})$，$\mathrm{V}(\mathcal{R}_{\mathrm{C}\beta_j})$ とする（図 4 (a) における V_2 と V'_2）．主鎖ペア $\{i,j\}$ と側鎖ペア $\{i',j'\}$ の差ベクトルの大きさをそれぞれ $\mathrm{D}^{\mathrm{local}}_{ij}(\mathrm{C}\alpha)$（図 4 (a-1)），$\mathrm{D}^{\mathrm{local}}_{i'j'}(\mathrm{C}\beta)$（図 4 (a-2)）とし，下式を用いて計算する．

$$\begin{aligned}\mathrm{D}^{\mathrm{local}}_{ij}(\mathrm{C}\alpha) &= ||\mathrm{V}(\mathcal{Q}'_{\mathrm{C}\alpha_{i,i+1}}) - \mathrm{V}(\mathcal{R}_{\mathrm{C}\alpha_{j,j+1}})|| \\ &\quad (i=1,\ldots,l_{\mathrm{que}}-1,\ j=1,\ldots,l_{\mathrm{ref}}-1) \\ \mathrm{D}^{\mathrm{local}}_{i'j'}(\mathrm{C}\beta) &= ||\mathrm{V}(\mathcal{Q}'_{\mathrm{C}\beta_{i'}}) - \mathrm{V}(\mathcal{R}_{\mathrm{C}\beta_{j'}})|| \\ &\quad (i'=1,\ldots,l_{\mathrm{que}},\ j'=1,\ldots,l_{\mathrm{ref}}). \end{aligned} \quad (2)$$

ここで l_{que} と l_{ref} はそれぞれ $\mathcal{Q}$ と $\mathcal{R}$ の長さ（すなわち Cα 原子数，$l_{\mathrm{que}} \le l_{\mathrm{ref}}$）である．

次に，Indi の $\mathcal{T}'$ によって $\mathcal{R}$ へ平行移動させた $\mathcal{Q}'$ を $\mathcal{Q}''$ とする（図 4 (b)）．$\mathcal{Q}''$ の $\mathrm{C}\alpha_i$ と $\mathcal{R}$ の $\mathrm{C}\alpha_j$ の位置ベクトルをそれぞれ $\mathrm{V}(\mathcal{Q}''_{\mathrm{C}\alpha_i})$，$\mathrm{V}(\mathcal{R}_{\mathrm{C}\alpha_j})$ とする．原子ペア $\{i,j\}$ の差ベクトルの大きさ $\mathrm{D}^{\mathrm{global}}_{ij}$ は下式を用いて計算する．

$$\begin{aligned}\mathrm{D}^{\mathrm{global}}_{ij}(\mathrm{C}\alpha) &= ||\mathrm{V}(\mathcal{Q}''_{\mathrm{C}\alpha_i}) - \mathrm{V}(\mathcal{R}_{\mathrm{C}\alpha_j})|| \\ &\quad (i=1,\ldots,l_{\mathrm{que}},\ j=1,\ldots,l_{\mathrm{ref}}). \end{aligned} \quad (3)$$

3.3.2 対応部分構造集合と対応原子集合の決定方法

$(l_{\mathrm{que}}-1)\times(l_{\mathrm{ref}}-1)$ 行列の要素 w_{ij} を次のように決め，Smith–Waterman の DP [29] を用いてサブ最適パス集合を求める．

$$w_{ij} = \begin{cases} 1, & \mathrm{D}^{\mathrm{local}}_{ij}(\mathrm{C}\alpha) \le \mathrm{cutoff}_1 \\ & \cap\ \mathrm{D}^{\mathrm{local}}_{ij}(\mathrm{C}\beta) \le \mathrm{cutoff}_2 \\ & \cap\ \mathrm{D}^{\mathrm{local}}_{i+1j+1}(\mathrm{C}\beta) \le \mathrm{cutoff}_2 \\ 0, & \text{otherwise} \end{cases}$$
$$(i=1,\ldots,l_{\mathrm{que}}-1,\ j=1,\ldots,l_{\mathrm{ref}}-1). \quad (4)$$

ここで cutoff_1 と cutoff_2 は閾値である．

サブ最適パスは上記行列の左上から右下へ n^{cons} 個以上続く値 1 の非重複パスであり，主鎖と側鎖が類似する対応部分構造を表す．以下，サブ最適パス集合の要素（つまり，$w_{ij}=1$ のベクトルペア）を $\mathrm{EQ}^{\mathrm{local}}_k (k=1,\ldots,m)$ とする．

また，$l_{\mathrm{que}}\times l_{\mathrm{ref}}$ 行列の要素 $w'_{i'j'}$ を次のように決め，Needleman-Wunch の DP [28] を用いて絶対最適パスを求める．

$$w'_{i'j'} = \begin{cases} 1, & \mathrm{D}^{\mathrm{global}}_{i'j'}(\mathrm{C}\alpha) \le \mathrm{cutoff}_3 \\ 0, & \text{otherwise} \end{cases}$$
$$(i'=1,\ldots,l_{\mathrm{que}},\ j'=1,\ldots,l_{\mathrm{ref}}). \quad (5)$$

ここで cutoff_3 は閾値である．

絶対最適パスは主鎖の重なりが類似する対応原子の集合を指す．以下，絶対最適パスの構成要素を

☆ Cβ が存在しないアミノ酸は文献 43) に従って C-Cα-N による擬似 Cβ を用いる．擬似 Cβ は SA 計算に影響しない．

$\mathrm{EQ}_{k'}^{\mathrm{global}}(k'=1,\ldots,m')$ とする．

3.3.3 溶媒接触可能性

溶媒接触表面とはある残基に存在する全原子の Van der Waals 表面を溶媒（水分子，半径 1.4Å）が移動したときの軌跡である．溶媒接触可能性（Solvent Accessibility，SA）とは，ある残基の溶媒接触表面に n^{water} 個の水分子をプロットしたとき，他残基の溶媒接触表面に入らない水分子の割合である（$0 \leq \mathrm{SA} \leq 1.0$）．

SA はタンパク質における各残基の疎水性指標と埋もれ度を表すため，タンパク質構造比較におけるアミノ酸の環境プロファイルとして使われている[13),44)]．ここでは MOLMOL パッケージ[45)] を用いて（$n^{\mathrm{water}}=66$），$\mathcal{Q}$ の残基 i の SA（$\mathrm{SA}_{\mathcal{Q}_i}$）と $\mathcal{R}$ の残基 j の SA（$\mathrm{SA}_{\mathcal{R}_j}$）を計算する．

3.3.4 合計類似度スコアと評価関数

ある対応部分構造集合の要素と対応原子を，それぞれ $\mathrm{EQ}_h^{\mathrm{local}}=\{i,j\}$，$\mathrm{EQ}_{h'}^{\mathrm{global}}=\{i',j'\}$ としたとき，類似度スコア $\mathrm{S}(\mathrm{EQ}_h^{\mathrm{local}})$ と $\mathrm{S}'(\mathrm{EQ}_{h'}^{\mathrm{global}})$ を次のように求める．

$$\begin{aligned}&\mathrm{S}(\mathrm{EQ}_h^{\mathrm{local}})\\&=\exp(-1.0\times\alpha^{\mathrm{frog}}\times\mathrm{D}_{ij}^{\mathrm{local}}(\mathrm{C}\alpha)) &(6)\\&+\exp(-1.0\times\alpha^{\mathrm{frog}}\times\mathrm{D}_{ij}^{\mathrm{local}}(\mathrm{C}\beta)) &(7)\\&+\exp(-1.0\times\beta^{\mathrm{frog}}\times|\mathrm{SA}_{\mathcal{Q}_i}-\mathrm{SA}_{\mathcal{R}_j}|), &(8)\end{aligned}$$

$$\begin{aligned}&\mathrm{S}'(\mathrm{EQ}_{h'}^{\mathrm{global}})\\&=\exp(-1.0\times\alpha^{\mathrm{frog}}\times\mathrm{D}_{i'j'}^{\mathrm{global}}(\mathrm{C}\alpha)) &(9)\\&+\exp(-1.0\times\beta^{\mathrm{frog}}\times|\mathrm{SA}_{\mathcal{Q}_{i'}}-\mathrm{SA}_{\mathcal{R}_{j'}}|). &(10)\end{aligned}$$

ここで α^{frog} と β^{frog} は定数である．

もし，$\mathrm{EQ}_h^{\mathrm{local}}=\{i,j\}$ が対応部分構造の最後なら，

$$\begin{aligned}&\mathrm{S}'(\mathrm{EQ}_h^{\mathrm{local}})\\&=\mathrm{S}(\mathrm{EQ}_h^{\mathrm{local}})\\&+\exp(-1.0\times\alpha^{\mathrm{frog}}\times\mathrm{D}_{i+1j+1}^{\mathrm{local}}(\mathrm{C}\beta)) &(11)\\&+\exp(-1.0\times\beta^{\mathrm{frog}}\times|\mathrm{SA}_{\mathcal{Q}_{i+1}}-\mathrm{SA}_{\mathcal{R}_{j+1}}|) &(12)\end{aligned}$$

となる．

類似度スコアの合計 lSOS と gSOS は，

$$\mathrm{lSOS}=\sum_{k=1}^{m}\mathrm{S}(\mathrm{EQ}_k^{\mathrm{local}}) \quad (13)$$

$$\mathrm{gSOS}=\sum_{k'=1}^{m'}\mathrm{S}'(\mathrm{EQ}_{k'}^{\mathrm{global}}) \quad (14)$$

である（m と m' はそれぞれ対応部分構造集合と対応原子集合の要素数）．

$\mathcal{Q}$ の長さで正規化した LFA 類似度と GSA 類似度を，それぞれ fl，fg とする．ここで，主鎖と側鎖と SA 類似度を用いる $fl(\mathrm{C}\alpha+\mathrm{C}\beta+\mathrm{SA})$，$\mathrm{C}\alpha$ 原子ペアと SA 類似度を用いる $fg(\mathrm{C}\alpha+\mathrm{SA})$，$\mathrm{C}\alpha$ 原子ペアと SA 類似度とギャップペナルティを用いる $fg(\mathrm{C}\alpha+\mathrm{SA}-GAP)$ を次のように定義する．

$$fl(\mathrm{C}\alpha+\mathrm{C}\beta+\mathrm{SA})=\frac{\mathrm{lSOS}}{l_{\mathrm{que}}\times 3-1} \quad (15)$$

$$fg(\mathrm{C}\alpha+\mathrm{SA})=\frac{\mathrm{gSOS}}{l_{\mathrm{que}}\times 2} \quad (16)$$

$$fg(\mathrm{C}\alpha+\mathrm{SA}-GAP)=\frac{\mathrm{gSOS}-GAP}{l_{\mathrm{que}}\times 2}. \quad (17)$$

ここでギャップ n^{gap} 個（$=l_{\mathrm{que}}-m'$）に対するペナルティは，

$$\begin{aligned}GAP=&\exp(-1.0\times\alpha^{\mathrm{frog}}\times(\mathrm{cutoff}_3+0.1))\\&\times n^{\mathrm{gap}} &(18)\end{aligned}$$

とする．

3.4 パラメータセット

GA パラメータの実験的推奨値は，世代数=2000，集団サイズ=100，$n^{\mathrm{undx}}=50$，$\alpha^{\mathrm{undx}}=0.5$，$\beta^{\mathrm{undx}}=0.35$，$n^{\mathrm{rep}}=10$ である[46)]．ヘリックスの Cα-Cα 距離が 1.54Å であり，ストランドの Cα-Cα 距離が 3.2Å–3.4Å であることから，中間値 2.24Å をとるように cutoff_1 を設定する[11),47)]．そこで，ヘリックスの 1 回ターンが約 4 残基であることから $\mathrm{cutoff}_1=0.56$，$n^{\mathrm{cons}}=4$ に設定する．また，ストランドの Cα-Cα 距離を考慮して $\mathrm{cutoff}_2=3.5$，$\mathrm{cutoff}_3=3.5$ とする．

α^{frog} によって LFA 評価値と GSA 評価値のバランスが調節される．相同タンパク質ペアに対する先行実験結果[48)] に基づいて fl と fg の相関係数が 0.8 以上となるような $\alpha^{\mathrm{frog}}=0.5$，$\beta^{\mathrm{frog}}=11.0$ に設定する．

3.5 非同期並列化と計算時間

FROG は Grid RPC システムである Ninf（`http://ninf.apgrid.org/`）によって非同期並列化され，32 CPU の Linux クラスタ（Pentium3 1 GHz）に実装された．

管理ノードは各計算ノードに重複しない親 3 個体を渡す．そして，ある計算ノードから 3 個体が返却されたときに世代を更新し，新たな親 3 個体をその計算ノードへ渡す．

非同期並列化によって平均長 270.5 のタンパク質ペアに対する計算時間が約 97 分から約 15 分に短縮された．全体的におおよそ 7 倍から 10 倍の高速化に成功した[49)]．

4. 実　　験

ここでは，1. 3つのタンパク質ペアを用いてFROGの大域的探索性能を観察し，2. タンパク質ペアの類似度分布を用いて評価関数の有効性を確認する．その後，3. 相同タンパク質の認識性能を既存手法と比較し，4. 部分構造と全体構造の保存・変異関係の可視化を行う．最後に実験結果のまとめを行う．実験1以外の提案手法の結果は10試行における最良LFA類似度の試行とする．

タンパク質ペアはSCOP☆の分類に従って用意する．SCOPのタンパク質構造空間はFamily，Superfamily，Fold，Classの階層構造になっている．Familyは構造—機能相関の明らかなタンパク質グループであり，Family どうしの弱い進化的関連性はSuperfamilyとして分類される．Foldは全体構造の形が保存されている機能の異なるSuperfamilyのグループであり，Foldの支配的な二次構造によってClassが決定される．

ここでは，同じFamilyに属する相同（ホモロジ）タンパク質ペアをHペア，同じSuperfamilyの異なるFamilyに属するリモートホモロジペアをRHペア，同じFoldの異なるSuperfamilyに属する類似（アナロジ）ペアをANペアとする．

4.1 探索挙動

4.1.1 実験目的

FROGの大域的探索性能を確認するために，集団の平均評価値曲線を用いてGAの探索挙動と世代交代モデルの効果を観察する．世代交代モデルは家族における最良LFA個体の回転行列を探索集団へ戻すモデルと戻さないモデルを比較する．

4.1.2 実験準備

比較されるタンパク質ペアによって以下の3つの回転行列がありうる．

GSA類似度の高いタンパク質ペアの対応部分構造は順序LFAである．したがって最適LFAにおける回転行列と最適GSAの合同変換における回転行列は類似する（Case1）．GSA類似度の低いタンパク質ペアにおける対応部分構造が順序LFAである場合，最適LFAの回転行列は最適GSAのと類似する（Case2）．一方，GSA類似度の低いタンパク質ペアにおける対応部分構造が非順序LFAである場合，最適LFAの回転行列は最適GSAと異なる（Case3）．最適順序LFAをGSAへ拡張する方法や最適GSAを探索する方法はCase1とCase2に適している．しかし，Case3は非順序LFAを集中的に探索しなければならない．

ここで，新たな世代交代モデルを採用する「gFROG」を用いてCase3におけるFROGの性能を比較する．gFROGは3.2.4項のStep 4を行わない最適GSAの探索方法である．

SCOPから次のような3つのタンパク質ペアを用意した．Retroviral proteaseである1idaa☆☆と1mvpaはHペアである（Case1）．Heat-labile toxin（1tiid）とtRNA synthetase（1eqra）はOB-foldを共有するANペア[50]である（Case2）．4-helical bundleトポロジのCytochrome（1bbha）と5-helical bundleトポロジのApolipophorin-III（1aep）は最適GSAにおける順序LFAが3つのヘリックスであるが，最適GSAの角度を約180度（Upside-down）回転させると4つのヘリックスが非順序LFAとして一致する（Case3）[51]．

4.1.3 実験結果・考察

異なる乱数系を用いた100試行における集団の平均評価値曲線の推移を**図5**に示す．

Case1（図5(a)）とCase2（図5(b)）における最適GSAは順序LFAで構成されるため，gFROGとFROGとの探索性能に違いはほとんどない．集団の収束様子について，最良順序LFA情報を参照するFROGの集団収束がgFROGより早い．

Case3に関して，gFROGは非順序LFAに対応しないため，集団は類似度の低い合同変換へ収束する（図5(c-1)）．一方，FROGにおける集団は非順序LFAの回転行列へ70試行収束する（図5(c-2)）．

以上の結果から，FROGはGSAおよび順序・非順序LFAの最適化に効果的であることを確認した．

4.2 評価関数における類似度スコア

4.2.1 実験目的

評価関数における主鎖ペア・側鎖ペアおよび対応原子の距離類似度とSA類似度の影響を観察する．4種類の評価関数をそれぞれ最適化し，同じフォールドを共有するタンパク質ペアにおける類似度分布の変化を示す．

4.2.2 実験準備

式(6)および式(9)の類似度スコア関数をそれぞれ$S_{C\alpha}$，$S'_{C\alpha}$とすると，主鎖ペアと原子ペアの距離類似度を評価する$fl(C\alpha)$と$fg(C\alpha)$は，

☆ Structural Classification of Proteins（1.63 release）: http://scop.mrc-lmb.cam.ac.uk/scop/

☆☆ Protein Data Bank（PDB）の4文字IDコード，5文字の場合は右1文字がチェーンIDを示す．

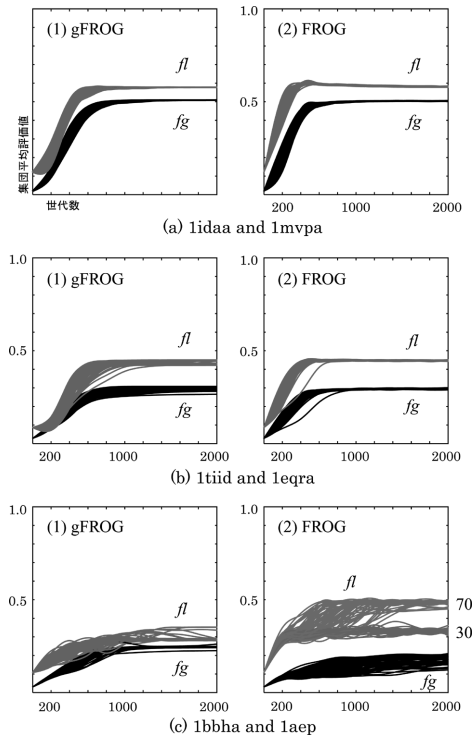

図 5 100 試行における集団の平均評価値曲線：gFROG は最適な GSA を探索する．FROG は GSA および順序・非順序 LFA を探索する．相同タンパク質ペア (a) および類似タンパク質ペア (b) における gFROG と FROG の性能に差はほとんどない．3 つのヘリックスが順序 LFA として，4 つのヘリックスが非順序 LFA として存在するタンパク質ペア (c) において FROG は非順序 LFA へ 70 試行収束する

Fig. 5 Population convergences of 100 trials in three cases.

$$fl(\mathrm{C}\alpha) = \frac{\sum_{k=1}^{m} \mathrm{S}_{\mathrm{C}\alpha}(\mathrm{EQ}_k^{\mathrm{local}})}{l_{\mathrm{que}} - 1} \tag{19}$$

$$fg(\mathrm{C}\alpha) = \frac{\sum_{k'=1}^{m'} \mathrm{S'}_{\mathrm{C}\alpha}(\mathrm{EQ}_{k'}^{\mathrm{global}})}{l_{\mathrm{que}}} \tag{20}$$

である．

式 (6) と式 (8)，もしフラグメントの最後なら式 (12) を加えた類似度スコアを $\mathrm{S}_{\mathrm{C}\alpha+\mathrm{SA}}$ とする．主鎖ペアの距離類似度と SA 類似度を評価する $fl(\mathrm{C}\alpha+\mathrm{SA})$ は，

$$fl(\mathrm{C}\alpha+\mathrm{SA}) = \frac{\sum_{k=1}^{m} \mathrm{S}_{\mathrm{C}\alpha+\mathrm{SA}}(\mathrm{EQ}_k^{\mathrm{local}})}{l_{\mathrm{que}} \times 2 - 1} \tag{21}$$

である．ただし，式 (4) では $\mathrm{D}_{ij}^{\mathrm{local}}(\mathrm{C}\alpha) \leq \mathrm{cutoff}_1$ だけを用いる．

上記の評価関数群と式 (15)，式 (16)，式 (17) による類似度分布の変化を観察するために，SCOP から OB-fold 2415 ペアを用意した．OB-fold は配列類似度 25% 以下の β バレル構造を共有する多糖結合タンパク質である．OB-fold の β バレル構造は 7–8 個のストランドからなっており，1 つのヘリックスによってキャップされている特徴がある．

4.2.3 実験結果・考察

$fl(\mathrm{C}\alpha)$ と $fg(\mathrm{C}\alpha)$ によるタンパク質ペアの類似度分布は（**図 6** (a-1)），$fl(\mathrm{C}\alpha+\mathrm{SA})$ と $fg(\mathrm{C}\alpha+\mathrm{SA})$ によって全体的に左へ移動する（図 6 (a-2)）．すなわち，SA 類似度は GSA 探索より LFA 探索に影響を与える．

ギャップペナルティは fg に影響するため，$fl(\mathrm{C}\alpha+\mathrm{SA})$ と $fg(\mathrm{C}\alpha+\mathrm{SA})$ による類似度分布（図 6 (b-1)）は全体的に下へ移動する（図 6 (b-2)）．特に GSA 類似度の低い RH ペア・AN ペアに対して顕著である．

側鎖ペアの類似度を加えた $fl(\mathrm{C}\alpha+\mathrm{C}\beta+\mathrm{SA})$ は H ペアの類似度分布をよりコンパクトにまとめると同時に，RH ペア・AN ペアにおける類似部分構造を強調する（図 6 (c-2)）．

以上の結果より，タンパク質ペアの類似度分布は評価関数 $fl(\mathrm{C}\alpha+\mathrm{C}\beta+\mathrm{SA})$ と $fg(\mathrm{C}\alpha+\mathrm{SA})$ によって生化学的に有意な方向へシフトされることが分かった．また，H ペアと RH ペア・AN ペアにおける $\mathrm{C}\alpha$ 原子類似度の重なりが（図 6 (a) の拡大図）著しく分離されることが観察された（図 6 (c) の拡大図）．

4.3 相同タンパク質の認識性能

4.3.1 実 験 目 的

類似機能の H ペアと RH ペアおよび異なる機能の H ペアと AN ペアにおける FROG の相同タンパク質認識性能を既存手法と比較実験する．

4.3.2 実 験 準 備

Sensitivity（感度）と Specificity（特異度）は下式で与えられる[52]．

$$\mathrm{Sensitivity}(s) = \frac{N_{\mathrm{h}}(s)}{N_{\mathrm{h}}} \tag{22}$$

$$\mathrm{Specificity}(s) = \frac{N_{\mathrm{h}}(s)}{N(s)}. \tag{23}$$

ここで，s は降順にソートされたスコア，N_{h} と N はそれぞれ H ペアの総数とタンパク質ペアの総数，$N_{\mathrm{h}}(s)$ はスコア s 以上の H ペアの数，$N(s)$ は s 以上のタンパク質ペアの数である．特異度が下がり始めるスコアの H ペアから最後の H ペアまでのペア数を M_{h}，このスコア領域に存在する RH・AN ペアの数を $M_{\neg\mathrm{h}}$ とする．このような構造的類似度と機能との関係の曖

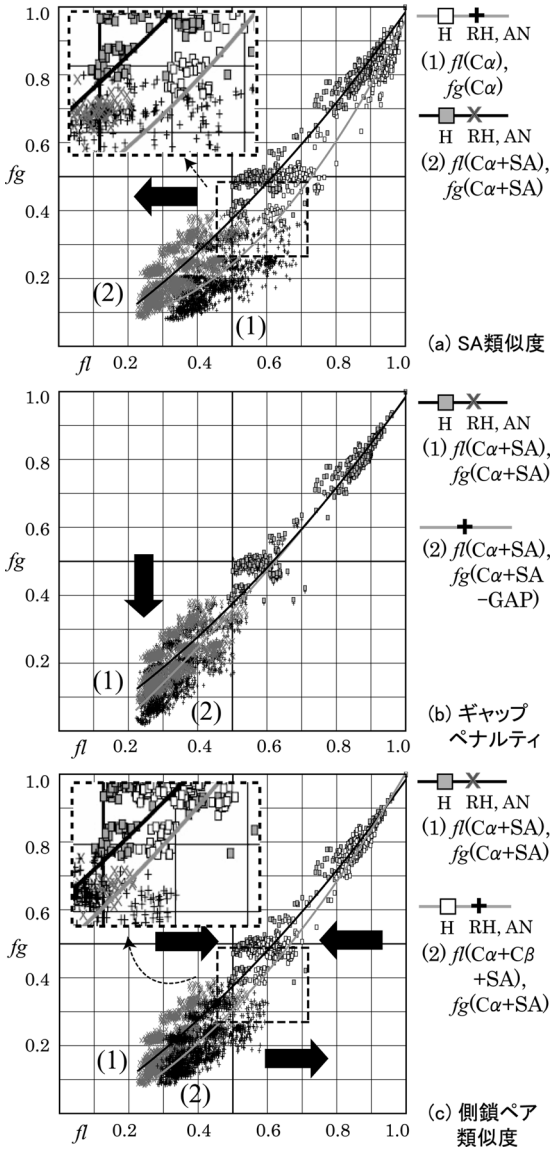

図 6 OB-fold 2415 ペアにおける類似度分布：4 種類の評価関数をそれぞれ最適化し，LFA 類似度と GSA 類似度を横軸と縦軸としてタンパク質ペアの類似度をプロットした．OB-fold 2415 ペアは 622 のホモロジ（H）ペアと 1793 のリモートホモロジ（RH）ペア・アナロジ（AN）ペアから構成されている．各類似度分布の見やすさを考慮して多項式近似曲線を示したが，LFA 類似度と GSA 類似度が多項式で近似可能という意味ではない．Cα の幾何学的類似度関数 (a-1) に SA 類似度を反映すると（a-2）LFA 類似度が低下し，さらに側鎖ペア類似度を加えると (c-2)H ペアと RH ペア・AN ペアの類似度分布は顕著に分離される（カラー図面は WEB サイト[41] を参照すること）

Fig. 6 Distributions of similarities for the different objective functions.

昧な領域を OR（Overlapping Region）と呼ぶことにする．OR に存在する H ペアと RH・AN ペアの割合

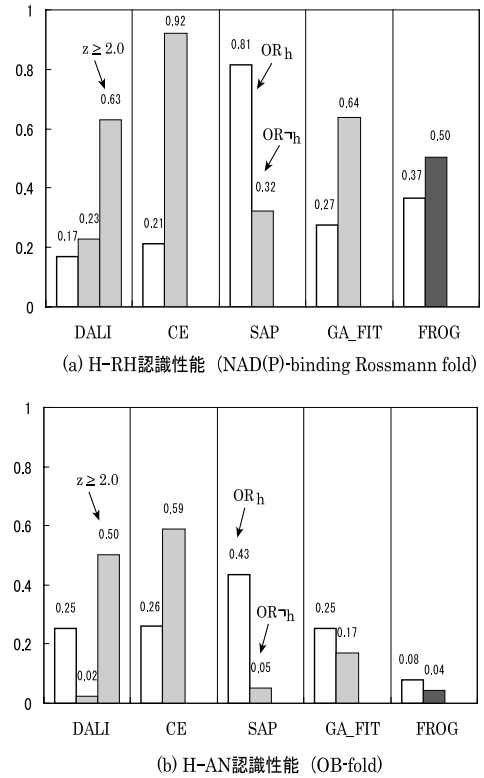

図 7 相同タンパク質の認識性能：$\mathrm{OR_h}$ と $\mathrm{OR_{\neg h}}$ は構造的類似度の曖昧な領域に存在するホモロジペアの割合とリモートホモロジ・アナロジペアの割合を示す．ホモロジペアとリモートホモロジペアが混在するタンパク質ペアにおける相同タンパク質ペアの認識性能は DALI が最も優れており (a)，ホモロジペアとアナロジペアにおける認識性能は FROG が最も良い (b)

Fig. 7 Performances of recognizing homologues.

は下式で定義される．

$$\mathrm{OR_h} = \frac{M_\mathrm{h}}{N_\mathrm{h}} \tag{24}$$

$$\mathrm{OR_{\neg h}} = \frac{M_{\neg \mathrm{h}}}{N - N_\mathrm{h}}. \tag{25}$$

H-RH 認識性能比較に用いるタンパク質ペアセットは，NAD(P)-binding Rossmann fold に属する 946 ペア（8 種類の Family，$N_\mathrm{h} = 142$）とする．H-AN 認識性能比較のためには，OB-fold 3486 ペア（13 種類の Family からなる 5 種類の Superfamily，$N_\mathrm{h} = 652$）を用いる．比較手法は DALI（z-score），SAP，CE（raw score），GA_FIT とする．FROG におけるスコア s は $fl(\mathrm{C}\alpha + \mathrm{C}\beta + \mathrm{SA})$ と $fg(\mathrm{C}\alpha + \mathrm{SA})$ との論理積，すなわち，スコア $s(fl, fg)$ と $s'(fl', fg')$ が $(fl \leq fl' \ \wedge \ fg \leq fg')$ のときに $s \leq s'$ とする．

4.3.3 実験結果・考察

図 7 (a) において DALI は H-RH 認識性能に最も優れている．DALI における H の 24 ペア（17%）のス

表 2 NAD(P)-binding Rossmann fold における LFA 類似度と GSA 類似度：* タンパク質における Cα 原子数，†*fl* と *fg* は FROG の LFA 類似度と GSA 類似度，DALI は z-score，‡RMSD における対応原子数

Table 2 Comparison of structure similarity between the proposed method and existing methods.

タンパク質ペア	ツール †	スコア	RMSD	N‡
2hdha and 1xel	SARF2		2.36	117
(286* and 338)	*fl*	0.2095	2.46	179
	DALI	0.3	9.20	91
	fg	0.3742	2.04	118
1c1da and 1kola	SARF2		2.52	115
(249 and 396)	*fl*	0.1749	2.17	183
	DALI	0.5	5.60	90
	fg	0.3192	2.17	132
1xel and 1bg6	SARF2		2.58	119
(338 and 349)	*fl*	0.1697	3.21	178
	DALI	0.5	4.30	112
	fg	0.3145	2.27	118
1keva and 1bgva	SARF2		2.54	113
(351 and 449)	*fl*	0.1685	2.93	201
	DALI	0.8	4.60	112
	fg	0.3230	2.13	123
1ae1b and 2hdha	SARF2		2.59	99
(258 and 286)	*fl*	0.2118	2.56	144
	DALI	1.1	4.20	103
	fg	0.3383	2.11	110
1fmca and 1c1da	SARF2		2.43	117
(255 and 349)	*fl*	0.218	2.90	157
	DALI	1.6	4.90	84
	fg	0.3625	1.88	103
1ae1b and 1hyha	SARF2		2.34	112
(258 and 297)	*fl*	0.2286	2.81	149
	DALI	1.9	5.60	102
	fg	0.3636	2.15	121

コア領域に RH の 183 ペア（23%）が存在する．構造的類似性が存在する（z-score≥ 2.0）RH ペアは 505 ペア（63%）である．H-RH 認識は代表 PDB 構造における統計的有意性を用いる方法が適しているといえる．実際，NAD(P)-binding Rossmann fold は 1 つの Superfamily しか存在しない特殊な構造–機能関係のタンパク質群である．

しかしながら，DALI z-score が 2.0 より低い RH ペアでも著しい構造的類似度を示すペアが存在する．FROG は，**表 2** に示す RH ペアの 35%以上の Cα 原子ペアを RMSD 3.5Å 以下で重ねる．

図 7 (b) の H-AN 認識について，DALI における H の 166 ペア（25%）のスコア領域に AN の 68 ペア（2%）のスコアが散在する．そのうえ，z-score≥ 2.0 の AN ペアの割合は 50%にものぼる．FROG における M_h はわずか 51 ペア（8%）であり，最も優れた H-AN 認識性能を示している．

以上の結果から，FROG は H-RH 認識における

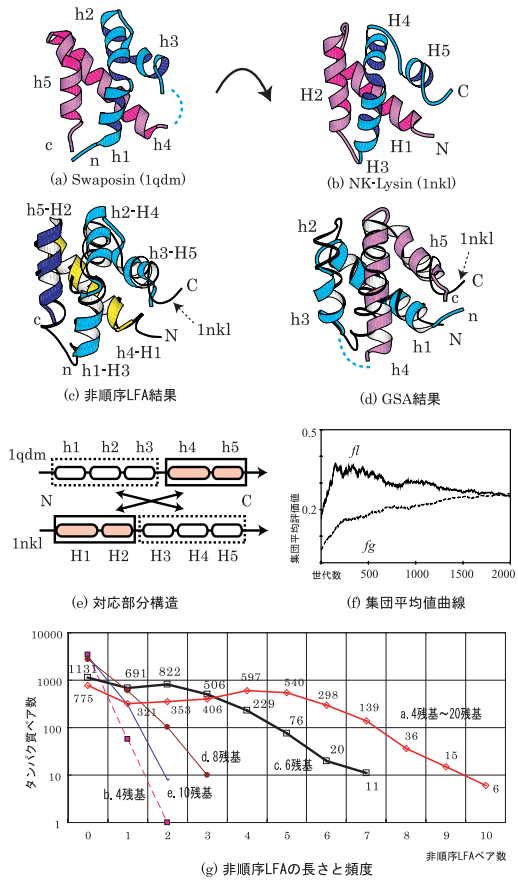

図 8 非順序対応部分構造と GSA との関係：Swaposin と NK-lysin は Saposin Superfamily に属する Circular Permutation のタンパク質ペアである．提案手法 FROG は非順序対応部分構造 (e) の LFA 結果 (c) と GSA 結果 (d) を同時に発見する．LFA と GSA における部分構造の順序が変化するときに FROG は特徴的な集団探索挙動 (f) を見せる．OB-fold 3486 ペアにおいて非順序対応部分構造を 4 つ持っているペアが最も多く (g-a)，長さ 6 残基の非順序対応部分構造の出現頻度が最も高い (g-c)

Fig. 8 Effectiveness of the proposed method on capturing the structural flexibility of proteins.

DALI の補完的な役割を果たすとともに，同じフォールドを共有するタンパク質群における Family 同定に最も優れているということが分かった．

4.4 非順序 LFA—GSA 関係の可視化

4.4.1 実 験 目 的

集団の平均評価値曲線を用いて非順序 LFA の部分構造と GSA との関係を可視化する．

4.4.2 実 験 準 備

明らかな非順序対応部分構造が存在するタンパク質ペアを用意する．Saposin Superfamily に属する Swaposin タンパク質の 1qdm（残基番号 1 s-104 s，**図 8** (a)）と NK-lysin タンパク質の 1nkl（図 8 (b)）は N

末と C 末のつながりが異なる Circular Permutation の RH ペアである[53]．

4.4.3 実験結果・考察

FROG は図 8 (c) の LFA 結果と図 8 (d) の GSA 結果を同時に発見する．LFA は 1qdm の第 1 番目のヘリックスが 1nkl の第 3 番目のヘリックスと，1qdm の第 4 番目のヘリックスが 1nkl の第 1 番目のヘリックスと一致する非順序 LFA である（図 8 (e)）．

図 8 (f) の平均評価値曲線において，LFA 類似度推移はある世代以降減少し続ける特徴的な振舞いを見せる．すなわち，最良非順序 LFA 個体の回転行列へ向かっていた集団の移動方向が GSA 最適化に有効な順序 LFA の回転行列へ切り替わったことを意味する．

実験 4.3 の OB-fold 3486 ペアにおいて，図 8 (f) のような探索挙動を示したペア数は 676（19%）であり（最終平均評価値の差 $|fl - fg > 0.01| \wedge fg > 0.2$ のとき），そのうち AN ペアが 583（86%）であった．非順序対応部分構造を共有する 2711 ペア（78%）のうち，H ペアが 41（1%），AN ペアが 2670（77%）であった．非順序対応部分構造を 4 つ持っているペア数が最も多く（図 8 (g-a)），長さ 6 残基の非順序対応部分構造の出現頻度が最も高い（図 8 (g-c)）．

以上の結果より，FROG の集団平均評価値曲線は全体構造における部分構造と非順序類似部分構造の保存・変異関係をとらえることが可能であるといえる．

4.5 実験のまとめ

4 つの実験を行って提案手法 FROG の性能を確認した．大域的探索性能に優れている FROG は生化学的に有効な評価関数を最適化し，順序・非順序 LFA と GSA を同時に行う新規 PSA 手法である．FROG は相同タンパク質の認識性能に優れており，タンパク質全体構造の非順序 LFA の順序が変化するときに特徴的な集団平均評価値曲線を示す．したがって，全体構造における部分構造の保存と変異が観察できるユニークなツールである．

部分構造と全体構造における LFA の順序が変化するタンパク質ペアの多くは，フォールドが保存されていて機能が異なるアナロジペアであった．自然界に存在するフォールドの数が限られているとすると[1]，部分構造の出現順位の変化はフォールド保存と機能進化の関わりを示唆していると考える．

5. ま と め

本論文では，タンパク質の柔軟性を正確にとらえられない既存のタンパク質立体構造比較手法の問題点を指摘したうえで，部分構造比較と全体構造比較を同時に最適化する 2 層比較を提案した．実数値 GA によって実装された提案手法は部分構造と全体構造との保存・変異関係を示すことが可能であり，タンパク質の柔軟性を考慮した構造—機能相関解析における効果的なツールであることを示した．

今後，様々なフォールドにおける部分構造と全体構造との関係を網羅的に解析し，タンパク質の柔軟性と機能保存・獲得について議論したい．

謝辞 査読者の方々と神戸大学理学部の千見寺浄慈博士には適切なご意見とご討論をいただきました．ここに感謝の意を表します．

参 考 文 献

1) Chothia, C.: One Thousand Families for the Molecular Biologist, *Nature*, Vol.357, pp.543–544 (1992).
2) Blomberg, N., Baraldi, E., Nilges, M. and Saraste, M.: The PH Superfold: a Structural Scaffold for Multiple Functions, *Trends Biochem. Sci.*, Vol.24, pp.441–445 (1999).
3) Russell, R.B. and Sternberg, M.J.E.: Two New Examples of Protein Structural Similarities within the Structure-function Twilight Zone, *Protein Eng.*, Vol.10, pp.333–338 (1997).
4) Murzin, A.G.: How Far Divergent Evolution Goes in Proteins, *Current Opinion in Structural Biol.*, Vol.8, pp.380–387 (1998).
5) Getz, G., Vendruscolo, M., Sachs, D. and Domany, E.: Automated Assignment of SCOP and CATH Protein Structure Classification from FSSP Scores, *Proteins*, Vol.46, pp.405–415 (2002).
6) Goldsmith-Fischman, S. and Honig, B.: Structural Genomics: Computational Methods for Structure Analysis, *Protein Sci.*, Vol.12, pp.1813–1821 (2003).
7) Gibrat, J.F., Madej, T. and Bryant, S.H.: Surprising Similarities in Structure Comparison, *Current Opinion in Structural Biol.*, Vol.6, pp.377–385 (1996).
8) Eidhammer, I., Jonassen, I. and Taylor, W.R.: Structure Comparison and Structure Patterns, *J. Computational Biol.*, Vol.7, pp.685–716 (2000).
9) Koehl, P.: Protein Structure Similarities, *Current Opinion in Structural Biol.*, Vol.11, pp.348–353 (2001).
10) May, A.C.W. and Johnson, M.S.L.: Improved Genetic Algorithm-based Protein Structure Comparisons: Pairwise and Multiple Superpositions, *Protein Eng.*, Vol.8, pp.873–882 (1995).
11) Singh, A.P. and Brutlag, D.L.: Hierarchical

Protein Structure Superposition Using Both Secondary Structure and Atomic Representations, *Proc. Intelligent Systems for Mol. Biol.*, pp.284–293 (1997).
12) Shindyalov, I.N. and Bourne, P.E.: Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path, *Protein Eng.*, Vol.11, pp.739–747 (1998).
13) Taylor, W.R.: Protein Structure Comparison Using Iterated Double Dynamic Programming, *Protein Sci.*, Vol.8, pp.654–665 (1999).
14) Holm, L. and Sander, C.P.: Protein Structure Comparison by Alignment of Distance Matrices, *J. Mol. Biol.*, Vol.233, pp.123–138 (1993).
15) Alexandrov, N.N. and Go, N.: Biological Meaning, Statistical Significance, and Classification of Local Spatial Similarities in Nonhomologous Proteins, *Protein Sci.*, Vol.3, pp.866–875 (1994).
16) Lehtonen, J.V., Denessiouk, K., May, A.C.W. and Johnson, M.S.: Finding Local Structural Similarities among Families of Unrelated Protein Structures: a Generic Non-linear Alignment Algorithm, *Proteins*, Vol.34, pp.341–355 (1999).
17) Szustakowski, J.D. and Weng, Z.: Protein Structure Alignment Using a Genetic Algorithm, *Proteins*, Vol.38, pp.428–440 (2000).
18) Hamada, D. and Goto, Y.: The Equilibrium Intermediate of Beta-lactoglobulin with Nonnative Alpha-helical Structure, *J. Mol. Biol.*, Vol.269, pp.479–487 (1997).
19) Riechmann, L. and Winter, G.: Novel Folded Protein Domains Generated by Combinatorial Shuffling of Polypeptide Segments, *Proc. Natl. Acad. Sci. USA*., Vol.97, pp.10068–10073 (2000).
20) Chothia, C., Gough, J. and C. Vogel, S.A.T.: Evolution of the Protein Repertoire, *Science*, Vol.300, pp.1701–1703 (2003).
21) Wierenga, R.K.: The TIM-barrel Fold: a Versatile Framework for Efficient Enzymes, *FEBS Letters*, Vol.492, pp.193–198 (2001).
22) Krem, M.M., Prasad, S. and Cera, E.D.: Ser(214) is Crucial for Substrate Binding to Serine Proteases, *J. Biol. Chem.*, Vol.277, pp.40260–40264 (2002).
23) Shatsky, M., Nussinov, R. and Wolfson, H.: Flexible Protein Alignment and Hinge Detection, *Proteins*, Vol.48, pp.242–256 (2002).
24) Ye, Y. and Godzik, A.: Flexible Structure Alignment by Chaining Aligned Fragment Pairs Allowing Twists, *Bioinformatics*, Vol.19, pp.ii246–ii255 (2003).
25) Goldberg, D.E.: *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley (1989).
26) Srinivas, N. and Deb, K.: Multiobjective Optimization Using Nondominated Sorting in Genetic Algorithms, *Evolutionary Computation*, Vol.2, pp.221–248 (1994).
27) Hendrickson, W.A.: Transformations to Optimize the Superposition of Similar Structures, *Acta Cryst.*, Vol.A35, pp.158–163 (1979).
28) Needleman, S.B. and Wunsch, C.D.: A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, *J. Mol. Biol.*, Vol.48, pp.443–453 (1970).
29) Smith, T.F. and Waterman, M.S.: Identification of Common Molecular Subsequences, *J. Mol. Biol.*, Vol.147, pp.195–197 (1981).
30) 山村雅幸，小林重信：遺伝的アルゴリズムの工学的応用，人工知能学会誌，Vol.9, pp.506–511 (1994).
31) 佐藤　浩，小野　功，小林重信：遺伝的アルゴリズムにおける世代交代モデルの提案と評価，人工知能学会誌，Vol.12, pp.734–744 (1997).
32) Eshleman, L. and Schaffer, J.: Real-Coded Genetic Algorithms and Interval-Schemata, *Foundations of Genetic Algorithms*, pp.187–202 (1992).
33) 小野　功，佐藤　浩，小林重信：単峰性正規分布交叉 UNDX を用いた実数値 GA による関数最適化，人工知能学会誌，Vol.14, pp.1146–1155 (1999).
34) 喜多　一，小野　功，小林重信：実数値 GA のための正規分布交叉に関する理論的考察，計測自動制御学会論文集，Vol.35, pp.1333–1339 (1999).
35) Chikenji, G., Fujitsuka, Y. and Takada, S.: A Reversible Fragment Assembly Method for De Novo Protein Structure Prediction, *J. Chem. Phys.*, Vol.119, pp.6895–6903 (2003).
36) 後藤祐児，谷澤克行：タンパク質の分子設計，共立出版 (2001).
37) Feng, Z.K. and Sippl, M.J.: Optimum Superimposition of Protein Structures: Ambiguities and Implications, *Folding & Design*, Vol.1, pp.123–132 (1996).
38) Chen, L., Zhou, T. and Tang, Y.: Protein structure alignment by deterministic annealing, *Bioinformatics* (2004).
39) Godzik, A.: The Structural Alignment between Two Proteins: Is There a Unique Answer?, *Protein Sci.*, Vol.5, pp.1325–1338 (1996).
40) Yang, A. and Honig, B.: An Integrated Approach to The Analysis and Modeling of

Protein Sequences and Structures. I. Protein Structural Alignment and a Quantitative Measure for Protein Structural Distance, *J. Mol. Biol.*, Vol.301, pp.665–678 (2000).
41) FROG. http://www.es.dis.titech.ac.jp/snail/
42) Lee, B. and Richards, F.M.: The Interpretation of Protein Structure: Estimation of Static Accessibility, *J. Mol. Biol.*, Vol.55, pp.379–400 (1971).
43) Hiroike, T. and Toh, H.: A Local Structural Alignment Method that Accommodates with Circular Permutation, *Chem-Bio Informatics J.*, Vol.1, pp.103–114 (2001).
44) Jung, J. and Lee, B.: Protein Structure Alignment Using Environmental Profiles, *Protein Eng.*, Vol.13, pp.535–543 (2000).
45) Koradi, R., Billeter, M. and Wüthrich, K.: MOLMOL: A Program for Display and Analysis of Macromolecular Structures, *J. Mol. Graphics*, Vol.14, pp.51–55 (1996).
46) Park, S.J. and Yamamura, M.: GA-based Generic Method for Protein Structure Comparison, *Proc. 2003 IEEE Congress on Evolutionary Computation* (*CEC 2003*), pp.1528–1535 (2003).
47) Gerstein, M. and Levitt, M.: Using Iterative Dynamic Programming to Obtain Accurate Pairwise and Multiple Alignments of Protein Structures, *Proc. Intelligent Systems for Mol. Biol.*, pp.59–67 (1996).
48) Park, S.J. and Yamamura, M.: Two-layer Protein Structure Comparison, *Proc. 15th IEEE International Conference on Tools with Artificial Intelligence* (*ICTAI 2003*), pp.435–440 (2003).
49) Park, S.J. and Yamamura, M.: FROG (Fitted Rotation and Orientation of protein structure by means of real-coded Genetic algorithm) : Asynchronous Parallelizing for Protein Structure-based Comparison on the Basis of Geometrical Similarity, *Genome Informatics*, Vol.13, pp.344–345 (2002).
50) Russell, R.B., Saqi, M.A.S., Sayle, R.A., Bates, P.A. and Sternberg, M.J.E.: Recognition of Analogous and Homologous Protein Folds: Analysis of Sequence and Structure Conservation, *J. Mol. Biol.*, Vol.269, pp.423–439 (1997).
51) Dror, O., Benyamini, H., Nussinov, R. and Wolfson, H.: MASS: Multiple Structural Alignment by Secondary Structures, *Bioinformatics*, Vol.19, pp.i95–i104 (2003).
52) Rice, D.W. and Eisenberg, D.: A 3D-1D Substitution Matrix for Protein Fold Recognition that Includes Predicted Secondary Structure of the Sequence, *J. Mol. Biol.*, Vol.267, pp.1026–1038 (1997).
53) Kervinen, J., Tobin, G.J., Costa, J., Waugh, D.S., Wlodawer, A. and Zdanov, A.: Crystal Structure of Plant Aspartic Proteinase Prophytepsin: Inactivation and Vacuolar Targeting, *EMBO Journal*, Vol.18, pp.3947–3955 (1999).

(平成 16 年 3 月 23 日受付)
(平成 17 年 1 月 7 日採録)

朴　聖俊

1998 年専修大学経営学部情報管理学科卒業．2000 年東京工業大学大学院総合理工学研究科知能システム科学専攻修士課程修了．2001 年 4 月同大学院同専攻博士後期課程入学，2003 年 9 月神戸大学理学部技術補佐員，現在に至る．進化型計算，並列計算，バイオインフォマティクス等の研究に従事．

高田　彰二

1988 年京都大学理学部卒業．1990 年同大学院理学研究科化学専攻修士修了．1991 年岡崎国立共同研究機構技官，1995 年日本学術振興会研究員，1998 年神戸大学理学部講師，2001 年同大学助教授，現在に至る．生物物理，理論的タンパク質構造と機能解析，タンパク質立体構造予測の研究に従事．

山村　雅幸（正会員）

1982 年東京工業大学工学部制御工学科卒業．1987 年同大学院総合理工学研究科システム科学専攻博士後期課程満期退学．同年東京工業大学助手．1996 年同大学助教授，2004 年同大学教授，現在に至る．学習・進化システム，バイオインフォマティクス，分子コンピュータの研究に従事．人工知能学会，認知科学会，ソフトウェア科学会等の会員．