社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－MPS－48　(3)
2004／3／1

タンパク質立体構造の2層比較

朴　聖俊[†]　山村雅幸[††]

タンパク質は，三次元立体構造によって固有の生物学機能を発現するため，立体構造を比較・分類することは非常に重要である．立体構造は，配列変異に対してロバストであり，部分構造と全体構造は機能進化過程において，強く保存されることが知られている．本研究では，部分構造を全体構造のビルディングブロックと捉え，部分–全体構造相関に着目する立体構造比較ツールを開発する．提案手法は，非同期並列化された実数値遺伝的 GA を用いて，有意な部分構造を全体構造比較に用いる 2 層比較を実現する．

Two-layer Comparison of Protein Structures

Sung-Joon Park[†] and Masayuki Yamamura[††]

The proteins fold into the native structures that express biological functions, and therefore comparing three-dimensional protein structures and classifying them are extremely important to understand the nature of protein molecules. Generally, the local structure and global structure that will be related to the survival of the fitness are strongly conserved in the process of molecular evolution. In here, we suggest an approach to lifting the veil of the relationship between local structure and global structure on the basis of assumption that local structures play a crucial role in assembling the global topology. The idea, two-layer comparison, proposed in this study is based on a Real-coded GA asynchronously parallelized.

1. はじめに

ポリペプチド鎖であるタンパク質は，三次元空間で固有の立体構造へフォールディングして，生化学的機能を発現する．タンパク質機能進化における立体構造は，アミノ酸配列（一次構造）の変異に対してロバストであるため [1]，一次構造–機能相関のあいまいな **Twilight Zone** は [2]，立体構造の保存性を観察することで明らかになる [3]．しかし，立体構造–機能相関においても Twilight Zone が存在するため [4]，改善されたタンパク質立体構造比較手法の開発が望まれる．

局所的構造部分は，タンパク質の活性部位など機能発現に関わる．このような部分構造を全体構造のビルディングブロックと考えると [5]，タンパク質における LFA（Local Fragment Alignment）と GSA（Global Superposition Alignment）の保存・変異相関を観察することは，タンパク質機能進化の理解に重要である．しかし，既存手法は LFA と GSA を排他的・特異的に比較するため，LFA-GSA 相関を明らかにすることはできない．

本研究では，二つのタンパク質立体構造における LFA 類似度と GSA 類似度を同時発見する手法を提案する．提案手法は，平行移動ベクトルと回転行列によってコード化した実数値 GA の集団を用いて，タンパク質の Cα バックボーンの合同変換を比較する．潜在的な最適 LFA と GSA は，スコア非依存 DP を用いて定義し，溶媒接触可能性と幾何情報の類似度を評価する．その際，有意な LFA の情報が GSA 探索に用いられる **2 層比較**を実現する．また，非同期並列化された世代交代モデルを用いることで，実用的計算時間を得る．GA の集団探索性を活用する提案手法は，LFA-GSA 相関が観察できる唯一のツールである．

2. 提案手法

2.1 FROG の手続き

タンパク質立体構造比較とは，参照構造（$\mathcal{R}$）に対する問い合わせ構造（$\mathcal{Q}$）の最適な合同変換 $\mathcal{Q}' = \mathcal{Q} \times \mathcal{M} + \overrightarrow{T}$ をみつけることである．GA 集団を回転行列 $\mathcal{M}$ と並行移動ベクトル $\overrightarrow{T}$ でコード化する提案手法（FROG）は，図 1 の三つのステップからなる．そして，2 層比較とは，交叉によって生成された子集団の優れた $\mathcal{M}$ を次世代へ戻すことによって，最適な合同変換のビルディングブロックをプールすることを指す．

† 東京工業大学総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
email: park@es.dis.titech.ac.jp
†† 東京工業大学総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
email: my@dis.titech.ac.jp

1. 入力：
   1.1 $\mathcal{Q}$ と $\mathcal{R}$ から重心 c を計算
   1.2 $\mathcal{Q}$ と $\mathcal{R}$ を Cartesian 軸上へ平行移動（重心 c が原点）
   1.3 $\mathcal{Q}$ のランダムな n^{rep} 個の Cα 原子からなる代表構造 $\mathcal{Q}^R$ を定義
   1.4 原点から最も遠い $\mathcal{R}$ の Cα 原子の距離 $|\vec{r}|$ を計算
2. 最適化：
   2.1 初期集団を生成（$\overrightarrow{T}$ は $|\vec{r}|$ 直径球の内側にプロットした $\mathcal{Q}$ の重心 c）
   2.2 3 個体を親としてランダムに選択
   2.3 UNDX による平行移動ベクトル交叉で子集団の $\overrightarrow{T}$ を生成
   2.4 UNDX と LSQ-fitting を併用した回転交叉で子集団の $\mathcal{M}$ を生成
   2.5 親と子集団からなる家族を評価関数 f_1 と f_2 で評価
   2.6 家族の f_2 評価値を用いて 1 個体をルーレット選択
   2.7 家族の最良 f_1 個体の $\mathcal{M}$，最良 f_2 個体とルーレット選択した 1 個体の $\mathcal{M}$ と $\overrightarrow{T}$ を親 3 個体と置換
   2.8 指定世代数まで 2.2–2.7 を反復
3. 出力:
   3.1 最良 f_1 個体の LFA と RMSD（平均二乗距離の平方根）を出力
   3.2 最良 f_2 個体の GSA と RMSD を出力

図 1 実数値 GA による提案手法の手続き

2.2 交叉方法

FROG は 2 種類の交叉（平行移動ベクトル交叉と回転交叉）を UNDX（Unimodal Normal Distribution Crossover）[6] と最小二乗法（LSQ-fitting）[7] を併用して行う．特に，オイラー角で表現される回転角度は，三次元空間で連続的であるため，親の形質（回転角度）を適切に継承させることが非常に難しい．ここでは，UNDX による原子座標を LSQ-fitting で角度へ変換する交叉方法を設計する．

平行移動ベクトル交叉は，親 3 個体の $\overrightarrow{T}$ から定義される正規分布を利用して，一回の交叉（cross-time）で子 2 個体の $\overrightarrow{T}'$ を生成する．回転交叉は，親 3 個体の代表構造 $\mathcal{Q}^R$ における，第 1 番目の Cα 原子座標（$parent_{\{1|2|3\}}\ \mathcal{Q}_1^R$）を用いる UNDX を実行して，子の第 1 番目の Cα 原子（$child_{\{1|2\}}\ \mathcal{Q}_1^R$）をプロットする．そして，$child_{\{1|2\}}\ \mathcal{Q}_1^R$ と $parent_{\{1|2|3\}}\ \mathcal{Q}_1^R$ の距離比例関係によって，残りの原子座標が自動的に決定される．次に，$child_{\{1|2\}}\ \mathcal{Q}^R$ へ $\mathcal{Q}^R$ を回転させる $\mathcal{M}'$ を LSQ-fitting 法を用いて計算する．そして，$\mathcal{M}'$ は $\mathcal{Q}$ の回転に用いられる．

2.3 評価関数

2.3.1 対応フラグメントペアと対応原子ペア

ある個体が $\mathcal{M}'$ と $\overrightarrow{T}'$ を持っているとき，$\mathcal{Q}$ は $\mathcal{M}'$ によって回転される．$\mathcal{Q}$ の Cα_i→Cα_{i+1} バックボーンを表す長さ 1 の方向ベクトル $\vec{V}_{\mathcal{Q}_{i,i+1}}$ と，$\mathcal{R}$ の Cα_j→Cα_{j+1} バックボーンを表す長さ 1 の方向ベクトル $\vec{V}_{\mathcal{R}_{j,j+1}}$ の差ベクトルの大きさ d_{ij}^{vect} が $\{>\ \mathrm{cf}_1\ \text{or}\ \leq \mathrm{cf}_1\}$ とき，$\{w_{ij}=0 \text{ or } 1\}$ とする $(l_{que}-1)\times(l_{ref}-1)$ 行列 $\Re^{vect}$ を構築する（l_{que} と l_{ref} は，$\mathcal{Q}$ と $\mathcal{R}$ の長さ，cf_1 は閾値）．

次に，回転された $\mathcal{Q}$ は $\overrightarrow{T}'$ によって $\mathcal{R}$ へ重なる．そして，$\mathcal{Q}$ の第 i 番目の Cα（$\mathcal{Q}_i$）と $\mathcal{R}$ の第 j 番目の Cα（$\mathcal{R}_j$）間の距離 d_{ij}^{dist} が $\{>\mathrm{cf}_2\ \text{or}\ \leq \mathrm{cf}_2\}$ とき，$\{w_{ij}=0 \text{ or } 1\}$ とする $l_{que}\times l_{ref}$ 行列 $\Re^{atom}$ を構築する（cf_2 は閾値）．

Smith-Waterman アルゴリズム [8] を用いて，$\Re^{vect}$ に存在する n^{cons} 個以上続く $w_{ij}=1$ の非重複サブパスを求め，対応するベクトルペアを $EQ_k^{vect}(k=1,\ldots,m)$ とする．また，Needleman-Wunsch アルゴリズム [9] を用いて，$\Re^{atom}$ の絶対最適パスを求め，対応する原子ペアを $EQ_{k'}^{atom}(k'=1,\ldots,m')$ とする．

2.3.2 溶媒接触可能性

溶媒接触可能性（Solvent Accessibility，SA）[10] 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に，他の残基の溶媒接触表面に入らない水分子の割合である（$0\leq \mathrm{SA}\leq 1.0$）．ここでは MOLMOL[11] を用いて SA を計算し（$n^{water}=66$），EQ_k^{vect} と $EQ_{k'}^{atom}$ における ΔSA を次のように求める．

$\mathcal{Q}$ の残基 h の SA を $\mathrm{SA}(\mathcal{Q}_h)$ とし，$\vec{V}_{\mathcal{Q}_{h,h+1}}$ の SA を

$$\mathrm{SA}(\mathcal{Q}_h^{vect})=\frac{\mathrm{SA}(\mathcal{Q}_h)+\mathrm{SA}(\mathcal{Q}_{h+1})}{2} \tag{1}$$

とする．EQ_k^{vect} がベクトルペア $\{i,j\}$ のとき，

$$\Delta\mathrm{SA}_{ij}^{vect}=|\mathrm{SA}(\mathcal{Q}_i^{vect})-\mathrm{SA}(\mathcal{R}_j^{vect})| \tag{2}$$

であり，$EQ_{k'}^{atom}$ が Cα 原子ペア $\{i',j'\}$ のとき，

$$\Delta\mathrm{SA}_{i'j'}^{atom}=|\mathrm{SA}(\mathcal{Q}_{i'})-\mathrm{SA}(\mathcal{R}_{j'})| \tag{3}$$

とする．

2.3.3 評価関数 f_1，f_2

$k=\{i,j\}$，$k'=\{i',j'\}$ のとき，各対応ベクトルペアと対応原子ペアの幾何情報による類似度スコアは，

$$s_k^{vect\_d}=exp(-1.0\times\alpha^f\times d_{ij}^{vect}) \tag{4}$$
$$s_{k'}^{atom\_d}=exp(-1.0\times\alpha^f\times d_{i'j'}^{dist}) \tag{5}$$

であり，ΔSA の類似度スコアは，

$$s_k^{vect\_sa}=exp(-1.0\times\beta^f\times\Delta\mathrm{SA}_{ij}^{vect}) \tag{6}$$
$$s_{k'}^{atom\_sa}=exp(-1.0\times\beta^f\times\Delta\mathrm{SA}_{i'j'}^{atom}) \tag{7}$$

である（α^f と β^f は定数）．

LFA 類似度（f_1）と GSA 類似度（f_2）は幾何情報と SA を反映して，

$$f_1^{d+sa}=\frac{\sum_{k=1}^{m}[s_k^{vect\_d}+s_k^{vect\_sa}]}{(l_{que}-1)\times 2} \tag{8}$$

$$f_2^{d+sa-gap}=\frac{\sum_{k'=1}^{m'}[s_{k'}^{atom\_d}+s_{k'}^{atom\_sa}]-gap}{l_{que}\times 2} \tag{9}$$

とする．ここでギャップ n^G 個（$=l_{que}-m'$）に対する

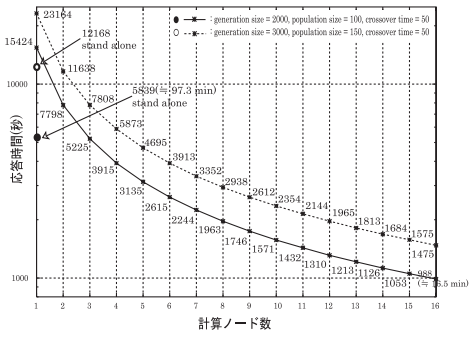


図 2 平均長 270.5 のタンパク質における計算時間

ペナルティは,

$$gap = exp(-1.0 \times \alpha^f \times (\mathrm{cf}_2 + 0.1)) \times n^G \qquad (10)$$

と定義する.

2.4 非同期並列世代交代モデル

FROG は，グリッド RPC システムである Ninf(`http://ninf.apgrid.org/`) を用いて非同期並列化した世代交代モデルを利用する．マスタノードは，重複しない親 3 個体を各計算ノードへ渡し，計算ノードが交叉と選択を行う．あるノードが選択された 3 個体を返すと，マスタノードは世代を更新し，新たな親をそのノードへ渡す.

16 ノード LINUX クラスタにおける提案モデルは，下記のパラメータ環境で平均長 270 のタンパク質ペアを約 15 分で比較し（図 2），全体的に約 7–10 倍の高速化された計算時間を示した [12].

2.5 パラメータセット

実験的･経験的パラメータセットは，世代数=2000, 集団数=100, cross-time=50, UNDX（$\alpha = 0.5$，$\beta = 0.35$），$\mathrm{cf}_1 = 0.56$ と $\mathrm{cf}_2 = 3.5$，$n^{cons} = 4$，$n^{rep} = 10$ であり [13),14)]，$\beta^f = 11.0$，$\alpha^f = 0.5$ に設定する.

3. 実 験 結 果

3.1 2 層比較における FROG の挙動

構造と機能が共通する相同なタンパク質 3 ペア（SCOP（`http://scop.mrc-lmb.cam.ac.uk/scop/`）における all α，all β，$\alpha + \beta$ クラス）を異なる乱数系を用いて 100 試行し，その結果を図 3 に示した.

タンパク質機能に関わるループは，一般に溶液で不安定であるため，座標情報に欠ける場合が多く，正確に比較することは，極めて困難である．図 3a-c が示すように，FROG の 2 層比較は，類似する部分構造と全体構造を一回の比較で発見するため，ループ部位の相違性（LFA の切れ目）とその周辺構造の保存性を LFA レベルで示すと同時に，ループ部位を含めた最適な重ね合わせを GSA レベルで提示する新規性がある.

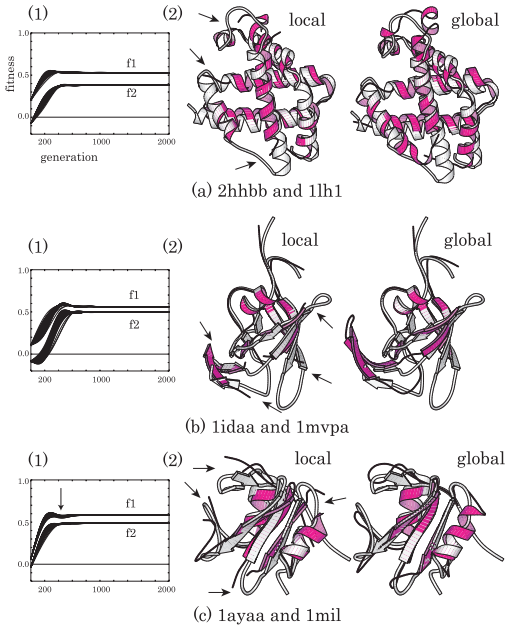


図 3 FROG の 100 試行結果と非線形フラグメント：(1) 集団平均評価値の遷移, (2) 100 試行における最良結果のローカルアラインメント（左）とグローバルアラインメント（右）を “問合せ構造（黒） and 参照構造（灰）” で表示.

3.2 FROG の汎用性と統計的有意性

既存手法と FROG の結果における，%equiv(= $n/l_{que}\times$ 100) と RMSD を用いて（n は対応原子数），

$$S(\%equiv) = \mathrm{F}^e - \mathrm{M}_i^e \qquad (11)$$

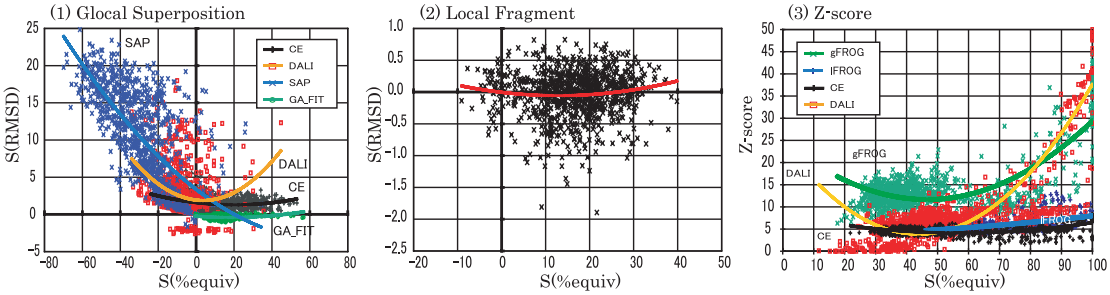
$$S(\mathrm{RMSD}) = \mathrm{M}_i^r - \mathrm{F}^r \qquad (12)$$

を計算する．ここで F^e と F^r は，それぞれ，FROG の%equiv と RMSD．M_i^e と M_i^r は，既存手法 i における%equiv と RMSD である．$S(\%equiv) > 0$, $S(\mathrm{RMSD}) > 0$ は，既存手法に対する FROG の優位性を示す．その際，下式の統計的有意性（Z-score）を CE[15]，DALI[16] と比較実験する.

$$\mathrm{Z-score} = \frac{f - \mu}{\sigma}. \qquad (13)$$

ここで f は現在のペアに対する類似度スコア，μ と σ はそのペアのランダム比較から計算する平均スコアと標準偏差である.

NAD(P)-binding Rossmann fold スーパーファミリー 946 ペアの全体立体構造比較結果に関して（図 4a-1），DALI は 766 ペア，SAP は 876 ペアにおいて，FROG の GSA より多くの対応原子ペアを発見しているが，S(RMSD) は非常に大きく，無意味な重ね合わせが数多く存在することがわかる．そのうえ，CE による 814 ペアの構造比較結果は，明らかに FROG に劣っている．一方，部分構造比較結果について，FROG の LFA は SARF2 と比べ，856 ペアから最大 40%多い対応原子を発見するが，S(RMSD) はわずか 2.0Å 以下の許容範囲にある（図



(a) Homologous / Remote Homologous Protein Pair

図 4 NAD(P)-binding Rossmann fold スーパーファミリーにおける FROG の汎用性と統計的有意性：gFROG は FROG の全体構造比較結果，lFROG は FROG の部分構造比較結果．各分布は二次多項式で近似された．

4a-2）．

図 4a-3 に示すように，FROG は Z-score=4.0 以上の統計的有意な立体構造比較結果を示す．GA はランダムな試行（初期集団）を統計的有意な評価値の解（最良個体）へ，徐々に収束させる探索方法である．従って，統計的有意性が常に高いことは自明である．

4. おわりに

本稿では，タンパク質の Cα バックボーン構造の幾何類似度と溶媒接触可能性に基づく評価関数を設計し，有意な部分構造と最適な全体構造の重ね合わせを一回の実行で発見する，全く新しい手法を設計した．そして，提案手法の新規性と優位性，統計的有意性について実験，考察した．

今後，アミノ酸側鎖情報を考慮した大量のタンパク質ペアの構造比較を行い，基質結合部位・活性部位などのより詳細な部位について比較解析を行いたい．

参 考 文 献

1) Chothia, C.: One Thousand Families for the Molecular Biologist, *Nature*, Vol. 357, pp. 543–544 (1992).
2) Doolittle, R. F.: *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences*, University Science Books (1986).
3) Murzin, A. G.: How Far Divergent Evolution Goes in Proteins, *Current Opinion in Structural Biol.*, Vol. 8, pp. 380–387 (1998).
4) 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).
5) Simons, K. T., Kooperberg, C., Huang, E. and Baker, D.: Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions, *J. Mol. Biol.*, Vol. 268, pp. 209–225 (1997).
6) 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).
7) Hendrickson, W. A.: Transformations to Optimize the Superposition of Similar Structures, *Acta Cryst.*, Vol. A35, pp. 158–163 (1979).
8) Smith, T. F. and Waterman, M. S.: Identification of Common Molecular Subsequences, *J. Mol. Biol.*, Vol. 147, pp. 195–197 (1981).
9) 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).
10) Lee, B. and Richards, F. M.: The Interpretation of Protein Structure: Estimation of Static Accessibility, *J. Mol. Biol.*, Vol. 55, pp. 379–400 (1971).
11) 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).
12) 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).
13) Park, S.J. and Yamamura, M.: Two-layer Protein Structure Comparison, *Proc. of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2003)*, pp. 435–440 (2003).
14) Park, S.J. and Yamamura, M.: GA-based Generic Method for Protein Structure Comparison, *Proc. of the 2003 IEEE Congress on Evolutionary Computation (CEC 2003)*, pp. 1528–1535 (2003).
15) 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).
16) Holm, L. and Sander, C. P.: Protein Structure Comparison by Alignment of Distance Matrices, *J. Mol. Biol.*, Vol. 233, pp. 123–138 (1993).