アルゴリズム 52－5
（1996. 7. 24）

配列類似度に基づく大規模配列集合の高速分類アルゴリズム

岩出智行、石原達也、松田秀雄、橋本昭洋
大阪大学大学院 基礎工学研究科 物理系専攻 情報工学分野

あらまし　分子生物学におけるアミノ酸配列集合の解析手法として，マルチプルアライメントによる解析がある．しかし，マルチプルアライメントは対象となる配列集合が大規模（おおむね 100 以上）であると計算時間が膨大になり，また互いの類似度が高いような配列集合でないと意味のある結果が導き出せないことが知られている．そのため，あらかじめ配列集合をある程度互いの類似度の高い部分集合に分類しておく必要がある．そこで本研究では，配列間の近似的な類似度を計算する **FASTA** 等の結果を用いることを想定した配列分類尺度を提案した．また，その分類尺度を基に，大規模配列集合の分類に耐え得る高速性を持ったアルゴリズムを作成，実装，性能評価を行なった．その結果十分高速で，かつ既存の方法と比較して，よりうまく類縁関係に分類できることがわかった．

Classifying method of large size set of sequences based on pairwise similarities

Tomoyuki Iwade, Tatsuya Ishihara, Hideo Matsuda, Akihiro Hashimoto
Physical Science Course, Graduate School of Engineering Science, Osaka University

ABSTRACT　In molecular biology, several multiple alignment methods have been used for analyzing amino acid sequences. But, these methods are too expensive to align more than 100 sequences and cannot lead to meaningful results if the similarities of sequences are not enough. Therefore, there is a demand for classifying a large size set of sequences into small size sets. In this paper, we propose a method to classify sequences based on an approximate value of pairwise similarity (e.g. **FASTA**'s score). The effectiveness of this method is demonstrated by measuring the performance on classifying hundreds of amino acid sequences.

1　はじめに

分子生物学では，遺伝情報を持つ DNA の塩基配列やそれを翻訳したアミノ酸配列の決定，および解析が盛んに行なわれている．機能解析の手法の一例としては，モチーフの抽出，分子系統樹作成，そしてマルチプルアライメント [1] などが挙げられる．ところが，これらの手法は計算量が大きく，また対象となるアミノ酸配列集合が互いにある程度似通っていないと有用な結果が得られない場合が多い．ゆえに，大規模かつ既存のモチーフを持たないような雑多な配列集合を解析する場合，まず何らかの方法である程度互いに類似した小さな配列集合のグループに分類することが行われている。

このような手法として平均距離法 [2] などの計算量の小さなアルゴリズムを用いたクラスター分析がある．クラスター分析は配列間の類似度などを基に，階層的に類縁の部分集合に分類する手法で，計算量が小さい（$O(n^2 \log n)$ 程度．n は配列数）．ここで類似度としては，２つの配列がなるべく揃うよう

にギャップと呼ばれるすき間を挿入して並べた（これをアライメントという）時の，揃えられたアミノ酸配列間の類似度が良く用いられる．この類似度はダイナミックプログラミングを用いて $O(m^2)$（m は配列長）の計算量で求めることができる [3]（この類似度が生物学的な意味において最適かどうかは議論の余地があるが，以下便宜上「最適な類似度」といえばこのダイナミックプログラミングによって求まった類似度とする）．しかし，大規模配列を対象とする場合においては，データーベースの類似度検索等に用いられる FASTA[4] 等の高速（$O(m)$）に近似度を求めることのできるアルゴリズムを用いる方が望ましい．ところが，一般のクラスター分析においては類似度が近似値である場合について考慮されていない．また階層的手法では 2 種類以上のモチーフを持つ配列を含む集合を分類する場合に，モチーフを共有する複数の集合と個別に関連付けられるのではなく，あるひとつの集合との関連しか見つけられないため，同一モチーフを持つ集合の関連が見えてこない，という望ましくない事態が起こり得る．

本研究では，まず，

1. 配列間の類似度がある特性を持った近似アルゴリズムによって計算されていることを想定
2. 複数のモチーフを持つ集合をモチーフごとに重なりを持つような部分集合へ別々に分類できるよう考慮

するような配列分類尺度を提案した．ここで 1 のある特性とは以下のとおりである．

- ある確率 p $(0 \leq p \leq 1)$ で最適な類似度を返す．
- 最適値より大きな値を返すことはない．

この特性を「類似度信頼率が p である」と定義する．FASTA をはじめとする多くの近似アルゴリズムは，最適値を与えるアライメントと同じ部分をうまくアライメントできればほぼ同等の値を返すが，違う部分をアライメントしてしまうと大きく下回った値を返すという性質を持つ，ゆえに上記の特性を仮定することは妥当であると考えられる．

提案する分類尺度の基本方針は，信頼率が p であるような配列間の類似度と，ある定数 b が与えられた時に以下を満たす部分集合で，与えられた配列集合を被覆することである．

- 集合内の任意の 2 配列間の類似度が，確率 p で b 以上の値をとる．
- 要素数が極大である．すなわち，任意の集合外の配列を加えると上の条件を満たさない．

本研究では，上記の方針に基づいた分類尺度を 3 節で提案し，それが NP 完全であることを 4 節で示す．5 節ではその分類尺度に対して近似解を求める，平均時間計算量 $O(n^3 \log n)$，必要空間量 $O(n^2)$(n は配列数) のアルゴリズムを示す．最後に 6 節ではそのアルゴリズムの性能を，平均距離法との比較を交えながら行う．それにより本アルゴリズムが，

- 配列間の類似度が FASTA 等の近似アルゴリズムで計算されたものであっても比較的良く類縁の関係に分類出来ること
- 平均距離法ではうまく分類できなかった配列集合をうまく分類する能力を持つこと

を示す．

2 諸定義

- $G = (V, E)$ は，頂点集合が $V(|V| = n)$，辺集合が E であるようなグラフ．各頂点は $v_i, (1 \leq i \leq n)$，頂点 v_i, v_j 間の辺は $\{v_i, v_j\}$ と表記する．
- seq_i $(1 \leq i \leq m)$ はアミノ酸配列．Seq はアミノ酸配列集合，すなわち $Seq = \{seq_i \mid 1 \leq i \leq m\}$
- sim_{ij} はアミノ酸配列 seq_i, seq_j 間の類似度．なお，$sim_{ii} = \infty$ と定義する．Sim はアミノ酸配列間の類似度の集合，すなわち $Sim = \{sim_{ij} \mid 1 \leq i, j \leq m, i \neq j\}$
- 集合 Sに対し，$\rho(S)$ は S のすべての部分集合の集合 (巾集合)．

3 アミノ酸配列集合の分類尺度

この節では本研究で提案する，ある特性を持った近似アルゴリズムによって計算された類似度を基にアミノ酸配列を分類するための尺度を紹介し，その特性をいくつか示す．

まず、p-準完全グラフを定義する。

定義 1 任意の頂点の次数が $\lceil p(|V|-1) \rceil$ 以上であるようなグラフ (V, E) を，p-準完全グラフと呼ぶ．

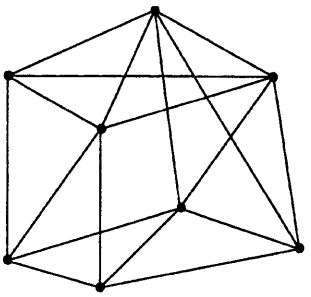

図 1: 0.5-準完全グラフの例. 頂点の次数$\geq \lceil 0.5 \cdot (8-1) \rceil = 4$

以下にこの p-準完全グラフを用いた, 配列集合 Seq, 配列間の類似度集合 Sim, 基準類似度 b, 類似度信頼率 $p(0.5 \leq p \leq 1)$ が与えられたときの配列分類尺度を示す.

1. 各配列を頂点とし, 類似度が b 以上の場合に辺を引いたグラフ (V, E) (v_i は $seq_i \in Seq$ に対応. $E = \{\{v_i, v_j\} \mid sim_{ij} \geq b, i \neq j\}$) を考える.
2. まず, グラフ (V, E) の部分グラフかつ要素数が極大な p- 準完全グラフを構成する頂点集合の集合 ($S_p \subset \rho(V)$ とする) を求める.
3. 次に, 以下を満たす S_pの部分集合 $Cover_p = \{C_i \mid 1 \leq i \leq k, C_i \in S_p\}$ を求める.
   - $\bigcup_{1 \leq i \leq k} C_i = V$
   - 上記を満たすもののうち, kが最小

この分類尺度に基づき求まった集合 $Cover_p$を配列集合 Seqの, 類似度信頼率 p, 基準類似度 b における分類と呼ぶ.

ここで, 集合 $C_i \in Cover_p$は, p-準完全グラフ, つまり集合内の任意の配列と集合内の他の配列との類似度が b 以上である割合が p 以上であることを意味している. これは集合内の任意の 2 配列間について, おおよそ類似度が信頼出来る部分だけを見たときに, その部分に関して類似度が b 以上であるような集合を選び出して来ることを意図している.

また, 分類尺度の定義において, 類似度信頼値 p の範囲は $0.5 \leq p \leq 1$ に限っているが, これは $0 \leq p < 0.5$ である時に全く類似性の見られない 2 つの集合を結合した部分集合を結果として出す可能性があるためである。

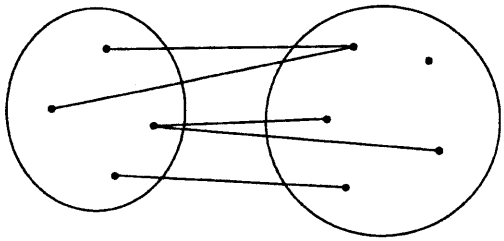

図 2: $|V| = 9$ なる p-準完全グラフの例. $0.5 \leq p \leq 1$ ならばどんな 2 分割に対しても, 集合間に 4 本以上の辺が存在し, 要素数の小さい方の集合の全ての要素から他方の集合内へ少なくとも 1 本以上の辺が伸びている.

一方, $0.5 \leq p \leq 1$ であれば, 類似性を持つ配列が 2 つ以上の集合に分離しにくいことがわかっている (図 2). つまりある程度類似度の点でまとまりがあると言える.

尚, 次節で示すように, 頂点数最大の 0.5-準完全グラフを求める問題は我々により NP 完全であることが証明できている. それ以外 $(0.5 < p < 1)$ についてはまだ証明できていないが, おそらくNP 完全であると予想される.

4 0.5-準完全グラフ問題の NP 完全性の証明

以下の証明の手順は文献 [5] に従っている.
まず, 定義を行なう.

定義 2 以下の問題を **0.5-準完全グラフ問題**と定義

インスタンス

グラフ (V, E), 正定数 k

質問

グラフ (V, E) が, サイズ k 以上の 0.5-準完全グラフを部分グラフとして持つか.

定義 3 以下の問題を制限つきクリーク問題と定義

インスタンス

グラフ (V, E), 定数 $k \geq \left\lfloor \frac{|V|}{2} \right\rfloor + 1$

質問

グラフ (V, E) が, サイズ k以上のクリークを持つか.

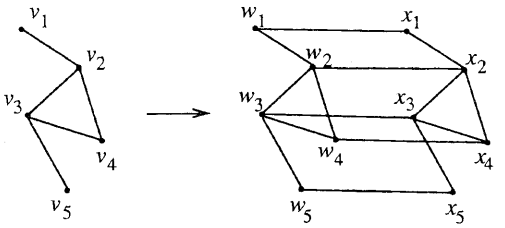

図 3: 制限つきクリーク問題 $((V,E),k)$ から，0.5-準完全グラフ問題 $((V_1,E_1),k_1)$ への変換例

制限つきクリーク問題は，NP 完全問題のひとつであるクリーク問題のインスタンスに制限を加えたものであり，NP 完全である．

以下，この制限つきクリーク問題を用いて，0.5-準完全グラフ問題が NP 完全であることを示す．

定理 1 *0.5-準完全グラフ問題は NP 完全である．*

[証明]

ある一つの部分集合 $V' \subset V$ が与えられた時，それがサイズ k 以上の 0.5-準完全グラフであるかどうかの判定は明らかに多項式時間で可能である．よって 0.5-準完全グラフ問題はクラス NP に属する．次に，制限つきクリーク問題が，0.5-準完全グラフ問題へ多項式時間還元可能であることを示す．

まず，以下のように，制限つきクリーク問題のインスタンス $((V,E),k)$ を，0.5-準完全グラフ問題のインスタンス $((V_1,E_1),k_1)$ に変換する．

$$
\begin{aligned}
V_1 &= W \cup X \\
&\quad (|W| = |X| = |V|, \\
&\quad\ w_i \in W, x_i \in X \text{はそれぞれ } v_i \in V \text{に対応}) \\
E_1 &= \{\{w_i, w_j\}, \{x_i, x_j\} \mid \{v_i, v_j\} \in E\} \\
&\quad \cup \{\{w_i, x_i\} \mid 1 \le i \le |V|\}\} \\
k_1 &= 2k
\end{aligned}
$$

この変換は明らかに多項式時間で行なえる．

ここで，制限つきクリーク問題が Yes インスタンスを持つ時，またその時に限って 0.5-準完全グラフ問題が Yes インスタンスを持つということを示す．

(1) 制限つきクリーク問題において，ある $((V,E),k)$ が Yes インスタンスであるとき

このとき，あるサイズ k 以上のクリーク V' が存在する．ここで，0.5-準完全グラフ問題への変換後のグラフ (V_1,E_1) において，以下のような部分グラフ (V_1',E_1') を考える．

$$
\begin{aligned}
V_1' &= W' \cup X' \\
&\quad (|W'| = |X'| = |V_1'|, \\
&\quad\ w_i \in W', x_i \in X' \text{はそれぞれ } v_i \in V_1' \text{に対応}) \\
E_1' &= \{\{v,w\} \mid v,w \in V_1', \{v,w\} \in E_1\}
\end{aligned}
$$

ここですべての $v \in V_1'$ は $|V'| - 1 + 1 = |V'| = |V_1'|/2 \ge \lceil 0.5 \cdot (|V_1'| - 1) \rceil$ 本の辺をグラフ (V_1',E_1') 内に持つ．一方，$|V_1'| = 2|V'| \ge 2k = k_1$ である．

以上より，グラフ (V_1',E_1') はサイズ k_1 以上の 0.5-準完全グラフである．つまり，変換後のインスタンス $((V_1,E_1),k_1)$ は 0.5-準完全問題における Yes インスタンスである．

(2) 0.5-準完全グラフ問題において，ある $((V_1,E_1),k_1)$ が Yes インスタンスであるとき

このとき，頂点数 k_1 以上の 0.5-準完全グラフ (V_1',E_1') が存在する．ここで，

$$
\begin{aligned}
|V_1'| &\ge k_1 \\
&\ge |V| + 1
\end{aligned}
$$

より，V_1' は集合 W,X の両方の要素を少なくとも 1 つ以上含む．よって，

$$
\begin{aligned}
V_1' &= W' \cup X' \ (W' \subset W, |W'| \ge 1, \\
&\quad\ X' \subset X, |X'| \ge 1, |W'| + |X'| = |V_1'|)
\end{aligned}
$$

とおく．ここで，$|W'| \le |X'|$ として一般性を失わない．すると，

$$\lceil 0.5 \cdot (|V_1'| - 1) \rceil \ge |W'| \qquad (1)$$

が成立する．また，$w_i \in W'$ の次数 (m とする) はグラフ (V_1',E_1') が 0.5-準完全グラフであることより

$$m \ge \lceil 0.5 \cdot (|V_1'| - 1) \rceil \qquad (2)$$

であり，式 (1)(2) より

$$m \ge |W'| \qquad (3)$$

が成立する．つまり，少なくとも一本以上の辺が集合 X内の頂点に伸びている．一方 w_iから集合 X内の要素への辺は，x_iに対しての辺 $\{w_i, x_i\}$ のみである (*) から，

$$m \leq |W'| \tag{4}$$

である．よって式 (3)(4) より $m = |W'|$ であり，(*) をあわせて考えると，W'内の頂点は W'内の他の頂点に対してちょうど $|W'|-1$ 本の辺が伸びている，すなわち W'はクリークである．また，式 (1)(2)(3)(4) より，

$$\begin{aligned} |W'| &= \lceil 0.5 \cdot (|V_1'|-1) \rceil \\ &= k \end{aligned}$$

であるから，W'はサイズ k以上のクリークである．すなわち，変換元の制限つきクリーク問題において $V' = \{v_i \mid w_i \in W'\}$ はサイズ k以上のクリークであり，変換元のインスタンスは Yes インスタンスである．

以上より，制限つきクリーク問題が，0.5-準完全グラフへ多項式時間還元可能であることが示せた．

よって，0.5-準完全グラフ問題は NP 完全である．

[証明終わり]

5 近似分類アルゴリズムの概要

4 節で証明したように，0.5-準完全グラフ問題は NP 完全である．ゆえに，$0.5 \leq p \leq 1$ であるときに，p-準完全グラフ問題を解くためにはなんらかの近似的アプローチが必要であると考えられる．この節では，p-準完全グラフ問題に対する近似アルゴリズムについて述べる．

まず，定義を行う．

定義 4 配列集合 Seq内のある要素 seq_iについて，$sim_{ij}, (seq_j \in Seq - \{seq_i\})$ を降順に並べた時，$\lceil p(|Seq|-1) \rceil$ 番目の値を $Sim_p(seq_i)$ と定義．

定義 5 割合 p が与えられた時，$eval(Seq, p) = min_{seq_i \in Seq} Sim_p(seq_i)$ を配列集合 Seqの評価値と定義．

これより，配列集合 Seqについて以下は同値となる．

- 各配列を頂点とし，類似度が b 以上の場合に辺を引いたグラフ (V, E) (v_i は $seq_i \in Seq$ に対応．$E = \{\{v_i, v_j\} \mid sim_{ij} \geq b,\ i \neq j\}$) が p-準完全グラフである．
- $eval(Seq, p) \geq b$

つまり，評価値 $eval(Seq, p)$ は類似度信頼率が p である場合において，配列集合 Seqがどの程度類縁集合であるかを表していると考えられる．

5.1 近似分類アルゴリズム

入力

配列集合 Seq, 配列間の類似度集合 Sim, 基準類似度 b, 類似度信頼率 $p(0.5 \leq p \leq 1)$.

アルゴリズム

[ステップ 1]

まず，全ての $seq_i \in Seq$について以下を行う

i. 初期集合を $M_i = \{seq_i\}$ とする．

ii. $Seq - M_i = \phi$ならば終了．そうでないならば次へ．

iii. 以下の手順で，$Seq - M_i$から配列 s を一つ選び出す．

(a) すべての $t \in Seq - M_i$について，$sim(t, u_k)$ $(u_k \in M_i, 1 \leq k \leq |M_i|)$ を降順に並べた時の $\lceil p|M_i| \rceil$ 番目の値を求め，それを t と M_iとの近縁度とする．

(b) $Seq - M_i$の中から，(a) で求めた近縁度の高いものを上位 20 個まで取り出す．($T = \{t_1, \cdots, t_{20}\}$ とする)

(c) $eval(\{t_j\} \cup M_i, p), (1 \leq j \leq 20)$ が最も大きくなるような t_j を s とする (2 つ以上ある場合はその内のどれか)．

iv. $eval(\{s\} \cup M_i, p) \geq b$ なら $M_i \leftarrow M_i \cup \{s\}$ として ii. へもどる．そうでないなら終了．

[ステップ 2]

次に，ステップ 1 で得られた集合 M_iの集合 $S_p' = \{M_1, \cdots, M_{|Seq|}\}$ について以下を行い，最終結果である集合の集合 $Cover_p'$を得る

i. $seq_i \in Seq, (1 \leq i \leq |Seq|)$ に対し，bool 関数 $masked(seq_i)$ を用意し，$masked(seq_i) \leftarrow F$ と初期化しておく．ま

た，$unmaskedseq(M_i), (1 \leq i \leq |Seq|)$ を $|\{seq \mid seq \in M_i, masked(seq) = F\}|$ と定義する．

ii. $Cover'_p = \phi$と初期化する．

iii. $max_{M_i \in S'_p}\{unmaskedseq(M_i)\}$ を与える M_i (2つ以上ある場合は，要素数の多い方，それでも2つ以上ある場合はその内のどれか) を $Cover'_p$の要素として加え，$S'_p \leftarrow S'_p - \{M_i\}$ とし，M_iに属するすべての配列 seq について，$masked(seq) \leftarrow T$とする．

iv. $1 \leq i \leq |Seq|$ であるすべての i について，$masked(seq_i) = T$ ならば終了，そうでないならばiii. へ．

5.2 近似アルゴリズムの解説

ここでは，5.1節のアルゴリズムの性質等について述べる．ステップ1は，分類尺度の定義における S_p(極大な p-準完全グラフの集合) を求める部分に対応した近似アプローチである．ここで求まる集合は，極大性については妥協している．しかし，評価値が b 以上すなわち p-準完全グラフであるという条件は守られている．

ステップ1の手順の概要は，1) Seq内の個々の要素から出発して，2) 集合の評価値がなるべく高い値を維持するように，3) 要素を一つずつ結合していく，ことである．ここで，

- 1) により，最大 $|Seq|$ 個の集合を作りだすことが可能である．
- 2) は，集合を大きくしていく際に，なるべくその集合に近縁な配列を加えていくことを表している．要素間の類似度の相関関係は多くの場合ランダムではないことから，より要素数の多い集合を得る戦略としてある程度妥当であると考える．
- 2) は，評価値が「なるべく高く」であり，「一番高く」ではない．これはステップ1の iii. に於いて，(c) における計算量を軽減するためにまず (a)(b) によって候補を絞るからである．集合 M_iにある配列 seqを加えた場合の集合の評価値の変化は，$Sim_p(seq)$ による影響が大部分を占めると考えられるので，この絞り込みによる精度への影響はかなり小さいのではないかと考えられる．
- 3) の部分について，実際は，「どんな一つの要素を付け加えても評価値が b を下回るが，ある2つ以上の要素を同時に付け加えた場合に評価値が b を越える」場合は存在する (図4)．しかし，本アルゴリズムでは，計算量の観点から1つずつ結合していく方法を取ることにした．これにより極大な集合は得られなくなるが，例えば図4のような場合に集合 A と集合 $\{c_1, c_2, c_3\}$ に分けてしまうこと自体は大きな問題ではないと考えている．

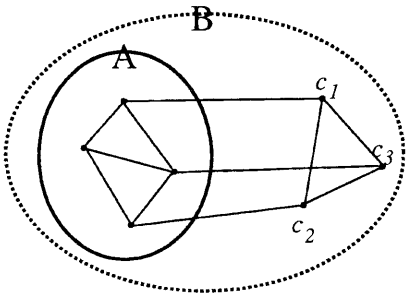

図 4: 類似度信頼率は 0.5，辺が存在する頂点間のみ類似度は b 以上とする．集合 $A \cup \{c_i\}(1 \leq i \leq 3)$ の評価値は b 未満であるが，集合 Bの評価値は b 以上．

ステップ2は，ステップ1によって得られた近似極大集合の集合から最小被覆を求める部分に対応した近似アプローチである．この方法では必ずしも最小被覆が求まるわけではない．しかし，出来るだけ要素数が最大に近い集合で与えられた配列集合を覆いたいという，最小被覆を選ぶことにした目的には合致している．

5.3 時間計算量，必要空間量

時間計算量について，$n = |Seq|$ とすると，ステップ1での計算量は平均 $O(n^3 \log n)$ で，ステップ2では平均 $O(n^3)$ となる．一方，必要空間量は，全体で $O(n^2)$ となる．

6 アルゴリズムの性能評価

5.1 節で示したアルゴリズムを実装し評価を行った.

6.1 類似度信頼率について

グロビン族タンパク質から 236 本, チトクローム族タンパク質から 627 本をタンパク質データーベース **SWISS-PROT**[6] から選びだし, それぞれについて配列間類似度を FASTA Ver2.0 の **opt スコア** と, Smith-Waterman [7] のローカルアライメントスコア (以下, **S-W スコア**) を比較した. その結果, i) FATSA はある程度の割合で最適値に近い値を返すこと, ii) 特に S-W スコアの高い配列間 (およそ 200 以上) では比較的良い結果が得られること, iii) 他方 S-W スコアが 100 より低いような配列間では, おおよそ 20 から 40 % の組合せについて, スコアが著しく下がる (S-W スコアの半分程度になる) ことがわかった.

以上のことから, 基準類似度を 100 以上に設定し, また 類似度信頼率を 0.6 ～ 0.8 程度に設定することができそうである.

6.2 平均距離法との比較

モチーフデータベース **PROSITE**[8] から, モチーフ DHFR(ジヒドロ葉酸レダクターゼ) 持つタンパク質と, モチーフ THYMIDYLATE_SYNTHASE (チミジル酸シンターゼ) を持つタンパク質を探しだし, そのアミノ酸配列をタンパク質データベース SWISS-PROT から得た (61 本). 得られたアミノ酸配列集合は以下のようなものであった.

- モチーフ DHFR のみを含む配列集合 (SWISS-PROT のエントリ ID が DYR で始まる. 以下 **DYR** と呼ぶ)
- モチーフ THYMIDYLATE_SYNTHASE のみを含む配列集合 (同エントリ ID が TYS で始まる. 以下 **TYS** と呼ぶ)
- モチーフ DHFR と THYMIDYLATE_SYNTHASE を両方含む配列集合 (同エントリ ID が DRT で始まる. 以下 **DRT** と呼ぶ)

一般に同一モチーフを持つものは互いに類似度が高いこともあり, これらを分類する場合, モチーフごとに分類されることが望ましい.

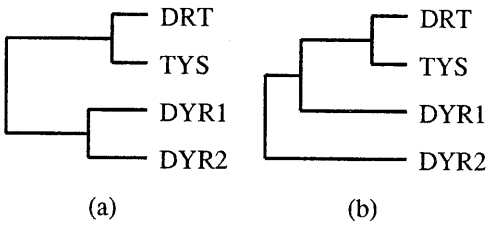

図 5: (a) S-W スコアと平均距離法による分類
(b) FASTA スコアと平均距離法による分類

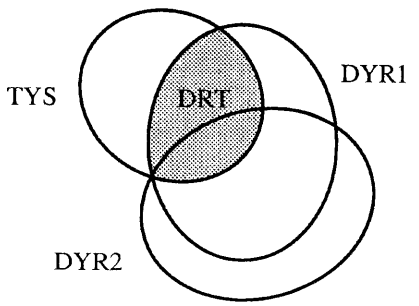

図 6: FASTA スコアと本アルゴリズム ($b = 100$, $p = 0.7$) による分類

この配列集合に対して以下の 4 通りの方法で分類を行なった.

1. 類似度計算は Smith-Waterman ローカルアライメントアルゴリズム, 分類は平均距離法.
2. 類似度計算は FASTA, 分類は平均距離法.
3. 類似度計算は Smith-Waterman ローカルアライメントアルゴリズム, 分類は本アルゴリズム (基準類似度 $b = 100$, 類似度信頼率 $p = 1$ と設定)
4. 類似度計算は FASTA, 分類は本アルゴリズム (基準類似度 $b = 100$, 類似度信頼率 $p = 0.7$ と設定)

この結果, 1. の場合は図 5(a), 2. の場合は図 5(b) のような結果が得られた. 一方, 4. の場合は図 6 のような結果となった. また 3. の場合でも 4. と同様の分類結果を得ることができた. これらの結果をまとめると以下のようになる.

- 平均距離法では, DRT と, DYR が類似している (同一モチーフを持っている) ことが全

く見えてこなかった．これは平均距離法に限らず，木構造を階層的に構築する分類手法に一般的に当てはまると考えられる．

- 2. では，DYR の結合関係が 1. と異なる結果となった．これは以下が原因と考えられる．i) DYR 内のある配列集合 (DRY1) がそれ以外の配列集合 (DRY2) との間の類似度が 100 前後と低く，FASTA では取りこぼしが多いこと，ii) TYS と DYR1 間の類似度 (かなり低い) と，DRT と DYR1 間の類似度 (ある程度大きい) の平均がたまたま 100 を越えたこと．
- 3. および 4. では，DRT が TYS と DYR の両方に類似していることが示されている．一方，DYR と TYS が類似していないことも示されている．

以上のことから，本アルゴリズムが以下の能力を持つことを示せた．

- 配列間の類似度として一部最適解より大幅に劣るような値を与えられてもうまく補完したかのように分類を行う．
- ある配列が 2 種類以上のモチーフを含むような配列集合を分類する場合にモチーフごとに分類を行う．

6.3 大規模配列集合の分類における時間評価

次にチトクローム族のタンパク質 627 本に対して，FASTA Ver2.0 で類似度を計算し，本分類アルゴリズムで基準類似度 100，類似度信頼度 0.6 に設定して分類を行なった．

尚，今回は分類アルゴリズムを C++でインプリメントしたものを，また FASTA と Smith-Waterman アルゴリズムについては FASTA Ver2.0 パッケージを，NEWS-5000，NEWS-OS4.2 上で，gcc-2.5.8 でコンパイルして利用した．

- 本分類アルゴリズムによる分類…約 7 分
- FASTA による類似度計算…約 16 分
- (参考)Smith-Waterman ローカルアライメントアルゴリズムによる類似度計算…約 5 時間

本分類アルゴリズムの計算時間は，分類で得られる集合のサイズに依存する．そこで，基準類似度 0 にして実行してみると，約 50 分かかった．しかし，それでも類似度の計算を含めて 1 時間程度で分類が可能である．また，分類結果については省略するが，このような大規模配列集合の場合でもモチーフごとに分類を行うことが可能であることが確かめられた．

以上より，数 100 本クラスの大規模配列集合を，類似度の計算も含めてかなり高速に，ほぼ所持モチーフごとに分類する能力を持つことが判明した．

7 おわりに

本論文では，配列間の類似度がある条件を満たす近似値であることを前提とした，大規模アミノ酸配列集合の分類尺度を提案し，それに基づくアミノ酸配列分類アルゴリズムを作成，性能評価した．

その結果，平均距離法など木構造を構成する分類アルゴリズムが不得手であった 2 種類以上のモチーフを持つアミノ酸配列を含む集合の分類に関して，モチーフごとに分類する能力を持つことを分かった．また，配列間の類似度を FASTA 等の高速な近似アルゴリズムで計算することができ，より多数の配列に対して高速に，互いの類似度が高いと思われる部分集合に分類できることがわかった．

以上よりこの手法は，大規模配列集合の解析の前処理として有用であると考える．

参考文献

[1] S. C. Chan, A. K. C. Wong, and D. K. Y. Chiu: "A Survey Of Multiple Sequence Comparison Methods," Bull. Math. Biol., Vol.54, No.4, pp.563-598(1992).

[2] P. H. A. Sneath, and R. R. Sokal, "Principles Of Numerical Taxonomy," W. H. Freeman and Company(1973).

[3] S. B. Needleman, and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequences of two proteins," J. Mol. Biol., Vol.48, pp.443-453(1970).

[4] W. R. Pearson, and D. J. Lipman, "Improved tools for biological sequence comparison," Proc. Natl. Acad. Sci. USA, Vol.85, pp.2444-2448(1988).

[5] M. R. Garey and D. J. Johnson, "Computers and intractability : a guide to the theory of NP-completeness," W. H. Freeman and Company(1979).

[6] A. Bairoch , and R. Apweiler, "The SWISS-PROT protein sequence data bank and its new supplement TREMBL," Nucleic Acids Res. Vol.24, No.1, pp.21-25 (1996).

[7] T. F. Smith, and M. F. Waterman, "Identification of common molecular subsequences," J. Mol. Biol., Vol.147, pp.195-197(1981).

[8] A. Bairoch, P. Bucher, and K. Hofmann, "The PROSITE database, its status in 1995," Nucl. Acids Res. Vol.24, No.1, pp.189-196(1996).