2008－BIO－12 (4)
2008／3／3

局所類似構造を用いた蛋白質機能部位発見に関する研究

今田 圭亮 † 尾崎 知伸 †† 大川 剛直 ‡

† 神戸大学大学院 自然科学研究科 †† 神戸大学 自然科学系先端融合研究環

‡ 神戸大学大学院工学研究科

近年，機能未知の蛋白質を対象に類似構造を探索することによる，新たな機能部位を自動抽出する試みが行われている．機能部位の多くは，凹構造を持つことが知られているが，その大きさは様々であり，直接的に比較することは困難である．そこで本論文では，各凹構造を細分化し最適な部分において比較する．しかし機能部位によって大きさが異なるため，細分化後大きさを変化させることでより柔軟な探索法を提案する．また実験により，提案手法により機能部位抽出の精度向上が確認された．

Extraction of functional sites in proteins by the search of similar local structures

Keisuke IMADA † Tomonobu OZAKI †† Takenao OHKAWA ‡

† Graduate School of Science and Technology, Kobe University

†† Organization of Advanced Science and Technology, Kobe University

‡ Graduate School of Engineering, Kobe University

There are many researches on the automatic extraction of new functional sites in proteins by the search of structures which are similar to the known functional sites.
While many functional sites consist of concave structures, it is difficult to compare those concaves directly due to the various sizes of concaves. To cope with these difficulties and to realize the detailed comparison between concaves, we propose a method of searching and comparing concaves by changing the size gradually. Through the experiments, we confirm that the accuracy for the extraction of functional sites is improved.

1 はじめに

蛋白質の機能解析は生体のメカニズムを解明する上で重要な研究領域である．しかし，実験による解析には膨大なコストがかかるため，計算機を用いた解析に期待が寄せられている．蛋白質の情報にはアミノ酸配列の情報や，構成原子の空間上の位置や特徴を示した立体構造情報があり，様々な視点から解析がなされている[1, 2]．

蛋白質の機能に関与する部位は機能部位と呼ばれており，これらの特定は機能解析の糸口になると考えられる．機能部位の中には他の蛋白質や化合物と結合することで発現するものもあり，結合部位や表面に注目した解析が有効であると考えられている[5, 7]．

例えば，既知の結合部位に類似する構造を探索することにより，新たに与えられた蛋白質の機能部位を推定することや，ある蛋白質ファミリ内で共通する類似部位を発見することにより，機能部位を特定することが可能となる．このとき，機能部位は蛋白質によって，大きさが異なるために，事前にそのサイズを決定しておくことが困難であり，類似性を評価する場合にも，どの範囲を評価対象にするかが問題になる．

そこで，本研究では表面データから結合部位の特徴を基に候補を抽出し，再度全領域において一定範囲ごとに部位近傍を抽出し類似性を評価する．これにより，候補部位における機能部位の位置や範囲に柔軟に対応した比較を図る．さらに，提案した類似性評価手法を用いてファミリ内で類似する局所表面を探索することにより，機能部位を特定する手法を定式化する．また，この手法を蛋白質の分類問題に対して適用した結果についても議

論する．

以降，2でポケットの有効性について，3で提案手法について，4で評価実験と考察について，5でまとめと今後の課題について述べる．

2 ポケットの有効性

蛋白質の機能に関与する部位の多くは，ファミリと呼ばれる同機能を持つ蛋白質群に共通して出現することが知られている．またこれらの部位は，類似した形状や特性を持っていることが分かっている．

一方で，例えば，代表的な酵素蛋白質であるセリンプロテアーゼは，触媒として働く際，機能部位付近の非極性ポケットに他の蛋白質が結合するといった特徴がある．このように，ある種の蛋白質においてポケットは機能部位抽出における重要な手がかりであることが知られている[4]．

そこで，ファミリ内における類似部位を抽出することによって機能部位を発見することを考える．しかしながら，立体構造データにおいて蛋白質全体を対象にそのような部位を探索するには膨大なコストがかかる．そこで，結合部位の特徴である凹構造，すなわちポケットを結合部位の候補として抽出し，ポケットのみを類似性比較の対象とする．以降，ポケットをモチーフ候補，類似する部位をモチーフと呼ぶことにする．

3 蛋白質のモチーフ抽出と分類

3.1 モチーフ抽出の概要

本節では蛋白質のモチーフ抽出方法の概要について述べる．本研究ではモチーフ及びモチーフ候補を表面データを用いて抽出する．表面データは表面を形成する頂点の位置や物性値，また頂点間の情報を持っている．表面データはef-site[1]に登録されているものを用いた．

モチーフ候補 a と蛋白質 B が持つモチーフ候補の集合 $mc(B) = \{b_1, b_2, ..., b_{n_B}\}$ において，最も類似するモチーフ候補 a，b_i のペアを $mcp(a, B)$ と表記する．モチーフ候補間の比較の具体的な方法に関しては後で述べる．次に蛋白質ファミリ内で類似している部位を抽出することを考える．n 個の蛋白質で構成されるファミリ $f = \{P_1, P_2, ..., P_n\}$ からモチーフ群を抽出する．まず2つの蛋白質 A，B 間において，お互いの蛋白質間類似モチーフ候補ペア集合を $pair(A, B)$ と表記する．すなわち，

$$pair(A, B) = \{\langle a, b\rangle | a \in mc(A), b = mcp(a, B)\} \quad (1)$$

ファミリ F 内の任意の2つの蛋白質間における類似モチーフ候補ペア集合を求め，これをもとにファミリ内類似モチーフ候補ペアの集合 $M(f)$ を得る．

$$M(f) = \bigcup_{A,B \in f} pair(A, B) \quad (2)$$

次にファミリ f 中の各蛋白質 P_i から1つずつモチーフ候補が組み合わせてできた，モチーフ候補群 $S(f) \in mc(P_1) \times mc(P_2) \times ... \times mc(P_n)$ を考える．ここでモチーフ候補群のスコア $score(S)$ を以下のように決定する．このスコアの大きいモチーフ候補群ほどモチーフ群である可能性が高いと考えられる．Fig. 1 にモチーフ抽出の概要を図示する．

$$score(S) = |\{\langle x, y\rangle \in S | \langle x, y\rangle \in M(f)\}| \quad (3)$$

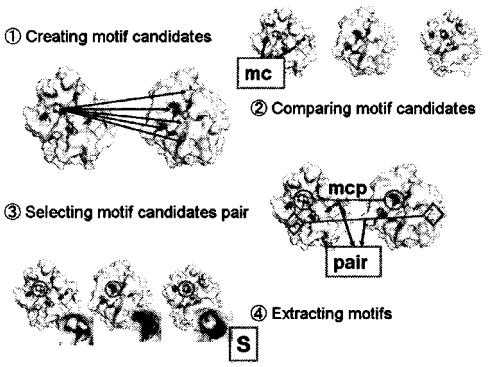

Fig. 1 Flow chart

3.2 モチーフ候補の抽出

本研究では表面データの曲率をもとにポケット部位を抽出する[8]．頂点の持つ曲率情報には近傍の最大曲率 k_{max} と最小曲率 k_{min} があり，これらをを用いてガウス曲率 K，平均曲率 H を以下のように定義する．

$$K = k_{max} k_{min} \quad (4)$$

$$H = \frac{1}{2}(k_{max} + k_{min}) \quad (5)$$

[1] http://ef-site.hgc.jp/eF-site/[3]

K, H の値によって Table 1 に示す 8 種類の曲面形状のうちのどの形状に属するかが決定される.

Table 1 Curved surface shape by Gaussian curvature and mean curvature

	$K<0$	$K=0$	$K<0$
$H>0$	concavity	valley	saddle (valley)
$H=0$		plain surface	saddle (equal)
$H<0$	convexity	crest	saddle (crest)

そして，ポケットに相当する凹型の形状に属する頂点群の集合を領域拡張法 [8] に基づき抽出し，これらをモチーフ候補と定義する．Fig. 2 は上記の方法によって抽出されたモチーフ候補群の一例である．なお，実際にモチーフ候補を抽出する際には，機能部位になる可能性の低い小さいポケットを削減するために頂点数の下限を設定する．また，蛋白質の内部には空洞が存在し，これらは表面上に現れないためここでは除外して考える．空洞形状の領域は，表面のポケット構造に比べて，サイズが大きな凹構造であるため，頂点数に上限を設けることで，これに対処する．

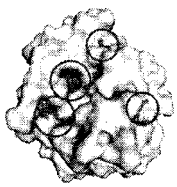

Fig. 2 candidates of motif(1fn8)

3.3 類似モチーフ候補ペア

ここでは蛋白質間における類似モチーフ候補ペアの抽出について説明する．機能部位は蛋白質によって，大きさが異なるために，事前にそのサイズを決定しておくことが困難である．従って，機能部位を比較しながら類似性を評価する場合にも，どの範囲を評価対象にするかが問題になる．

3.3.1 類似モチーフ候補ペア抽出の概要

本研究ではモチーフ候補を構成する各頂点の近傍頂点集合を用いて比較することで，モチーフ候補間における適当な類似部分における比較を実現する．k 個の頂点で構成されているモチーフ候補 $m_1 = \{u_1, u_2, ..., u_k\}$ と l 個の頂点で構成されている $m_2 = \{v_1, v_2, ..., v_l\}$ に対するモチーフ候補間の非類似度 $msim(m_1, m_2)$ を求めるのための手順を以下に示す．Fig. 3 に蛋白質の持つ要素を図示する．

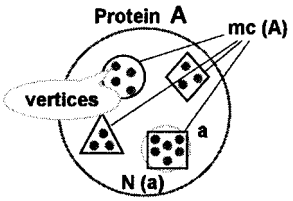

Fig. 3 Protein model

1. モチーフ候補 m_1, m_2 の各頂点 u_i, v_j における近傍 $d\mathring{A}$ 内の頂点集合 $n(u_i), n(v_i)$ の集合 $N(m_1)$, $N(m_2)$ を抽出する.
2. モチーフ候補間 m_1, m_2 において，近傍 $n(u_i)$, $n(v_j)$ の非類似度 $nsim(n(u_i), n(v_i))$ を全近傍について求める.
3. 最小の非類似度をモチーフ候補間の非類似度 $msim(m_1, m_2)$ を式 (6) で定義する．あるモチーフ候補 a の蛋白質 B における類似モチーフ候補ペア $mcp(a, B)$ を式 (7) で定義する.

$$msim(m_1, m_2) = \min(\{nsim(n(u_i), n(v_i)) \mid n(u_i) \in N(m_1), n(v_i) \in N(m_2)\}) \quad (6)$$

$$mcp(a, B) \Leftrightarrow b \in mc(B),\ s.t.\ \forall x \in mc(B), msim(a, b) \leq msim(a, x) \quad (7)$$

3.3.2 近傍間非類似度

次に，2 つの近傍 a, b 間の非類似度 $nsim(a, b)$ を導入する．a, b それぞれの物性値 (電荷，疎水性) の平均値を $ave_c(a)$, $ave_c(b)$, $ave_h(a)$, $ave_h(b)$ とおく．また，分散を $var_c(a)$, $var_c(b)$, $var_h(a)$, $var_h(b)$ とおく．2 つの近傍間の類似性を評価するため，さまざまな方法が想定されるが，ここでは代表的な評価尺度として，物性の平均と分散を用い，以下の 3 種類の非類似度を定義する.

$$nsim1(a,b) = \sum_{i\in\{c,h\}} |ave_i(a) - ave_i(b)| \quad (8)$$

$$nsim2(a,b) = \sum_{i\in\{c,h\}} |ave_i(a) * var_i(a) - ave_i(b) * var_i(b)| \quad (9)$$

$$nsim3(a,b) = \sum_{i\in\{c,h\}} (|ave_i(a) - ave_i(b)| + |var_i(a) - var_i(b)| \quad (10)$$

3.3.3 正規化

前節で導入した非類似度の定義式では単位の異なる物性値を用いており，そのまま計算したのでは数値の大きい種類に偏る可能性がある．そのため，各物性値が 0 から 1 の範囲に収まるように正規化する．蛋白質の各頂点が持つ物性値を V_{raw} とすると正規化後の物性値 $V_{normalized}$ は以下の式で求められる．また所属する蛋白質ファミリにおける，物性値の最高値を V_{max}，最小値を V_{min} とする．これを各頂点の物性値 (電荷，疎水性) に対して行う．

$$V_{normalized} = \frac{V_{raw} - V_{min}}{V_{max} - V_{min}} \quad (11)$$

3.3.4 提案手法の検証

本節では，頂点集合を比較するための非類似度の性能の検証を行い，それを用いて提案手法がモチーフ比較において有効であるかを検証する．同機能を持つ蛋白質内において，ある 1 つの蛋白質の機能部位に対応するモチーフ候補とペアをなすモチーフ候補を，他の残りの蛋白質それぞれのモチーフ候補集合から抽出する．抽出されたモチーフ候補がそれぞれの蛋白質の機能部位に対応しているかどうかを判断し，何個の蛋白質に対して，正しく機能部位に対応するポケットが抽出できたのかを評価した．また，各蛋白質は平均 40 個のモチーフ候補を持っている．

ファミリ情報は立体構造分類データベース (SCOP[2]) を用い，ウロキナーゼに属する 10 個の蛋白質 (1owe-A，1owd-A，1gjc-B，1sqt-A，1sqo-A，1sqa-A，1owi-A，1u6q-A，1gj7-AB，1owk-A，1owj-A) に対して実験を行った．機能部位データベース (PROSITE[3]) の情報を基に分子表面における機能部位の位置を特定し，近辺の結合部位に相当するポケットを目視で識別することで，これを検証用の正解データとした．

その結果，抽出されたモチーフ候補ペアが機能部位のペアであったファミリ内での平均は $nsim1$ が 58%，$nsim2$ が 68%，$nsim3$ が 88% であった．このことから $nsim3$ が非類似度として有効であり，以降これを近傍比較の指標として用いる．

次に，全頂点を対象としたモチーフ候補の比較方法と，提案手法を用いて抽出したペアの正解率の結果を Fig. 4 に表す．横軸は蛋白質の ID(PDB-ID) であり，縦軸は正解ペアであった割合である．All はモチーフ候補の全頂点を対象とした比較，Part は提案手法を用いた比較での結果である．提案手法において，この実験では頂点から 4Å の近傍について抽出し比較した．この結果からモチーフ候補の比較において提案手法が有効であることが確認された．

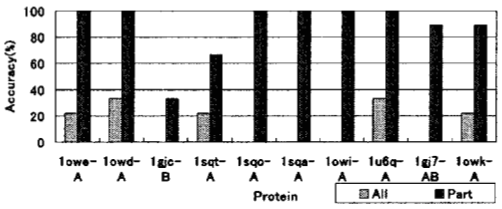

Fig. 4 Validation of method comparing motifs

3.4 近傍抽出

前節では近傍を固定し実験を行ったが，実際どれほどの範囲が機能部位に対して有効であるか自明ではない．そこで，近傍の大きさを変化させることによる，提案手法の精度の変化について検証を行った．用いる蛋白質によって有効な近傍範囲の大きさが異なる可能性があるため，3.3.4 節の実験と別の蛋白質ファミリ，トリプシンに所属する 10 個の蛋白質 (1gbt，1fn8-A，1f0t-A，1eb2-A，1fy5-A，1fn6-A，1fni-A，1bra，1co7-E，1fy8-E) を実験に用いた．Fig. 5 は各範囲での提案手法により抽出されたモチーフ候補ペアの精度を表したものである．これを見ると 1Å に近傍範囲を設定した場合が最適であるように見える．しかし，各範囲での抽出された正解ペアを確認すると，1Å で抽出されな

[2] http://scop.mrc-lmb.cam.ac.uk/scop/[6]

[3] http://www.expasy.org/prosite/[9]

かったペアが他の範囲では抽出されているということが確認された．加えて，全ての範囲で抽出されたペアを見るとウロキナーゼ，トリプシン共に正解率100%に達することも確認された．これらの結果から，1Åが必ず最適とは限らず，各モチーフペアにおいて最適な範囲が存在することがわかる．

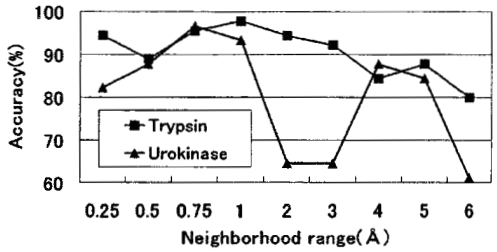

Fig. 5 Validation of neighborhood range

3.5 近傍における自動範囲設定

前節の結果からモチーフ抽出において，各モチーフ候補間ごとに最適な近傍の範囲を設定することが有効であると考えられる．そこで，モチーフ候補の比較において近傍を考慮した，類似モチーフ候補ペアの抽出方法を2種類提案する．Fig. 6に，近傍範囲を変化させることにより追加された要素を含めた，蛋白質の要素を図示する．

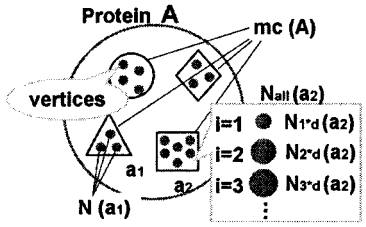

Fig. 6 Protein model2

3.5.1 類似近傍優先法 (Similar neighborhood priority method)

1つ目の提案手法 (SNP 法) は，モチーフ候補間において最も類似している近傍を優先する手法である．この方法は，モチーフにとって重要な部分というものは限定されているという考え方に基づいている．

1. モチーフ候補 m 内の全頂点の近傍の集合 $N(m)$ を一定間隔 dÅ ごとに $n * d$Å まで拡張し，範囲毎の近傍の集合 $N_{all}(m) = \bigcup_{i=1\ldots n} N_{i*d}(m)$ を抽出する．ここで $N_{i*d}(m)$ は $i * d$Å の近傍集合を表している．

2. 両モチーフ候補 $m1$，$m2$ の近傍集合を同範囲同士で比較し，最も類似していた範囲の近傍をモチーフ候補の類似部分とし，そのときの値 $msim(m1, m2)$ を式 (12) で定義する．このとき狭い範囲での類似性評価にはノイズが入る可能性があるので，広い範囲での非類似度を重視するために，非類似度に近傍の大きさ $i * d$Å を反映させる．この値が低いほど部位間で類似していると見なされる．

3. あるモチーフ候補 a の蛋白質 B における類似モチーフ候補ペアを $mcp(a, B)$ を式 (13).

$$msim(m_1, m_2) = \min(\{nsim(x, y)/(i * d) \mid x \in N_{i*d}(m_1), y \in N_{i*d}(m_2)\}) \quad (12)$$

$$mcp(a, B) = b \in mc(B)\ s.t.\ \forall x \in mc(B), msim(a, b) \leq msim(a, x) \quad (13)$$

Fig. 7 に本手法の概要を図示する．

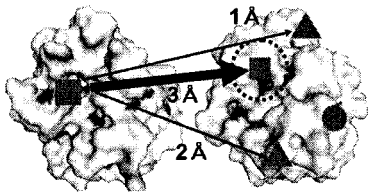

Fig. 7 Image of comparing motifs

3.5.2 類似モチーフ候補優先法 (Similar motif candidate priority method)

2つ目の手法 (SMP 法) は，モチーフ候補間で各近傍範囲において類似モチーフ候補を抽出し，最も多くの範囲において抽出されたモチーフ候補を優先する方法である．この方法は，モチーフにおける重要部位は多少範囲が変化しても，類似するという考えに基づいており，ある近傍において偶然生じるノイズを防ぐことができると考えられる．

1. モチーフ候補 m 内の全頂点の近傍の集合 $N(m)$ を一定間隔 dÅ ごとに $n * d$Å まで拡張し，範囲毎の近傍の集合 $N_{all}(m) = \bigcup_{i=1\ldots n} N_{i*d}(m)$ を抽出する．

2. ある範囲 $i*d\mathring{A}$ において両モチーフ候補 $m1$, $m2$ の全頂点における近傍非類似度の最小値 $msim(m1,m2,i)$ を式 (14) で定義する.

3. モチーフ候補 $m1$ が蛋白質 B における $m2$ を, 全範囲において類似モチーフ候補として抽出した回数 $g(m1,m2,B)$ を式 (15) で定義する. あるモチーフ候補 a の蛋白質 B における類似モチーフ候補ペアを $mcp(a,B)$ とする. また, 式 (17) で定義される関数 $f(a,b,i)$ を用いて, モチーフ候補間類似度 $msim(a,B)$ を式 (18) で定義する.

$$msim(m_1,m_2,i) = \min\left(\{nsim(x,y) \mid x \in N_{i*d}(m_1), y \in N_{i*d}(m_2)\}\right) \quad (14)$$

$$g(m_1,m_2,B) = |\{1 \le i \le n \mid \forall x \in mc(B), msim(m_1,m_2,i) \le msim(m_1,x,i)\}| \quad (15)$$

$$mcp(a,B) = b \in mc(B) \quad s.t.\ \forall x \in mc(B), g(a,x,B) \le g(a,b,B) \quad (16)$$

$$f(a,b,i) = \begin{cases} msim(a,b,i) & \forall x \in mc(B), msim(a,b,i) \le msim(a,x,i) \\ 0 & otherwise \end{cases} \quad (17)$$

$$msim(a,B) = \frac{\sum_{i=1}^{n} f(a,mcp(B),i)}{g(a,mcp(B),B)} \quad (18)$$

Fig. 8 に本手法の概要を図示する.

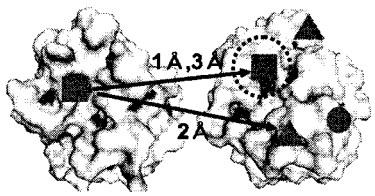

Fig. 8 Image of comparing motifs

3.5.3 近傍範囲の上限

本節では切り出す部分の妥当な上限を検証及び決定する. 上限を $2\mathring{A}$ から $6\mathring{A}$, 近傍の拡張幅を $0.25\mathring{A}$ に設定し, 3.3.1 節と同様の実験を行った. それぞれの結果を Fig. 9 に示す.

この結果から $5\mathring{A}$ が妥当な上限値であることがわかる. また, 近傍範囲を固定した場合の最高値程ではないが, 両手法とも平均的に高い正解率であり, 機能部位の範囲が特定できない蛋白質に対して有効な手法であることがわかる.

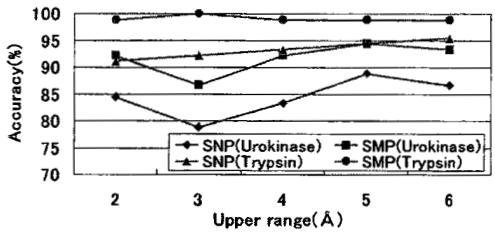

Fig. 9 Validation of step-width comparing motifs

3.6 蛋白質の分類

同じ機能を持つ蛋白質間では共通のモチーフ候補が見られる傾向にある. このことから, いくつかの蛋白質ファミリに対して提案手法を適用して得られたモチーフを利用することで, 新たに与えられた蛋白質の所属ファミリを決定することが可能であると考えられる. そこで, 3 で述べた手法で得られたモチーフ群の集合を用いて, 以下の枠組みにより, 蛋白質の分類を試みる.

ある蛋白質 T の所属ファミリを m 個の蛋白質ファミリの集合 $F = \{f_1, f_2, ..., f_m\}$ から探すことを考える. 蛋白質 T は n 個のモチーフ候補の集合を持つ. また, 各蛋白質ファミリはそれぞれモチーフ群の集合 $S(f)$ を持ち, k 個の蛋白質で構成されるファミリ $f_i = \{p_i^1, p_i^2, ..., p_i^k\}$ のモチーフ群の集合を以下のように表す.

$$S(f) = \{\{m_1, m_2, \ldots, m_k\} \mid m_j \in mc(p_i^j), j = 1...k\} \quad (19)$$

ここで, k 個のモチーフで構成されたモチーフ群 $s = \{m_1, m_2, ..., m_k\}$ と n 個のモチーフ候補を持つ蛋白質 $T = \{t_1, t_2, ..., t_n\}$ において, s 中の各要素 m_i の類似モチーフ候補ペア $mcp(m_i, T)$ を求め, 得られる集合を $Z(s,T)$ とする.

$$Z(s,T) = \{\langle a,b\rangle \mid a \in s, b = mc(a,T)\} \quad (20)$$

$Z(s,T)$ における T の要素 b に該当するモチーフの集合を $n(b, Z(s,T))$ とする.

$$n(b,Z(s,T)) = \{a \mid \langle a,b\rangle \in Z(s,T)\} \quad (21)$$

このとき $|n(b,Z(s,T))|$ が最大になるモチーフの集合を $Nmax(Z(s,T))$ とおく.

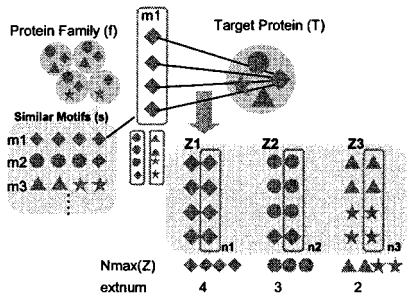

Fig. 10 Grouping Proteins

また，該当したモチーフ候補ペアの非類似度の平均 $ave(Z(s,T))$ を式 (22) により定義する．

$$ave(Z(s,T)) = \sum_{\langle a,b\rangle \in Nmax(Z(s,T))} \frac{msim(a,b)}{|Nmax(Z(s,T))|} \quad (22)$$

これらを用いて，モチーフ群非類似度 $nsimm(s,T)$ を式 (23) により定義する．また，選択された全てのモチーフ候補が異なる場合，その非類似度は採択しない．

$$nsimm(s,T) = \frac{ave(Z(s,T))}{(|n(b,Z(s,T))| * score(s))} \quad (23)$$

次に，ファミリ内全てのモチーフ群についてモチーフ群非類似度を求め，最小値をもつ類似モチーフ群をファミリ f と蛋白質 T との非類似度とする．ファミリの集合全体において，この非類似度が最小となるファミリに蛋白質 T を分類する．

4 実験結果と考察

提案手法の有効性を示すために，機能部位が既知である蛋白質データに対して蛋白質の構造情報を用いて機能部位を抽出する実験を行った．提案手法では，同一ファミリ内で共通するモチーフ候補を抽出するため，分類情報が必要である．そこで，SCOP に登録されている情報を基に実験を行った．蛋白質はセリンプロテアーゼに分類されているものを使用し，有効なサンプル数を得るために構成している蛋白質数が比較的多いファミリを採用した．実験に用いた蛋白質とそのファミリを Table 2 に示す．なお実験は，CPU2.6GHz，主記憶 2GB を搭載した Debian Linux 環境の計算機上で行った．

Table 2 Family and member proteins

Family	Protein
chymotrypsin	2gmt,1cho-E, 1gcd,1gl0-E,1acb-E
Protease B	1sgq-E,1sgp-E, 1ds2-E,1ct4-E,1ct2-E
Trypsin	1gbt,1fn8-A,1f0t-A, 1eb2-A,1fy5-A
alpha-Lytic protease	1ssx-A,1qq4-A, 1p12-E,1qrw-A,1qrx-A
Urokinase	1owe-A,1owd-A, 1gjc-B,1sqt-A,1sqo-A
Coagulation factor VIIa	1dva-H,1o5d-H, 1klj-H,1dan-H,1kli-H

4.1 モチーフ抽出

まず，提案手法の出力結果である各ファミリにおけるモチーフ集合について評価する．この実験では，拡張幅を 0.25Å に設定した．結果を Table 3 に示す．RANK は，モチーフ集合を，その score 順に並べたときの機能部位に対応するものの順位を，また，SCORE はそのモチーフ集合の score の値をそれぞれ表す．なお，括弧の数字は機能部位に対応するモチーフ集合と同じ score 値をとるモチーフ集合の個数である．

Table 3 Result of extracted motifs

	RANK (SNP)	SCORE (SNP)	RANK (SMP)	SCORE (SMP)
chymotrypsin	2(3)	17	1(2)	20
Protease B	1 (2)	20	1 (10)	20
Trypsin	1(0)	19	1(0)	19
alpha-Lytic protease	160 (284)	10	549 (378)	10
Urokinase	1(0)	16	1(2)	20
Coagulation factor VIIa	1(1)	17	3(1)	17

4.2 蛋白質分類

次に抽出されたモチーフ集合を用いて，SCOP に登録されている分類情報既知の蛋白質群に対して分類実験を行い，その分類結果をもとに提案手法の評価を行った．なお，分類器の学習には抽出精度が比較的良い範囲の SCORE が 16 以上のモチーフ集合のみを用いた．この実験では，拡張幅

を0.25Åに設定した．Table 4は分類実験に用いたテスト用の蛋白質とその所属ファミリを表している．Table 5は各ファミリごとの分類結果である．

Table 4　Family and member proteins for experiment

Family	Protein
chymotrypsin	1gct-A,1gha-E,8gch, 1gg6-ABC,1gl0-E
Protease B	3sgb-E,4sgb-E,1cso-E, 2sgf-E,2sgq-E,1ct0-E
Trypsin	1fn6-A,1fni-A,1bra, 1co7-E,1fy8-E,1an1-E
alpha-Lytic protease	1gbj-A,6lpr-A,1p11-E, ,1gbb-A,1gba-A,1p01-A
Urokinase	1sqa-A,1owi-A,1u6q-A, 1gj7-AB,1owk-A,1owj-A
Coagulation factor VIIa	1jbu-H,1qfk-H,1cvw-H

Table 5　Result of clustering proteins

	SNP(%)	SMP(%)
chymotrypsin	60	60
Proteinase B	67	100
Trypsin	0	33
alpha-Lytic protease	100	83
Urokinase	17	100
Coagulation factor VIIa	33	33

4.3　考察

Table 3，5からモチーフ抽出と蛋白質の分類どちらにおいてもSMP法が良好な結果を示している．また，Table 5を見ると機能部位が高い順位で抽出できなかったα-溶菌プロテアーゼが比較的高い精度で分類されていることが確認できる．このことは，既知の機能部位以外に，α－溶菌プロテアーゼファミリに共通する局所的部位が存在することを示唆している．

5　おわりに

本論文では，蛋白質の機能部位抽出におけるモチーフ候補比較方法について提案した．モチーフ候補を構成している全頂点の近傍において比較するといった提案手法により，機能部位の抽出精度の向上を確認した．今後，比較対象となる近傍の特徴に物性値以外を付加することにより，共通部位抽出の精度を向上させたい．

参考文献

1) Mikita Suyama, and Osamu Ohara, "DomCut: prediction of inter-domain linker regions in amino acid sequences," Oxford Journals Life Sciences Bioinformatics Vol. 19 no. 5, pp.673-674 (2003)

2) Ahmet Sacan, Ozgur Ozturk, Hakan Ferhatosmanoglu, and Yusu Wang, "LFM-Pro: a tool for detecting significant local structural sites in proteins," Bioinformatics Vol. 23 no. 6, pp.709-716 (2007)

3) Kengo Kinoshita and Haruki Nakamura, "eF-site and PDBjViewer: database and viewer for protein functional sites" Bioinformatics Vol. 20 no. 8, pp.1329-1330 (2004)

4) Lubert Stryer 著, 入村 達郎, 岡山 博人, 清水 孝雄 監訳, "ストライヤー 生化学," 東京科学同人 (1996)

5) Nripendra L. Shrestha, Yohei Kawaguchi, Tadasuke Nakagawa, and Takenao Ohkawa,"A Method of Filtering Protein Surface Motifs Based on Similarity among Local Surfaces," Proc. 5th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'04), pp.39-45 (2004)

6) Alexey G. Murzina, Steven E. Brennera, Tim Hubbarda, and Cyrus Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," J. Mol. Biol. 247, pp.536-540 (1995)

7) Jie Liang, Herbert Edelsbrunner, and Clare Woodward, "Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design.," Protein Science 7, pp.1884-1897 (1998)

8) 高木 幹夫, 下田 陽久, "画像解析ハンドブック", 東京大学出版会, (1995)

9) Christian J. A. Sigrist, Lorenzo Cerutti, Nicolas Hulo, Alexandre Gattiker, Laurent Falquet, Marco Pagni, Amos Bairoch, and Philipp Bucher, "PROSITE: a documented database using patterns and profiles as motif descriptors," Brief Bioinform. 3, pp.265-274 (2002)