蛋白質分子表面モチーフの抽出とその並列化実装

清水 穣[†] シュレスタ ヌリペンドゥラ ラル[†,☆]
大川剛直[††]

蛋白質の機能に関連する特徴的な構造パターンはモチーフとして知られている．本論文では，蛋白質の機能に密接に関連した分子表面に着目して表面モチーフを定義し，これを抽出する方式 SUMOMO を提案する．蛋白質分子表面の表現には，属性付き法線ベクトルを用いることにより，表面の物性に加え，窪み・突起といった大局的な構造の取扱いを可能とする．さらに，属性付き法線ベクトルのペアとしてモデル化した微小な表面（単位表面）を単位として，その逐次的結合処理を繰り返し実行することで，任意の形状・規模を持つ表面モチーフの抽出を実現している．実装にあたり，マスタ・ワーカモデルを用いた並列処理により，入力蛋白質の増加にともなうモチーフ抽出処理の計算量増加に対応した．マスタがワーカごとに処理する蛋白質を別々に割り当て，使用メモリの削減を図るとともに，抽出されるモチーフを管理する効率的な並列処理を実現した．5 種類の既知のモチーフを持つ 18 個の蛋白質に SUMOMO を適用した結果，抽出された表面モチーフ中に 5 つのモチーフがすべて含まれていた．1 台のマスタと 5 台のワーカからなる並列 SUMOMO を用いて 30 個の蛋白質からモチーフ抽出を行った結果，約 3.1 倍の処理速度向上となり，メモリ使用量は 5 分の 1 に削減できた．

A Method for Extracting Protein Molecular Surface Motifs and Its Implementation on Parallel Computers

Yutaka Shimizu,[†] Nripendra L. Shrestha[†,☆] and Takenao Ohkawa[††]

A motif is known as a specific pattern of the local structure related to the function of proteins. In this paper, we introduce a surface motif focusing on the molecular surface that is strongly related to the function of proteins, and propose a method of extracting surface motifs named SUMOMO. Protein molecular surfaces are expressed by using normal vectors with attributes, which enable to express physical properties and an irregular surface perspectively. Merging small unit surfaces that consist of a pair of normal vectors with attributes, can create surface motifs with variable shape and size. SUMOMO is implemented on parallel computing environment using master-worker model in order to reduce processing time. In this implementation, the master allocates proteins upon which workers focus and manages extracting motifs. As a result of applying SUMOMO to eighteen proteins that have five known motifs, all known motifs are found out in extracted surface motifs. Processing time to extract surface motifs from thirty proteins by SUMOMO with one master and five workers was shortened by 33%, and the amount of memory used was reduced to 20%.

1. はじめに

蛋白質は鎖状に連なったアミノ酸が折りたたまれた立体構造をとり，その構造が蛋白質の機能を決定付けている．蛋白質の機能に関連する部位は，保存されやすい性質があるため，多数の蛋白質から保存性の高い共通部位（モチーフ）を機能部位として抽出することができる．こうした部位の特定により，機能未知の蛋白質に対する機能の注釈付けや，既知の蛋白質の新しい機能の発見などへの応用が期待される．

有名なモチーフのデータベースに PROSITE があげられる[1]．PROSITE は機能の類似した蛋白質に共通に含まれるアミノ酸配列に注目してモチーフを集めている．また，進化的に重要なアミノ酸にランク付けを行う手法を用いて，蛋白質を構成するアミノ酸と蛋白質の機能の関係を識別する試みもある[2]．

近年，3 次元構造が解明された蛋白質の数が増える

† 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
†† 神戸大学大学院自然科学研究科
Graduate School of Science and Technology, Kobe University
☆ 現在，株式会社日立製作所
Presently with Hitachi, Ltd.

に従い，配列だけではなく立体構造上の特徴から系統的に機能をとらえようとする試みが見られる．たとえば，Rosen らは幾何学的ハッシングを用いて蛋白質の分子表面の比較を行う手法を提案している[3]．既知のモチーフに対して幾何学的ハッシングを用いることによって局所的かつ原子の出現順序に依存しない比較が可能である．また，de Rinaldis らは異なる蛋白質の分子表面モチーフ抽出のために 3D プロファイルを用いている[4]．一般にプロファイルは蛋白質のアミノ酸配列を多重アラインメントすることにより得られるが，これを蛋白質の 3 次元座標に基づいて行うことにより 3D プロファイルを得ることができる．こうして得られた 3D プロファイルはモチーフと見なすことができる．

これらの手法が蛋白質分子表面に着目する理由として，蛋白質がさまざまな機能を発現するにあたり，その機能に直接的に関与する部位が蛋白質の分子表面に存在していることがあげられる．そして，蛋白質の機能を知るうえで機能部位がどのような物理的・形状的な特性を備えているかが重要となる．そこで，蛋白質の機能に強い関連のある蛋白質分子表面データを対象とし，その形状や物性に注目して表面モチーフを自動抽出する方式 SUMOMO（SUrface MOtif mining MOdule）を提案する[5]．

表面モチーフの抽出を行うために，属性付き法線ベクトル[6] を用いて蛋白質分子表面をモデル化する．属性付き法線ベクトルとは，蛋白質表面で高曲率な突起・窪みにおける法線ベクトルに，分子表面の曲率や物性などの属性を付加したものである．これにより，蛋白質表面形状を蛋白質の機能発現に大きくかかわる突起・窪みレベルで大局的に表現・比較することが可能となる．表面モチーフの大きさを一意に決定することは難しいので，分子表面を微小な部分領域に分割した単位表面を定義し，隣接する単位表面の結合を繰り返すことで，さまざまな形状・規模の表面モチーフの抽出を可能とする．

一方，この手法では，ある蛋白質内に類似したモチーフ候補が複数存在すると，それと類似した他の蛋白質のモチーフ候補との位置関係を対応付ける際にその組合せの数が増え，モチーフ抽出処理に必要な計算量が増大する．そこで実装に際して，マスタ・ワーカモデルによる並列化により処理の高速化とともに処理に必要な蛋白質表面データを各計算機に分散させることによって必要メモリを低減する．

なお，本論文は，文献 5) で提案した表面モチーフとその抽出手法の概念をもとに，表面モチーフ抽出問題を厳密に数理モデル化して記述するとともに，その並列化実装法を与える形で拡張したものである．

2. 蛋白質分子表面モチーフ抽出方式 SUMOMO

2.1 蛋白質表面モチーフ

蛋白質の構造の中で機能的に重要な部分は進化の過程で保存される傾向があり，このような領域中でよく見られるアミノ酸配列のパターンを配列モチーフという．しかし，アミノ酸配列の類似性からは，蛋白質の機能と強く関係する分子表面形状の類似性を必ずしも発見できるとは限らない[7)~9)]．つまり，蛋白質が機能を発現するうえで重要となるのは配列よりも，むしろその表面形状と物性である[10)~12)]．そこで，分子表面の観点から蛋白質どうしの機能における類似性を発見するために表面モチーフという概念を導入する[5]．表面モチーフとは，直感的には，比較的多数の蛋白質に共通して見られる表面形状や物性の局所的パターンと見なすことができる．厳密な定義については以降で述べる．

蛋白質立体構造情報データをもとに分子表面の形状と物性を計算することで構築された蛋白質機能部位表面データベース eF-site が公開されている[13),☆]．eF-site では，1 つの蛋白質ごとに XML 形式で分子表面の座標が電位，疎水性，曲率などの属性値とともに記述されている．SUMOMO は，eF-site に登録されている複数の分子表面データを入力とし，表面モチーフを出力する．

分子表面の突起や窪みといった形状の比較を大局的に行うために，分子表面を属性付き法線ベクトルでモデル化する．属性付き法線ベクトルとは，分子表面上のある頂点の近傍内で最も曲率の高い点を通る法線ベクトルに電位，疎水性[14] などの属性値を付加したものである[6]．**図 1** に属性付き法線ベクトルの概略図を

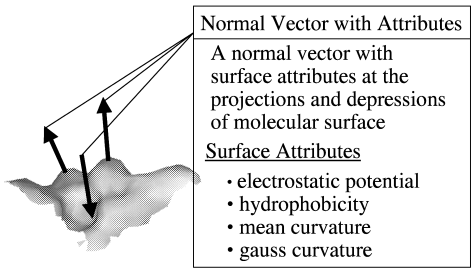

図 1 属性付き法線ベクトル
Fig. 1 Normal vector with attributes.

☆ http://ef-site.protein.osaka-u.ac.jp/eF-site/

示す．1つの蛋白質は，数十〜数千，平均的には約500の属性付き法線ベクトルから構成される．

抽出すべき表面モチーフの定義を以下に示す．まず，2つの属性付き法線ベクトルの間の類似性を，次のように定義する．

定義 2.1（属性付き法線ベクトル間の類似性）

2つの属性付き法線ベクトル v_1，v_2 に対し，ベクトルの4つの属性（電位，疎水性，平均曲率，ガウス曲率）それぞれの値が一定値以下のとき，v_1 と v_2 は類似性を持つといい，$v_1 \simeq v_2$ と表す．

□

このような v を用いて，ある蛋白質の分子表面は，次のように表現できる．

$A = \{v_{A_1}, v_{A_2}, \ldots, v_{A_N}\}$

このとき A の要素数 N は蛋白質の分子表面の規模によるため，蛋白質ごとに定まる．ここで，ある蛋白質分子表面 A から得られる部分表面 A' を次のように定義する．

定義 2.2（部分表面）

分子表面 A の部分集合を A' とする．ここで，A' 内のすべての属性付き法線ベクトルをノードとし，A' 内の任意の異なる2つの属性付き法線ベクトル v_1，v_2 が $d(v_1, v_2) < T_R$ を満たすときに両者をアークで結んだグラフを考える．なお，$d(v, v')$ はユークリッド座標上でのベクトルの始点間の距離を，T_R は事前に定めた定数を表す．このとき，グラフが連結グラフとなるような A' を A の部分表面という．また，A' が A の部分表面であるとき，A によって表現される分子表面を持つ蛋白質を部分表面 A' の抽出元蛋白質という．

□

さらに，ある2つの部分表面間の類似性について，以下のように定義する．

定義 2.3（部分表面の類似性）

2つの異なる蛋白質 A，B に対する部分表面を $A'(\subset A)$，$B'(\subset B)$ とする．A'，B' が以下の2つの条件を満たすとき，両者は類似しているといい，$A' \simeq B'$ と表す．

条件 1 A' 内のすべてのベクトル v_A に対して，$v_A \simeq f(v_A)$ が成立する $A' \mapsto B'$ の全単射写像 f が存在する．

条件 2 A' と B' をそれぞれ定義2.2の連結グラフで表したとき，両者は同型グラフとなり，対応するノードは写像 f によって求められる．

□

以上のように定義された類似する部分表面の集合から，表面モチーフ候補および表面モチーフが定義される．

定義 2.4（表面モチーフ候補）

事前に定めた定数 T_B に対し，部分表面 A_i を T_B 個以上集めて構成される集合を $\alpha = \{A_1, A_2, \ldots, A_M\}$ とする．α 内の任意の A_i，$A_j (i \neq j)$ について $A_i \simeq A_j$，かつ，(A_i の抽出元蛋白質) $\neq$ (A_j の抽出元蛋白質) が成り立つとき，α を表面モチーフ候補という．

□

定義 2.5（表面モチーフ）

ある表面モチーフ候補 α について，$\alpha \subset \beta$ を満たす表面モチーフ候補 β が存在せず，かつ，α のすべての要素 A_i について $A_i \subset B_j$ を満たす要素 B_j を持つ β が存在しないとき，α を表面モチーフという．

□

2.2 表面モチーフ抽出処理

2.1節で定義された表面モチーフを抽出するために，“蛋白質表面を単位表面に分割”，“表面モチーフ候補の抽出”，“表面モチーフ候補の結合” の各手続きからなる手法を導入する．以下にその詳細を述べる．

2.2.1 分子表面の分割

表面モチーフはさまざまな大きさや形が考えられるため，あらかじめその規模を見積もることができない．そこで，分子表面を小さな領域に分割し，類似する部分表面を抽出する．それらをある一定の条件のもとで，かつ規模が最大になるように結合することで，任意の大きさ・形状を持つ表面モチーフを構成する．

蛋白質の分子表面を微小領域に分割するには，属性付き法線ベクトルを単位とした表面分割が有効である．しかし，属性付き法線ベクトル1つを単位とする分割方法では，後に規模を拡大するために必要な分子表面上の位置情報を保持できない．そのために，距離 T_R 以内にある2つの属性付き法線ベクトルの組を分子表面の最小単位（単位表面）として定義する．**図2** に示すように，単位表面には属性付き法線ベクトルが持つ

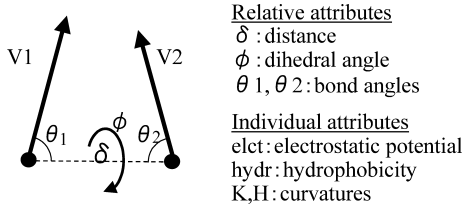

図 2 単位表面
Fig. 2 Unit surface.

物性値以外に，ベクトル間の距離や角度をその要素として付与することで，表面形状の相対位置関係も考慮に入れた比較を可能とする．

また，属性付き法線ベクトルどうしの相対位置関係の比較のために，分子表面上のベクトルすべての組を考慮すると，その数は膨大になる．しかし，あらかじめ表面モチーフの定義で示した距離 T_R 以内のベクトルの組に限定することで定義 2.2 に合わない部分表面を削除し，扱う単位表面の数を削減できる．たとえば，500 個のベクトルで表現された蛋白質の場合，すべての属性付き法線ベクトルの組を考慮した組合せは 250,000 であるが，単位表面として抽出される属性付き法線ベクトルの組は 2,000〜3,000 程度に減少する．なお，単位表面の総数は，入力蛋白質数 n に対して $O(n)$ となる．

2.2.2 表面モチーフ候補の抽出

定義 2.4 のような表面モチーフ候補を抽出するために，まず，単位表面を類似性によって分類し，その数をカウントする必要がある．単位表面どうしの類似性は，定義 2.3 による．しかしすべての単位表面について総当たりで類似性を比較すると，入力蛋白質数 n に対して $O(n^2)$ の時間を要する．そこで単位表面が持つ相対位置関係や物性値を要素とする多次元のバケットを利用することで単位表面を分類する[15]．これにより，類似性の比較は $O(n)$ となる．バケット分割の際はバケットの境界付近の類似単位表面が別々のバケットに分類されるのを防ぐため，バケットの境界に重複領域を定める．重複領域に分類された単位表面は境界をはさんだ両方のバケットに格納される．その結果，各バケットに格納された単位表面の集合は定義 2.3 を満たす類似部分表面の集合となる．

各バケットごとの単位表面集合から，定義 2.4 を満たす単位表面の部分集合を抽出する．このとき同一バケット内には，同じ抽出元蛋白質の単位表面が複数格納されている場合がある．これは，定義 2.4 において，他の蛋白質の単位表面と表面モチーフ候補 α を構成する単位表面が複数候補存在するということである．1 つの抽出元蛋白質について複数の単位表面がバケット内に存在する場合は，各単位表面すべての組合せについて表面モチーフ候補の抽出を行う．このため，入力蛋白質数を n，同一バケット内に含まれる 1 つの蛋白質あたりの単位表面数を k とすると，抽出される表面モチーフ候補数，ならびに抽出のための計算量は $O((1+k)^n)$ となる．このことから，現実的な時間で処理するためには，1 つの蛋白質から得られる単位表面数 2,000〜3,000 に対し，バケット数が比較的大きな値でなければならないことが示唆される．

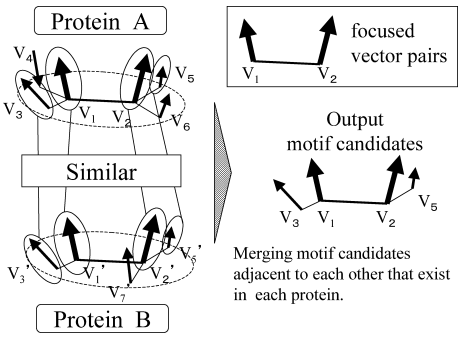

図 3 1 回目の表面モチーフ候補結合処理
Fig. 3 Motif candidates extraction.

2.2.3 表面モチーフ候補の結合

抽出した単位表面の集合の中から，定義 2.4 を満足したまま可能な限り規模を拡大させて，表面モチーフとして出力する．ある表面モチーフ候補 α のすべての要素 A_i において，$A_i \cap B_i \neq \emptyset$ を満たす B_i を要素として持つ β を α と結合し，新たな表面モチーフ候補 γ を作成する．具体的な処理は以下のとおりである．

ある単位表面をもとにモチーフが抽出される様子を**図 3** に示す．まず，ある 1 つの表面モチーフ候補に注目し，1 回目の表面モチーフ候補結合を行う．図 3 の例では，注目した蛋白質 A の単位表面 V_1-V_2 に対して近傍の単位表面 V_1-V_3，V_1-V_4，V_2-V_5，V_2-V_6 が抽出される．これらの単位表面と，V_1-V_2 と類似する単位表面 V_1'-V_2' の近傍の単位表面が比較され，対応関係のある単位表面 V_1-V_2，V_1-V_3，V_2-V_5 が結合される．また，V_1-V_2 の近傍のある単位表面に対して，複数の類似単位表面が存在する場合がある．このときは，すべての結合の組合せを表面モチーフ候補として出力することで，定義 2.3 を満たす部分表面を網羅する．こうして 1 回目の結合処理が完了する．

ある単位表面どうしが結合可能かどうかを判定するには，抽出元蛋白質の数だけベクトルの共有を調べる必要がある．抽出元蛋白質数は入力蛋白質数に比例するため，オーダは $O(n)$ となる．

さらに表面モチーフ候補に対して，結合処理を繰り返すことにより，表面モチーフを抽出する．**図 4** に表面モチーフ候補の結合の例を示す．表面モチーフ候補（ID: 1）と同（ID: 2）は，ともに抽出元蛋白質が A，B，C である．抽出元の蛋白質の組合せが同じであるので，次に共通して持つベクトルが存在するかを調べ，ベクトル V_3 を共有するので，結合して 1 つの表面モチーフ候補にする．表面モチーフ候補がこれ以

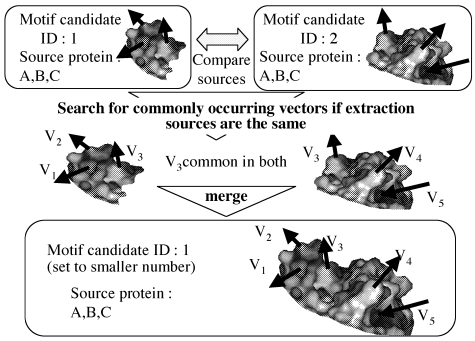

図 4 表面モチーフ候補の結合方法
Fig. 4 Motif candidates merging.

上結合できなくなるまで結合を繰り返す．結合した表面モチーフ候補を表面モチーフとして出力する．

分子表面の分割処理において，表面モチーフの定義 2.2 を満たす単位表面を抽出し，表面モチーフ候補の抽出によって定義 2.3，2.4 を満たす単位表面を抽出している．最後に，表面モチーフ候補の結合によって定義 2.5 を満たす，規模が最大の表面モチーフ候補を表面モチーフとして出力している．処理の各段階において表面モチーフの定義を満足する単位表面のみを網羅的に抽出しているため，定義した表面モチーフを過不足なく抽出することが可能である．ただし，単位表面を部分表面の最小単位として扱っているため，属性付き法線ベクトルを 1 つだけ要素として持つ表面モチーフは出力されない．

前述のとおり，表面モチーフ候補数は $O((1+k)^n))$ であるため，表面モチーフ候補結合処理全体の計算量は $O(n \times (1+k)^n)$ となる．

3. SUMOMO の並列化実装

3.1 概　　要

以上の手法をもとにプロトタイプシステムを構築し，モチーフ抽出実験を行った．30 個の入力蛋白質に対して約 1 時間の実行時間でモチーフ抽出処理が完了した．その際の処理時間の内訳を処理段階ごとに調べた結果を**表 1** に示す．表 1 は入力蛋白質数を 2 個から 30 個まで変えて 1 回ずつモチーフ抽出を行い，その処理にかかった時間割合の平均をとったものである．処理時間のほとんどをモチーフ候補の結合処理に費やしており，その高速化が望まれる．そこで，モチーフ候補結合処理を中心に分散化を図り，マスタ・ワーカモデルを用いて SUMOMO を並列化実装する．

図 5 に並列化方式の概要を示す．“分子表面の分割”と “モチーフ候補抽出” については，全体の処理時間に対する割合が少ないことから，マスタとワーカそれぞれが独自に処理を実行する．“モチーフ候補の結合” 処理に関しては，マスタがワーカに対し結合開始点，すなわち結合処理を行う際に注目するモチーフ候補を割り当てる．また，モチーフ候補にはそれぞれ一意の ID が付与され，マスタが割り当てる結合開始点の ID を指定することによって結合処理を行う．マスタによって割り当てられた結合開始点に対して結合作業が終了したワーカは，結合の終了したモチーフ候補の ID をマスタに返し，再びマスタによって結合開始点を割り当てられる．このようにして，すべての結合開始点についてモチーフ候補結合作業が完了するまで処理を繰り返す．

結合処理の際，必要なすべてのデータを各ワーカで共有すると，**図 6** に示すように，入力蛋白質を増加させた場合に必要になるメモリ量が飛躍的に増大する．ワーカの持つメモリ容量には限界があり，入力蛋白質の増加にともない処理時間とともに必要メモリの増加に対しても対応する必要がある．そこで，モチーフ候補の抽出元蛋白質に着目し，結合処理に必要なデータ

表 1 処理別必要時間
Table 1 Processing time of SUMOMO.

Processing phase	processing time (%)
Dividing into unit surfaces	0.2
Extracting motif candidates	3.5
Merging motif candidates	96.3

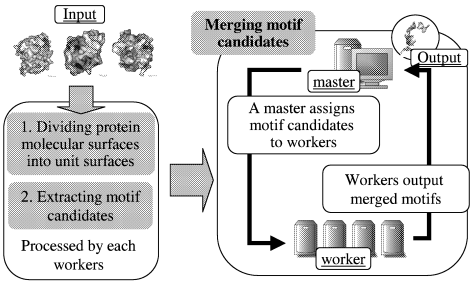

図 5 蛋白質表面モチーフ抽出並列化の概要
Fig. 5 Outline of parallel processing of motif extraction.

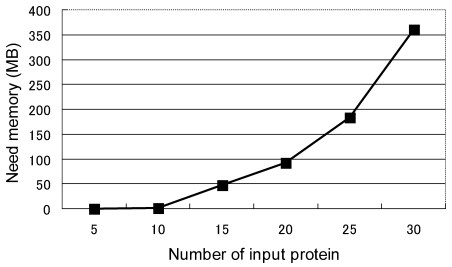

図 6 表面モチーフ抽出処理における必要メモリ
Fig. 6 Memory needed for surface motif extraction.

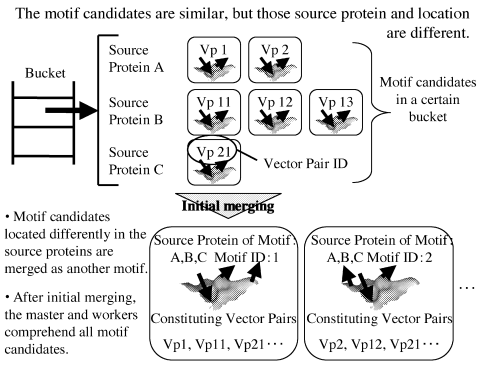

図 7 並列化における 1 回目の結合処理
Fig. 7 First merging procedure on parallel processing.

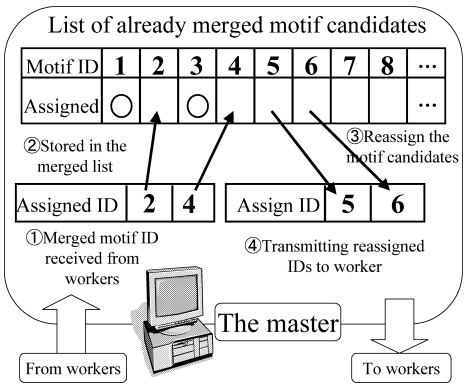

図 8 結合済みモチーフ管理
Fig. 8 Management of merged motif candidates.

のみを各ワーカに分割して持たせることで，使用メモリを削減する．

3.2 処理蛋白質を限定した割当て手法

3.2.1 ワーカの注目する蛋白質の決定と結合処理の開始

3.1 節で述べたように，単位表面分類処理の終了後，次の処理であるモチーフ候補結合処理に移行する．はじめにマスタは各ワーカが注目すべき蛋白質を指定することで，1 回目のモチーフ候補結合を行う．このとき，各ワーカに対して割り当てられる蛋白質について，結合処理の複雑さは均一ではない．たとえば，蛋白質の表面積が大きければ，ベクトルペアの数が増加し，モチーフ結合処理の負荷も増大する．このことから，蛋白質から抽出されたベクトルペアの総数がワーカ間でなるべく平均化するように蛋白質の割当てを決定する．

マスタから割り当てられた蛋白質に注目して，**図 7** に示すように，1 回目の結合処理によって抽出元蛋白質の組合せを考慮したモチーフ候補がすべて作られる．モチーフ候補に一意な ID が与えられ，以降の処理でマスタはモチーフ候補 ID によって結合開始点を指定する．また，ワーカは割り当てられた蛋白質が抽出元となっているモチーフ候補のデータをすべて保持する．ここまでの処理を行った後，マスタがすべてのモチーフ候補の情報を把握するために，各ワーカからモチーフ候補のデータをすべて受け取る．

3.2.2 マスタによる結合開始点の割当て決定

図 7 に示した例において，結合されたモチーフ候補の抽出元は，蛋白質 A，B，C である．このとき，蛋白質 A に注目したワーカ，蛋白質 B に注目したワーカ，蛋白質 C に注目したワーカがそれぞれ同一のモチーフ候補のデータを持つ．このため，これ以降の結合処理の際に同一のモチーフ候補に対して冗長に結合処理を行う可能性がある．この “モチーフ結合処理の重複” の問題を解決するために，**図 8** のようにマスタはすでに結合処理の終了したモチーフ候補リストを保持する．割り当てられた結合開始点について結合作業の終了したワーカはマスタに対して結合済みモチーフ候補のリストを送信するとともに，マスタから次の結合開始点を割り当てられる．ワーカから送信されてくる結合済みモチーフのリストは，結合開始点だけでなく結合されたすべてのモチーフ候補を含む．

マスタが結合開始点を割り当てる際に，結合済みリストを用いて結合済みのモチーフ候補を再び割り当てることのないようにすることで，出力モチーフの重複を防ぐ．

3.2.3 モチーフ抽出処理の終了したワーカへの蛋白質の割当て

モチーフ結合処理ではマスタによって指定された蛋白質のみについてモチーフ候補の結合を行うことで，処理の並列化と必要メモリの軽減を図る．しかし，各ワーカに割り当てられた蛋白質のモチーフ候補結合処理は蛋白質の種類やワーカの処理性能差によって処理が終了するまでに時間差が生じる．そこで，割り当てられた蛋白質について結合処理の終了したワーカについては他のワーカに割り当てられている蛋白質を重複して割り当てることによって結合処理途中の蛋白質について処理の効率化を図る．割り当てる蛋白質を決定するために，優先度を導入する．すなわち，結合処理の進行状況が遅い蛋白質は割当ての優先度が高いと考え，蛋白質 i の優先度 P_i を次式で定義する．

$$P_i = 1 - \frac{E_i}{A_i}$$

ただし A_i は蛋白質 i を抽出元とするモチーフ候補数，E_i は蛋白質 i を抽出元とするモチーフ候補のうち結合済みのモチーフ候補数である．

表 2 バケット分割数
Table 2 Number of buckets.

	Distance	Dihedral angle	Bond angles
Number	100	16	16
	Electrostatic potential	Hydrophobicity	Curvatures
Number	4	2	20

表 3 検証実験に用いた蛋白質
Table 3 Protein for an experiment of motif extraction.

Function/Classification	PDB ID
Serine protease	1avt, 1btw, 1cho, 4sga
Glycosyl hydrolases	1bga, 1bgg
Scorpion short toxins signature	1c55, 1sxm, 1tsk, 1jgk
Snake toxins signature	1cvo, 1fas, 2nbt, 1cod
Oxidoreductase	1dhr, 1e3s, 1bdm, 1gdh

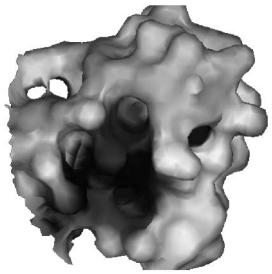

図 9 抽出された表面モチーフの例
Fig. 9 Example of extracted motifs.

4. 評価実験と考察

4.1 表面モチーフ抽出実験

並列化実装した SUMOMO を用いて，表面モチーフ抽出の検証実験を行った．実験では，$T_R = 10$Å，$T_B = 3$ とし，表面モチーフ抽出処理を行った．単位表面の持つ属性値ごとのバケット分割数は**表 2** のように設定した．

SUMOMO が出力する表面モチーフの正当性を評価するため，ここでは機能部位がすでに判明している 18 個の蛋白質の表面データを入力し，出力される表面モチーフについて検証した．入力した 18 個の蛋白質とこれらが有する機能分類を**表 3** に示す．表 3 に示すように，18 個の蛋白質は 5 つの機能に分けられ，対応する機能部位が知られている．SUMOMO によってこれらの機能部位に対応するすべての表面モチーフの抽出に成功しており，その有効性を確認した．抽出された表面モチーフの例を**図 9** に示す．

4.2 並列化の評価

SUMOMO の並列化実装に関する評価実験を行った．実験に用いた計算機は以下のとおりである．また，

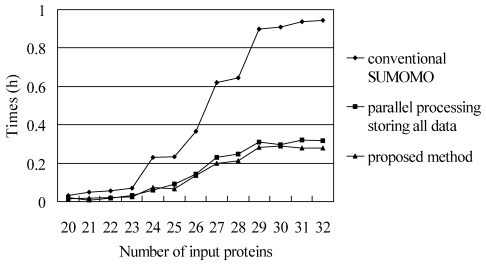

図 10 各手法の違いによるモチーフ抽出処理時間
Fig. 10 Total processing time of extracting motifs.

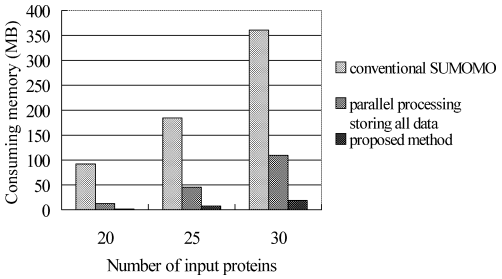

図 11 各手法における必要メモリ
Fig. 11 Amount of memory usage.

実装には Grid ミドルウェアである Ninf Ver.1☆を用いた[16),17)]．

- マスタ計算機
  CPU Pentium4 2.4 GHz，主記憶 1 GB
- ワーカ計算機
  CPU Pentium4 2.2 GHz，主記憶 512 MB

まず，計算機 1 台（Pentium4，2.2 GHz，主記憶 512 MB）による非並列型の SUMOMO と全データ保持型並列化手法（データ共有 SUMOMO）および 3 章の方式で並列実装した SUMOMO（並列 SUMOMO）について蛋白質の入力数を変えてモチーフ抽出にかかる時間を計測した．ここでデータ共有 SUMOMO とは，SUMOMO の並列化において，すべてのワーカがモチーフ候補のデータを冗長に保持する手法を指す．並列化には，マスタ 1 台とワーカ 5 台を用いた．結果を**図 10** に示す．

次に，入力蛋白質数を変えて並列 SUMOMO とデータ共有 SUMOMO におけるワーカの使用メモリ量を**図 11** に示す．どちらの並列化手法においても，マスタ 1 台とワーカ 5 台について実験を行った．また，参考として計算機 1 台による非並列化手法についても使用メモリを計測した．ワーカの使用メモリとは，全ワーカの使用メモリ量の平均である．

最後に，並列 SUMOMO の並列台数を 2 台，5 台，

☆ http://ninf.apgrid.org/

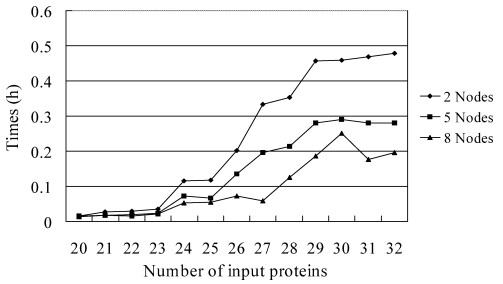

図 12 並列化台数による処理速度の変化
Fig. 12 Processing time by number of nodes.

8 台と変えたときの処理時間の変化を**図 12** に示す.

4.3 考　　察

まず，表面モチーフ抽出実験については，抽出すべきすべての既知モチーフを抽出することができ，SUMOMO の有効性が確認できた．しかし，5 つの既知モチーフを抽出するために出力されたモチーフ数は 1,233 個であり，false positive は実に 1,228 個（99.5%）にのぼる☆．膨大な false positive の削減方式については，局所構造の類似性に基づいた抽出元蛋白質のクラスタリングにより，出力された表面モチーフの中から，有意なモチーフのみに絞り込むフィルタリング方式について検討を進めており，すべての true positive を維持したまま，出力モチーフ総数の 86%に相当する false positive を削減することが可能となっている[18]．しかし，数個のオーダである ture positive を発見するために，抽出されているモチーフはフィルタリング後も百数十のオーダであり，さらなる false positive の削減方法を探ることを今後の課題としたい．

図 10 から，並列 SUMOMO によって非並列化時よりも高速に表面モチーフの抽出が行えることが確認された．また，データ共有 SUMOMO と比較すると，ほぼすべての入力蛋白質数についてわずかではあるが並列 SUMOMO が高速にモチーフ抽出を行っている．これはデータの分散によってマスタとワーカの間の通信コストが削減されたためと考えられる．しかし，図 11 に見られるようにワーカ 1 台あたりの持つモチーフ候補データはほぼ台数分減少しているが，これによって予想される通信量の削減効果に相当する高速化は実現されていない．つまり，マスタとワーカの通信以外にも処理時間のボトルネックとなっている部分が存在すると考えられる．

☆ false positive としてカウントした出力モチーフの中には，未知のモチーフが含まれている可能性もあるが，厳密な判定が難しいため，評価実験では既知モチーフに対応していないものはすべて false positive として扱っている．

図 11 から，データ共有 SUMOMO に比べ並列 SUMOMO のほうがモチーフ抽出における必要メモリ量が減少し，より多くの入力データについて表面モチーフの抽出が行えることが分かる．また，入力蛋白質が増えるに従ってワーカの数に対する入力蛋白質数の割合が増大すると，モチーフ結合処理においても出力モチーフの重複が生じにくくなり，より並列性が高まると思われる．

また図 12 に示すように，並列化台数を増やすことによりさらに高速化が可能である．20〜32 個の入力蛋白質について並列化に対する速度向上比と実際の並列化台数の比の平均をとると，2 台では 0.96，5 台では 0.61，8 台では 0.57 であった．5 台，8 台での並列化効率にあまり差が見られないが，処理時間は各ワーカへの注目蛋白質の割当てなどによって変動するため，実際には並列化台数を増やすにつれ効率は低下していると考えられる．並列化における速度向上比の低下原因として，大量の分子表面データを入力，出力する際のファイル I/O や，全ワーカで重複して行われるモチーフ候補抽出処理などがあげられる．並列化以外でも，分子表面データの格納方法や表面モチーフ抽出処理方法の改善により速度向上，必要メモリ量の低減が期待できる．特に，現在の SUMOMO では多数のモチーフ候補が抽出されるが，多数の蛋白質に共通して存在する単位表面は特異性が低いと考え，削除することでモチーフ候補数を削減し，処理を高速化することが考えられる．今後，アルゴリズム自体の改良を含め，モチーフ抽出処理のいっそうの性能向上について検討する予定である．

図 10，図 11 において入力蛋白質数が 23 以下においては，並列化の効果が現れていない．モチーフ抽出のためには抽出元蛋白質の組合せを考慮する必要があるが，蛋白質数が 23 以下においてはこの組合せに対する処理の計算量よりも，むしろデータ通信などにかかる処理時間が支配的であったためと考えられる．

5. 結　　論

蛋白質分子表面を属性付き法線ベクトルによってモデル化し，単位表面への分割，モチーフ候補の抽出，結合によって表面モチーフを抽出する SUMOMO を提案した．さらに，SUMOMO を並列処理することにより高速に蛋白質分子表面モチーフを抽出する方式を開発し，Ninf を用いた実装を行った．モチーフ結合処理の並列化にあたり，マスタによって各ワーカに対してモチーフ結合処理を行う蛋白質の限定を行うことにより，各ワーカの必要メモリ量の削減が実現できた．

謝辞 日頃よりご指導いただく，薦田憲久教授に感謝します．eF-site の開発者であり，また貴重な助言をいただいた大阪大学蛋白質研究所の中村春木教授に感謝します．

本研究の一部は文部科学省科学研究費補助金および科学技術振興機構バイオインフォマティクス推進事業（JST-BIRD）の助成による．

参 考 文 献

1) Bairoch, A. and Bucher, P.: Prosite: recent developments, *Nucl Acids Res*, Vol.19, pp.3583–3589 (1994).
2) Yao, H., Kristensen, D.M., Mihalek, I., Sowa, M.E., Shaw, C., Kimmel, M., Kavraki, L. and Lichtarge, O.: An accurate, sensitive and scalable method to identify functional sites in protein structures, *J. Mol. Biol.*, Vol.326, pp.255–261 (2003).
3) Rosen, M., Lin, S.L., Wolfson and Nussinov, R.: Molecular shape comparisons in searches for active sites and functional similarity, *Protein Engineering*, Vol.11, pp.263–277 (1998).
4) de Rinalds, M., Ausiello, G., Cesareni, G. and Helmar-Citterich, M.: Three-dimensional profiles: A new tool to identify protein surface similarities, *J. Mol. Biol.*, Vol.284, pp.1211–1221 (1998).
5) Shrestha, N.L., Kawaguchi, Y. and Ohkawa, T.: A method for extraction of surface motifs from protein molecular surface database using normal vectors with attributes, *The 7th Joint Conference on Information Sciences*, pp.911–914 (2003).
6) Kaneta, Y., Shoji, N., Ohkawa, T. and Nakamura, H.: A method of comparing protein molecular surface based on normal vectors with attributes and its application to function identification, *Information Sciences*, Vol.146, pp.41–54 (2002).
7) Branden, C. and Tooze, J.: *Introduction to Protein Structure*, Garland Publishing (1999).
8) Kinoshita, K., Sadanami, K., Kidera, A. and Go, N.: Structural motif of phosphate-binding site common to various protein superfamilies, *Protein Engineering*, Vol.12, pp.11–14 (1999).
9) Jackson, R.M. and Russell, R.B.: The serine protease inhibitor canonical loop conformation: Examples found in extracellular hydrolases, toxins, cytokines and viral proteins, *J. Mol. Biol.*, Vol.296, pp.325–334 (2000).
10) Goto, S., Nishioka, T. and Kanehisa, M.: Ligand: Chemical database for enzyme reactions, *Bioinformatics*, Vol.14, pp.591–599 (1998).
11) Gabb, H.A., Jackson, R.M. and Sternberg, M.J.E.: Modelling protein docking using shape complementarity, electrostatics and biochemical information, *J. Mol. Biol.*, Vol.272, pp.106–120 (1997).
12) Ester, M., Kriegel, H. and Wirth, S.: Feature based classification of protein docking sites: An algorithm for large databases and experimental results, *German conf. on Bioinformatics*, pp.193–196 (1996).
13) 木下賢吾，中村春木：タンパク質分子表面形状と物性のデータベース ef-site による分子機能類似性検索，生物物理，Vol.42, pp.20–23 (2002).
14) Kyte, J. and Doolittle, R.: A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.*, Vol.157, pp.105–132 (1982).
15) Wang, J.T.L., Shapiro, B. and Shasha, D.: *Pattern Discovery in Biomolecular Data: Tools, Techniques, and Applications*, Oxford University Press (1991).
16) Seymour, K., Nakada, H., Matsuoka, S., Dongarra, J., Lee, C. and Casanova, H.: Overview of gridrpc: A remote procedure call api for grid computing, LNCS, Vol.2536, pp.274–278 (2002).
17) Sato, M., Nakada, H., Sekiguchi, S., Matsuoka, S., Nagashima, U. and Takagi, H.: Ninf: A network based information library for a global world-wide computing infrastracture, *HPCN'97*, pp.491–502 (1997).
18) Shrestha, N.L., Kawaguchi, Y., Nakagawa, T. and Ohkawa, T.: A method of filtering protein surface motifs based on similarity among local surfaces, *Intelligent Data Engineering and Automated Learning*, LNCS, Vol.3177, pp.39–45 (2004).

(平成 16 年 12 月 22 日受付)
(平成 17 年 5 月 20 日再受付)
(平成 17 年 7 月 15 日採録)

清水 穣

昭和 57 年生．平成 16 年大阪大学工学部電子情報エネルギー工学科卒業．同年大阪大学大学院情報科学研究科マルチメディア工学専攻博士前期課程入学．蛋白質 DB からの情報抽出に関する研究に従事．

シュレスタ ヌリペンドゥラ ラル

1978 年生．2003 年大阪大学大学院情報科学研究科マルチメディア工学専攻博士前期課程入学．バイオインフォマティクス関連の研究に従事．2005 年同課程修了．同年（株）日立製作所に入社．情報通信グループに所属．

大川 剛直（正会員）

昭和 38 年生．昭和 63 年大阪大学大学院工学研究科通信工学専攻博士前期課程修了．大阪大学助手，講師，助教授を経て，平成 17 年神戸大学大学院自然科学研究科教授．工学博士．知的ソフトウェア，バイオインフォマティクスに関する研究に従事．IEEE 等の会員．