2006－BIO－6 （6）
2006／9／15

配列プロファイルを利用したドメインリンカー予測

滝沢雅俊[†‡]　山名早人[†]　野口保[‡]

† 早稲田大学大学院理工学研究科　〒169-8555 東京都新宿区大久保 3-4-1

‡ 産業技術総合研究所 生命情報科学研究センター　〒135-0064 東京都江東区青海 2-42

あらまし　ドメインリンカー予測は、タンパク質の立体構造を迅速に決定するうえで重要な役割を果たす。従来のドメインリンカー予測法では、スライディングウィンドウにより予測を行うため、ドメインリンカーに出現するアミノ酸の位置依存性を十分に考慮することができない。本稿では、ドメインリンカーの両端部分に着目し、配列プロファイルから抽出した各アミノ酸の出現位置依存性をもとに、サポートベクターマシーンを用いて予測する方法を提案する。提案手法を DSSP で決定したコイル領域に対して適用させた結果、従来手法に比べ、Sensitivity、Specificity 共に約 20%向上することが可能であると確認した。

Domain linker prediction based on position-specific scoring matrix

Masatoshi Takizawa[†‡]　Hayato Yamana[†]　Tamotsu Noguchi[‡]

† Graduate School of Science and Engineering, Waseda University

3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555 Japan

‡ Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Japan

2-42 Aomi, Koto-ku, Tokyo 135-0064 Japan

Abstract The domain linker prediction plays an important role in efficient protein structure analysis. Since previous domain linker prediction methods have employed sliding window, these methods do not explicitly consider the position dependence of amino acids within domain linkers. In this paper, we propose a novel domain linker prediction method, focusing on both ends of the domain linker. Our method employs Support Vector Machines, which train on position dependence of amino acids extracted from the position-specific scoring matrix. As a result of the experiment using data set of coil regions determined by DSSP, we confirmed that our method could achieve about 20 percent increase in sensitivity and specificity over the previous ones.

１．はじめに

分子量の大きなタンパク質の立体構造を解析する場合、構造の基本単位であるドメインごとに分割し、解析を行う必要がある。従って、アミノ酸配列からドメインに相当する領域を正確に決定することが重要となる。

アミノ酸配列からドメイン相当領域を予測するアプローチの一つとして、ドメインリンカー予測が挙げられる。ドメインリンカーとは、ドメイン同士をつなぐコイル領域のことである。ドメインリンカー予測では、ドメインリンカー領域の有する配列的特徴をもとに、ドメインリンカー領域の識別を行う。現在までに、ドメインリンカー予測法はいくつか提案されており、Web システムとして公開されているものとして DLP[1]や DomCut[2]などが挙げられる。

従来のドメインリンカー予測法では、共通してスライディングウィンドウ形式を採用している。そのため、ウィンドウ内の相対位置に基づくアミノ酸の位置依存性は考慮することができるが、ドメインリンカー領域全体における位置依存性を十分に考慮できない。

本稿では、ターゲット配列に対し二次構造予測を行い、決定したコイル領域に対し、ドメインリンカーの両端におけるアミノ酸の出現位置依存性を考慮したドメインリンカー予測法を提案する。提案手法では、ドメインリンカーの両端部分の配列プロファイルから抽出したアミノ酸の位置依存性をもとに、サポートベクターマシーンに学習させて

予測するというアプローチを採用する。また、提案手法の予測精度は、コイル領域の決定を行う二次構造予測法の精度に最も左右される。そこで、二次構造予測法の精度の影響を除いた提案手法の予測精度を測定するため、予備評価実験として、DSSP[3]で決定したコイル領域に対して、提案手法のドメインリンカー予測を適用させた。その結果、従来手法に比べ、Sensitivity、Specificity 共に約 20%向上することが可能であると確認した。

本稿では、以下の形を取る。２章では関連研究について述べる。３章では、配列プロファイルからアミノ酸の出現位置依存性を抽出する方法について述べる。４章では、新たなドメインリンカー予測法を提案する。５章では、データセットの作成方法について述べ、６章では、提案手法の予備評価実験について報告する。最後に本研究のまとめと考察を行う。

２． 関連研究

従来手法に共通している点として、スライディングウィンドウ方式を採用しているという点がある。具体的には、ウィンドウという固定長の部分配列単位で、予測対象領域の N 端から C 端まで１残基ずつスライドさせて、該当ウィンドウの中央部分がドメインリンカーに相当するか否かを予測する。

従来手法は、２種類に大別することができる。１つ目は、DLP に代表される、ウィンドウ内の各アミノ酸の種類および位置をもとにニューラルネットで予測を行う手法である[1]。この手法では、図１に示すように、アミノ酸Ｘとアミノ酸Ｙは１残基間隔で頻出するというような、ウィンドウ内の相対位置に基づく位置依存性を抽出できる。しかし、図２に示すように、ドメインリンカー領域のＮ端にはアミノ酸Ｘの出現頻度が高く、C 端にはアミノ酸Ｙの出現頻度が高いというような、ドメインリンカー領域全体における位置依存性が存在する場合を考える。各リンカーによって長さは異なるため、N 端のアミノ酸Ｘと C 端のアミノ酸Ｙの間に存在するアミノ酸の数も異なる。このため、N 端のアミノ酸Ｘと C 端のアミノ酸Ｙとの距離はリンカーによって異なり、固定残基長のウィンドウを利用する従来手法では、位置依存性をうまく抽出できない。

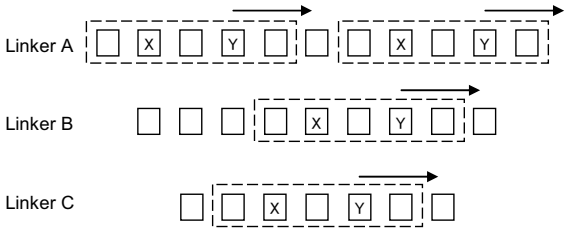

図 1　スライディングウィンドウにより抽出できる位置依存性

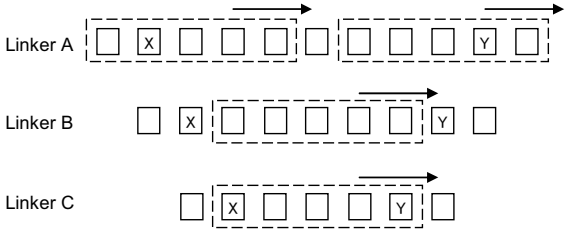

図 2　スライディングウィンドウでは抽出ができない位置依存性

２つ目は、DomCut に代表される、ウィンドウ内の各アミノ酸に対し、ドメインリンカー中における該当アミノ酸の出現頻度の重みをかけて算出したスコアの平均値を用いて予測を行う手法である[2,4-6]。この手法では、ドメインリンカー中における各アミノ酸の出現頻度のみに着目しており、各アミノ酸の位置依存性を考慮していない。

３．配列プロファイルからの出現位置依存性の抽出

配列プロファイルをもとにドメインリンカーの両端に出現するアミノ酸の出現位置依存性を抽出する方法を述べる。

３．１．配列プロファイルの作成方法

PSI-BLAST[7]を利用して、配列プロファイルを作成する。プロファイル作成には、NCBI-nr[8]データベースを用い、PSI-BLASTのパラメータは、E-value閾値を 10^{-3}、繰り返し検索回数を 20 と設定した。

３．２．出現位置依存性の抽出方法

本研究では、ドメインリンカー領域とループ領域の両端の４残基および直前または直後の５残基を含めた、ドメインとドメインリンカーの境界領域に着目して、アミノ酸の位置依存性の抽出を図る。ドメインリンカー領域とループ領域のＮ端と C 端における対象領域を図 4, 5 に示す。

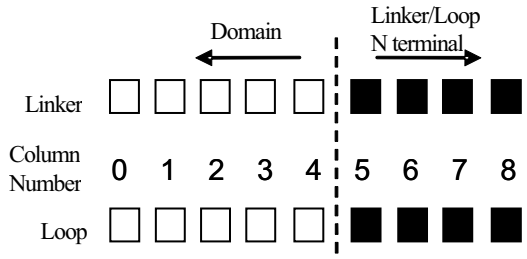

図４　N 端における各カラムと構造の位置の対応

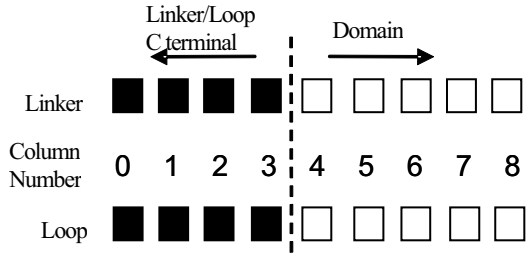

図５　C 端における各カラムと構造の位置の対応

PSI-BLAST による配列プロファイルの各アミノ酸のスコアの値が一般的には-7～7 となることに注意し、位置依存性を抽出する際、以下の２点を考慮した。１つ目は、マルチプルアラインメントの各カラムに頻出するアミノ酸情報のみを利用するため、プロファイルの各カラムについてスコアが１より大きいもののみ利用することである。２つ目は、２や３などの小さい値のスコアの影響も汲み取ることである。

以上を踏まえ、プロファイルの各カラムにおけるアミノ酸*j*のスコアx_jを（１）式により補正する。

$$\left.\begin{array}{l} \text{if } (x>1)\quad f(x_j)=\dfrac{1}{1+2\exp(-x_j/2)} \\ \text{if } (x\leqq 1)\quad f(x_j)=0 \end{array}\right\} \quad (1)$$

次に、補正したスコアを用いてリンカーまたはループの出現頻度を算出することにより、位置依存性を抽出する。具体的には（２）式を用いる。

$$P_{Xij}=\frac{\sum_{k=1}^{\#Coils} f(x_{ijk})}{\sum_{k=1}^{\#Coils}\sum_{j=1}^{20} f(x_{ijk})} \quad (2)$$

ここで、x_{ijk}はループ*k*またはドメインリンカー*k*のカラム*i*におけるアミノ酸*j*のプロファイルのスコア、*#Coils*はループ領域またはドメインリンカー領域の数を表す。

４．配列プロファイルを利用したドメインリンカー予測法の提案

リンカーとループの同一カラムにおいて、出現するアミノ酸が異なる場合、３ 章で抽出した位置依存性を利用してリンカーとループを識別することができると考えられる。そこで、３ 章の位置依存性の抽出方法を利用して、リンカーとループを識別するための特徴量を算出し、サポートベクターマシーン（SVM）に学習させることにより、ドメインリンカーを予測する方法を提案する。

４．１．特徴量算出

３．２．節で述べた、位置依存性の抽出方法をもとに、特徴量を算出する。具体的には、予測に利用する境界領域の９残基部分のプロファイルを利用して、各カラムに対応した特徴量の算出を行う。

カラム *i* における特徴量は、以下のように定義される。

$$F_i=\sum_{j=1}^{20}\left(L_{Xij}\times f(x_{ij})\right) \quad (3)$$

ここで、L_{Xij}はリンカー傾向性スコアであり、（４）式のようになる。

$$L_{Xij}=\log\left(\frac{P_{Xij}^{DL}}{P_{Xij}^{NL}}\right) \quad (4)$$

P_{Xij}^{DL} はドメインリンカーのカラム *i* におけるアミノ酸 Xj の出現頻度、P_{Xij}^{NL} はノンリンカー（ループ）のカラム *i* におけるアミノ酸 *Xj* の出現頻度である。

４．２．学習と特徴量選択

境界領域であるドメインリンカー領域とループ領域の両端部分の各カラムに対応した特徴量をもとに、SVM を利用して学習を行う。SVM のカーネルには RBF カーネルを利用する。

学習の前に特徴量選択を行う。まず、（２）式から計算されるリンカー・ループにおける各々のアミノ酸の出現頻度について相関係数を取り、相関係数が 0.6 を超える特徴量

を削除する。残った特徴量の中から、2 個の特徴量を選択するという方法を取る。具体的には、総当りで 2 個の特徴量を選び学習させ、最終的に最も精度が高くなる組み合わせを採用した。

5．データセット

6 章の予備評価実験に利用するデータセットの作成方法、および、作成したデータセットの解析結果について述べる。

5．1．データセットの作成

Tanaka らの論文[4]に基づき、ドメインリンカーを有するタンパク質のデータセットを作成した。

・手順 1（マルチドメインタンパク質の抽出）

タンパク質立体構造データベース SCOP[9]で定義されたドメインを複数有し、ドメインリンカーの候補となるドメイン境界部分のループの長さが 7 残基以上のタンパク質を抽出する。なお、DSSP[3]の定義に基づいて決定した二次構造に対し、本操作を行う。

・手順 2（構造的に独立でないマルチドメインタンパク質の除去）

手順 1 で抽出したタンパク質のうち、各ドメインが物理化学的相互作用しているタンパク質を除去する。ここで、物理化学的相互作用とは、疎水性相互作用、または、水素結合のことを指す。具体的には、各ドメイン間で 4.663Å 以内に疎水性残基の側鎖の炭素原子が 3 つ以上存在するか、2 つ以上の水素結合が存在するタンパク質を削除する。

・手順 3（代表タンパク質の決定）

手順 2 で抽出したタンパク質に対し、冗長性を取り除くために、代表タンパク質を決定する。具体的には、各タンパク質同士のBLAST[10]におけるp-valueが 10^{-7}以下となるように、代表タンパク質を決定する。

5．2．データセットの解析

データセットを作成した結果、106 個のマルチドメインタンパク質が得られ、ドメインリンカー領域は 112 個、7 残基以上のループ領域は 482 個であった。

ドメインリンカー領域とループ領域の長さの平均や標準偏差を測定した結果を表 1 に示す。また、ドメインリンカー領域とループ領域における各アミノ酸の出現頻度を図 3 に示す。リンカー領域とループ領域における各アミノ酸の出現頻度についてカイ二乗検定を行った結果、Asn、Gly、Lys、Pro の出現頻度には、有意水準 5%で有意差が検出された。

一方、Tanaka らの論文[4]では、上記の 4 種類のアミノ酸に加え、Asp、His の出現頻度においても、有意水準 5%で有意差が検出された。本データセットと論文[4]と比較して、リンカー領域とループ領域における出現頻度に有意差が検出されたアミノ酸に相違がある原因としては、SCOP や PDB の更新による違いに加え、代表タンパク質の決定方法や対象としたループの長さの違いが考えられる。論文[4]においては、代表タンパク質を決定する際に、nr-PDB[11]の定義する代表タンパク質のみを対象としていたのに対し、本研究では全てのタンパク質を対象とした。また、論文[4]では 4 残基以上のループを対象にしていたのに対し、本研究では 7 残基以上のループを対象にした。

表 1　リンカー領域とループ領域の長さの平均と標準偏差

	Linker	Loop
平均	12.21	10.07
標準偏差	5.61	4.24

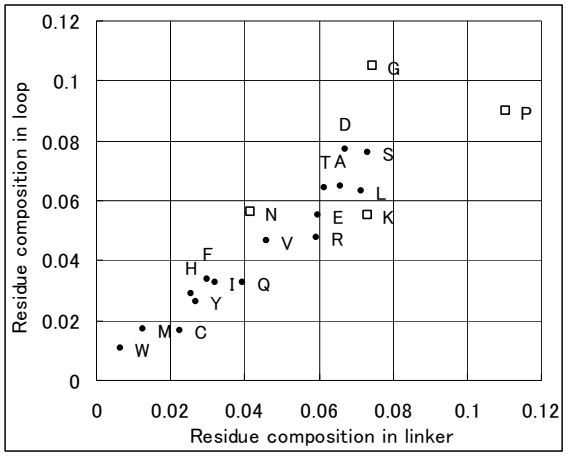

図 3 リンカー領域とループ領域における各アミノ酸の出現頻度（□はリンカー領域とループ領域での出現頻度についてのカイ二乗検定で有意差が検出されたことを表す）

6．予備評価実験

提案手法では、ターゲット配列に対し二次構造予測を行い、決定したコイル領域に対し、ドメインリンカーの両端におけるアミノ酸の出現位置依存性を考慮したドメインリンカー予測を行う。そのため、提案手法の予測精度は、コイル領域の決定を行う二次構造予測法の精度に最も左右される。二次構造予測法の精度の影響を排除した形で提案手

法の予測精度を確認するため、予備評価実験として、DSSPで決定したコイル領域に対して提案手法のドメインリンカー予測を行った。

６．１．予備評価実験条件

予測に利用する１本のアミノ酸配列以外の全てを学習に利用するというジャックナイフテスト形式により、評価する。また、αhelix続きの各端とβsheet続きの各端に、それぞれ１つのSVMを割り当て、リンカーまたはループであるか識別を行う。リンカー領域とループ領域の各端の内訳を、表２と３に示す。なお、調整可能なSVMのパラメータは、最適化問題の誤差項のCとRBFカーネルのγである。Cの値を 10^{-1}～10^{3}の範囲を 10^{-1}間隔で、また、γの値については 10^{-6}～10^{-1}に 10^{-6}間隔で変化させて、Sensitivityよりも Specificityを重視して最適パラメータを求める。

表２　リンカー領域の各端の内訳

	Helix 続き	Sheet 続き
N 端	78	34
C 端	59	53

表３　ループ領域の各端の内訳

	Helix 続き	Sheet 続き
N 端	206	276
C 端	222	260

また、精度の指標として利用する Sensitivity と Specificity は以下のように定義する。Sensitivity は、データセット中に含まれるリンカーの総数に対して正確に予測できたリンカーの総数の比率と定義する。Specificity は、予測したリンカーの総数に対して正確に予測できたリンカーの総数の比率と定義する。なお、予測したリンカーと実際のリンカーの位置が完全に一致した場合、正確に予測できたとみなす。

６．２．提案手法の予備評価実験結果

DSSP で決定した二次構造に対し、提案手法によるドメインリンカー予測を敢行した。なお、パラメータは、６．１で述べたように Sensitivity よりも Specificity を重視して設定し、精度測定を行った結果を表 4-6 に示す。

表４　最適パラメータの決定

	N 端		C 端	
	Helix 続き	Sheet 続き	Helix 続き	Sheet 続き
C	15	115	14	60
γ	6.0×10^{-6}	1.0×10^{-4}	1.0×10^{-6}	1.0×10^{-4}

表５　N 端の予測精度 (%)

	N 端	Helix 続き	Sheet 続き
Sensitivity	56.3	67.9	29.4
Specificity	44.3	46.9	34.5

表６　C 端の予測精度 (%)

	C 端	Helix 続き	Sheet 続き
Sensitivity	47.3	43.1	56.0
Specificity	45.2	44.6	45.9

また、Specificity の向上を図るためにリンカーの N 端側の予測結果とリンカーの C 端側の予測結果の間で積集合を取る操作を行う。具体的には、リンカーの N 端、C 端の両方をリンカーであると予測した場合に、対象コイル領域をリンカーであると予測する。さらに、Sensitivity の向上を図るため、同様にして和集合を取る操作も行う。具体的には、リンカーの N 端、C 端のいずれかをリンカーである予測した場合に、対象コイル領域をリンカーであると予測する。積集合と和集合を取った結果を表 7 に示す。

表７　N 端の予測結果と C 端の予測結果で積集合と和集合を取った場合の予測精度 (%)

	Intersection	Union
Sensitivity	25.0	78.6
Specificity	58.3	41.7

６．３．従来手法の精度評価

従来手法として、Web システムとして使用可能な DLP と DomCut の２手法を利用して精度の測定を行った。

提案手法では、DSSP で決定した二次構造に対して予測を行うため、ループとリンカーの識別に成功した場合、予測したリンカーと実際のリンカーの位置は完全に一致する。一方、DLP や DomCut では、アミノ酸配列の情報のみを利用して予測するため、予測したリンカーと実際のリンカーの位置を完全に一致させるのは困難である。そのため、提案手法に適用した精度評価方法では、DLP と DomCut

の精度が著しく低くなってしまう。そこで、論文[6]に述べられた評価方法を利用して、DLP や DomCut の精度評価を行う。具体的には、実際のリンカーが出現する前後 *n* 残基の領域を含めた範囲内に、予測したリンカーが１残基でも存在している場合は、正確に予測できているとみなし、精度を測定する。前後 *n* 残基の領域を含めて、論文[6]の精度評価方法を適用した DLP と DomCut の予測精度を表８に示す。なお、DLP のパラメータはデフォルトのままとしたが、DomCut の閾値は Sensitivity、Specificity ともに高精度な予測が可能な-0.09 に設定して精度を測定した[2]。

表８　前後 *n* 残基の領域を含めた精度評価方法を適用した DLP と DomCut の予測精度 (%)

	0 残基	10 残基	20 残基
Sensitivity DLP	27.7	34.8	39.2
Specificity DLP	19.0	23.9	27.0
Sensitivity DomCut	23.2	31.3	36.6
Specificity DomCut	15.2	20.6	24.1

６．４．提案手法と従来手法の比較

実際のリンカーが出現する位置の範囲内で、論文[5]の精度評価方法を利用して、提案手法を評価した場合、６．１．節で述べた精度評価方法と同じ結果となる。実際のリンカーが出現する位置の範囲内で、論文[6]の精度評価方法を適用した提案手法と従来手法の精度を表９に示す。この結果から、スライディングウィンドウを採用した従来手法に比べ、提案手法における各端の予測精度は Sensitivity、Specificity 共に約20%向上することが可能であると確認した。また、提案手法の Intersection 操作により Specificity が最も高く、Union 操作により Sensitivity が最も高くなることが可能であると確認した。

表９　提案手法と従来手法の精度 (%)

	N 端	C 端	IS	Union	DLP	DC
Sensitivity	56.3	47.3	25	78.6	27.7	23.2
Specificity	44.3	45.2	58.3	41.7	19	15.2

IS は InterSection、DC は DomCut を表す。

７．まとめと考察

本稿では、ドメインリンカーの両端部分に着目し、配列プロファイルから抽出した各アミノ酸の位置依存性をもとに、サポートベクターマシーンを用いて予測する方法を提案した。

提案手法の有用性として以下の２点が挙げられる。１点目は、プロファイルを利用することにより、アミノ酸の位置依存性や構造類似性をより正確に把握した予測を行うことができる点である。プロファイルを利用することにより、進化的に遠縁なタンパク質まで考慮したアミノ酸の出現位置依存性を抽出できると考えられる。また、ドメインリンカーの端、つまり、ドメインの端の部分は、ある程度構造が保存されているため、プロファイルには有意な情報が含まれていると考えられる。従って、ドメインリンカーの端部分のプロファイルをもとに SVM に学習させることで、ドメインリンカー同士の配列類似性が乏しい場合でも構造類似性が検出できると考えられる。

２点目は、ドメインリンカーの N 端と C 端の境界領域に着目することにより、ドメインリンカーの長さが多様であるという問題の克服やドメインリンカー領域全体における位置依存性の抽出が可能な点である。スライディングウィンドウ方式では、予測対象領域の N 端から C 端まで識別可能にする必要があるが、提案手法では N 端と C 端のみ識別可能にすればよいため、学習効率が向上すると考えられる。また、ドメインリンカー領域全体における位置依存性が存在する場合、ドメインリンカーの端部分を基準とした絶対位置に着目することで、より正確に抽出できると考えられる。

コイル領域の決定に利用する二次構造予測法の精度の影響を除いた提案手法の予測精度を確認するため、予備評価実験として、DSSP で決定したコイル領域に対して提案手法を用いてドメインリンカー予測を行った。その結果、従来手法に比べ、Sensitivity、Specificity 共に約 20%向上することが可能であると確認した。今後の課題としては、二次構造予測法で予測したコイル領域に対する提案手法の有効性について評価することが挙げられる。

謝辞

本稿の執筆にあたり、細部に至るまで熱心にご指導頂いた早稲田大学大学院理工学研究科の山田真介氏に感謝致します。

参考文献

[1] Miyazaki, S., Kuroda, Y., and Yokoyama, Y. : Characterization and prediction of linker sequences of multi-domain proteins by a neural network, J.Struct.Funct. Genomics, vol. 2, pp. 37-51 (2002).

[2] Suyama, M., and Ohara, O. : DomCut: prediction of inter-domain linker regions in amino acid sequences, Bioinformatics, vol. 19, pp. 673-674 (2003).

[3] Kabsch, W., and Sander, C. : Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers, vol. 22, pp. 2577−2637 (1983).

[4] Tanaka, T., Yokoyama, S., and Kuroda, Y. : Improvement of domain linker prediction by incorporating loop length-dependent characteristics, Biopolymers, vol. 84, 161-168 (2005).

[5] Dong, Q., Wang, X., Lin, L., and Xu, Z. : Domain boundary prediction based on profile domain linker propensity index, Comp.Biol.Chem., vol. 30, pp. 127-133 (2006).

[6] Dumontier, M., Yao, R., Feldman H.J., and Hogue C.W. : Armadillo: domain boundary prediction by amino acid composition, J.Mol.Biol., vol. 350, pp. 1061-1073 (2005).

[7] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. : Gapped BLAST and PSI-BLAST : a new generation of protein database search programs, Nucleic Acids Res., vol. 17, pp. 3389-3402 (1997).

[8] NCBI-nr : ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr..gz

[9] Lo Conte L., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. : SCOP database in 2002: refinements accommodate structural genomics, Nucleic Acid Res., vol. 30, pp. 264-267 (2002).

[10] Altschul, S..F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. : Basic local alignment search tool, J.Mol.Biol, vol. 215, pp. 403-410 (1990).

[11] nr-PDB: http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html