G-019

データ分布に基づく階層型多クラス分類を用いたタンパク質の細胞内局在部位予測

Hierarchical Multi-class Classification based on data distribution for Protein Subcellular Localization Prediction

久須田 樹哉 † **渡邉 真也** † Paul Horton ‡
Tatsuya Kusuda Shinya Watanabe Jianming Shi Paul Horton

1 はじめに

近年，実験技術の進歩により得られるようになったゲノム，トランスクリプトームなどの大量の実験データから何らかのマイニングを行うバイオインフォマティクスが新たな研究分野として注目されている．中でも，タンパク質の局在部位を予測することにより機能を推定する細胞内局在部位予測に基づく手法は，配列解析の典型的な問題として，数多くの手法が提案，公開されており，多くの研究者に実用的ツールとして利用されている [1, 2, 3, 4, 5].

アミノ酸配列に基づく細胞内局在部位予測としては，アミノ酸の出現頻度を利用した手法とアミノ酸配列に含まれる局在化シグナルに関する情報を利用した手法が提案されている．Horton らにより提案された WoLF PSORT[1, 2] は後者の代表的な手法であり，Cello[4] や MultiLoc[5] といった他の手法に比べ幅広い対象に対して優位な結果を示すことが報告されている [6].

しかしながら，WoLF PSORT を含めた既存の細胞内局在部位予測手法では比較的高い精度の多クラス分類が実現されている一方，分類結果からクラス間の関係および各クラスと特徴量の関係を推定することが困難なため，データの内部構造を理解するツールとしては不十分であるという問題点がある．

そこで本研究では，細胞内局在部位予測において詳細な問題分析を実現するための新たなアプローチとして各クラスのデータ分布に基づく階層型分類手法の提案を行う．提案する手法では，データの分布に基づき段階的に大まかな分割から徐々に詳細な分割を実現する階層的な分類メカニズムに基づいている．具体的には，類似データを同一のグループとして扱いグループ間分離を最大化するような分類を繰り返し，全体として木構造分類の生成を行っている．提案する分類分けの実現により，各クラス間の近接度合いや各クラスを特徴づけている特徴量の種類と重みづけの情報をより詳細に提示できると考えている．

提案する階層化の効果を検証するために Swiss Prot[7] から引用した菌類データ（14 クラス，データ数 2158）に対して通常の WoLF PSORT と階層化のメカニズムを組み込んだ場合の比較実験を行った．

2 タンパク質の局在化および局在部位予測

タンパク質は DNA から転写により mRNA の形に変換され，つぎに mRNA のもつ塩基配列情報に沿い翻訳の過程を経てタンパク質に合成される [1, 2, 8]．各タンパク質はその機能を果たすために種類に応じて細胞質, 細胞膜といった, 細胞内小器官の適切な場所に運ばれる．このことをタンパク質の局在化と呼ぶ [9]．局在部位は，局在化シグナルと総称されるタンパク質を構成するアミノ酸配列に含まれる特徴的な配列の情報により決定することが知られている．

ここでは，代表的な局在化シグナルの 1 つであるシグナルペプチド [8, 10, 11] を例に説明する．シグナルペプチドは，小胞体への輸送を指示する配列であり膜透過を伴う局在化において必須のものである．そして，シグナルペプチドは N 末端に存在する数十残基の長さを持つ配列であり次に示す 3 領域において共通した構造を持つ. (1)N 末端側には，1,2 残基の塩基性のアミノ酸があり，(2)7 から 15 残基の疎水性のアミノ酸が多い領域がある．最後に，(3)C 末端側には極性のあるアミノ酸配列が存在し，1 番目と 3 番目には小さく非極性のアミノ酸が含まれる．局在部位予測では，このような特徴に基づいて局在化シグナルを予測し利用している．

以上のように，局在部位予測とはアミノ酸配列からタンパク質の局在部位を予測しその働きを推定しようとするもので，一般にパターン分類問題として定式化される．基本的に局在部位を決定する情報はアミノ酸配列に含まれるため，理論上はアミノ酸配列が与えられれば局在部位を予測することが可能である．

† 室蘭工業大学，Muroran Institute of Technology
‡ 産業総合研究所 生命情報工学研究センター, AIST CBRC

3 WoLF PSORT

本論文では，既存の局在部位予測としては最も高精度なツールの 1 つである WoLF PSORT に注目し提案する手法を組み込む実験を行った．ここでは，WoLF PSORT[1, 2] の概略について説明する．

WoLF PSORT は，重み付き kNN 法（k-Nearest Neighbor）に基づく局在部位予測を実現しており，Nakai らにより開発された PSORT[12]，Bannai らにより開発された iPSORT[13] の特徴量および局在部位と相関するアミノ酸組成の 2 種類の情報を特徴量として利用している．

WoLF PSORT のアルゴリズムを図 1 に示す．

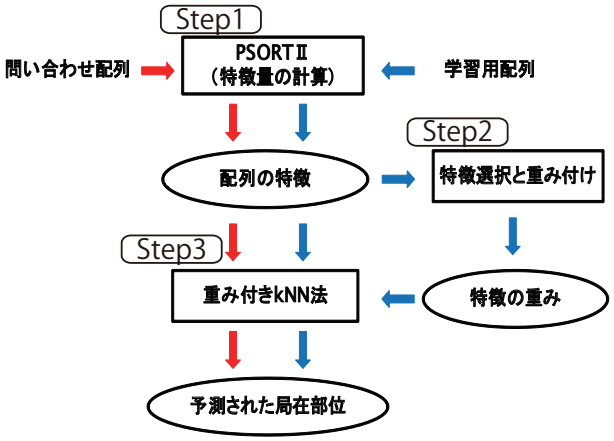

図 1: WoLF PSORT システム

図 1 から分かるように WoLF PSORT は次の 3 つのステップに基づいている．

Step1: 入力された問い合わせ配列と学習用データセットの特徴量を計算する．

Step2: ビームサーチを用いて特徴選択と重み付けを行う．

Step3: 重み付き kNN 法を用いて局在部位を予測する．

WoLF PSORT は K-NN 法に基づく 1 層型多クラス分類を行っている．WoLF PSORT についての詳細は参考文献 [1, 2] を参照．WoLF PSORT では比較的高精度な識別を実現しているが, 対象とするクラス数が多数であるため各クラスの関係を推定することが難しい上，唯一の定義に基づき各クラス間の距離（類似度）が決定されるため詳細なクラスと属性の関係を把握することができない．そのため，データの内部構造を理解するためのツールとしては不十分である．

そこで次章ではデータ分布に基づく階層型分類を提案し，上記の問題点の解決を試みる．

4 階層型多クラス分類

提案する分類手法では，クラス間の類似度に基づき段階的に大まかな分割から徐々に詳細な分割を実現する階層的な分類メカニズムを実現している．

提案手法の概念図を図 2 に示す．

図から分かるように，提案手法では類似している学習用データをグループ化し分割する事で一度に識別する数を減らし，1 つの分類自体を単純化している．図中の例では，1 層目においてクラス 1・3 のグループとクラス 2・3・4 のグループを変数増加法により求めた最適な特徴空間で分割し，2 層目においてクラス 1 とクラス 3，クラス 2 とクラス 3・4 を分類，最下層の 3 層目ではクラス 2 とクラス 3・4 を分類している様子が示されている．

重要なのは，階層ごと，分類器ごとに用いている特徴空間（類似度の定義）が異なっているため，それぞれの分類において特徴づけに重要な役割を果たしている属性を明確化できる点である．また，下の階層ではより詳細な分類が実現されることになるため，一度の多クラス分類ではデータ数の多いクラスに埋もれてしまうデータ数の少ないクラスも精度よく識別する事が期待できる．

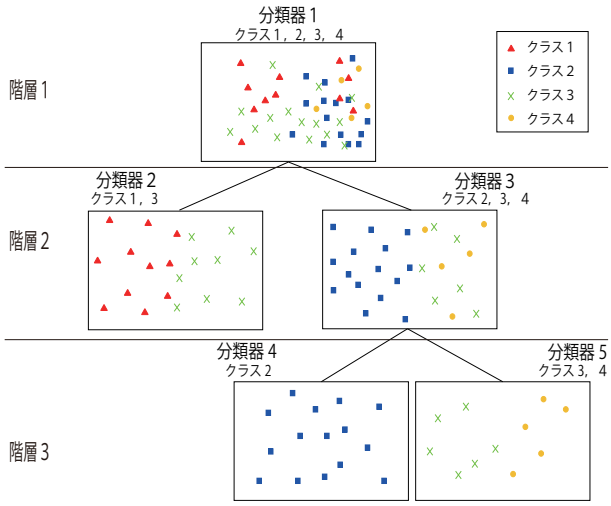

図 2: 提案手法の概念図

4.1 アルゴリズム

提案する階層型分類手法のアルゴリズムの流れについて説明する．本手法は，図 3 に示すように 7 つのス

テップに基づいている．各ステップの手順を以下に示す．

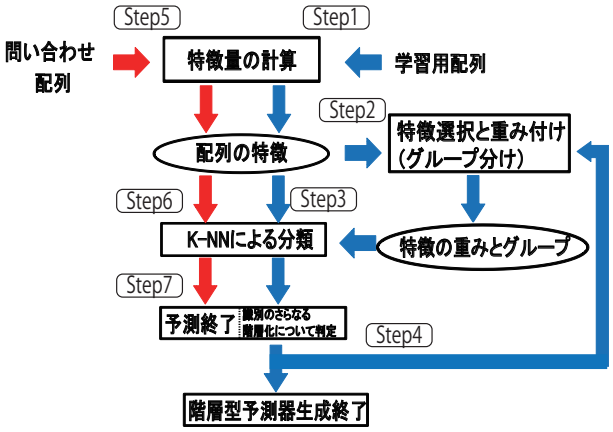

図 3: 階層型多クラス分類手法のシステム

Step1: 学習用データセットの特徴量を計算．

Step2: 得られた特徴量によるデータ分布からデータをグループ化し，グループ間の分割を実現する最良な特徴空間を算出（特徴選択とグループ分け）．

Step3: Step2 で求めた類似度（選択した特徴量および重み）に基づく K-NN 分類を実行．

Step4: Step3 で分類されたデータに対して，下階層の分類を行うかを判定．分類を行う場合には Step2 へ，そうでなければ階層型学習器の生成を終了．

Step5: 問い合わせ配列の特徴量を計算．

Step6: Step4 により得られた階層型学習器に対して，問い合わせ配列を入力．各階層において問い合わせ配列がどの分岐に属するか K-NN 分類により識別．

Step7: 最下層において得られた結果を最終的な推測結果として出力．

上述の識別手法は基本的に WoLF PSORT の手順を踏襲している．さらに，WoLF PSORT と同様に特徴量は全て「0〜1」に正規化したものを使用している．WoLF PSORT と異なっている部分は，階層化によって分割を複数回行っている事とグループ分けを行うことで分けやすいクラスから分割するようにした事である．次節では，Step2 のグループ分けについて説明する．

4.2 グループ分け

階層型多クラス分類手法では類似したクラスをグループに統合し，グループ間の分類を行う．グループ分けでは，クラス間の分離度から同一グループとなるクラス（同一グループとする分離度の値 β は任意の値を設定.）を探索する．分離度は 2 つのクラス間の重なり度合いを評価するための指標で，式（1）で求めている．中心間距離は各クラスの平均値間の距離を使用し，半径には 2 つのクラスの半径の内の最大値をとることにする．

階層型多クラス分類手法の Step2 のグループ分けの概念図を図 4 に示す．

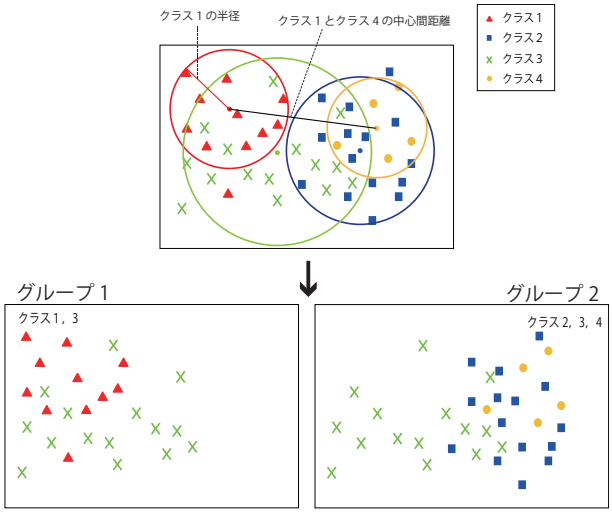

図 4: グループ分けの概念図

以下にグループ分けのアルゴリズムを示す．

Step1: 各クラスにおいて，データの α%（データ含有率）以上を含む円を作成．

Step2: 各クラスの円の中心間距離を求める．

Step3: クラス間の分離度を以下の式で求める．

$$\text{分離度} = \frac{\text{中心間距離}}{\text{半径}} \qquad (1)$$

Step4: 最も離れた 2 つのクラス（分離度が最大）それぞれに対して分離度が β (グループ判定，$0 \leq \beta \leq 1$) 以下のクラスを同一グループとする．

Step5: 両方のグループに対して同一グループとされたクラス（重複クラス）を探す．

Step6: 重複クラスの個数が，重複許容率以上の場合，

重複クラスの中で，分離度が高いクラスを片方のグループから削除する．

Step7: グループに含まれたクラスの内，重複してないクラスに対して分離度が β 以下のクラスも同様に同一グループとする．

Step8: 両グループに含まれなかったクラスで Step4〜7 を行う．

このグループ分けメカニズムでは，類似したクラスを同一のグループに統合しグループ間分類を行う事で，大まかな分割から段階的に詳細な分割を行う階層型分類を実現している．

5 数値実験

本実験では，階層型多クラス分類手法に対して Swiss Prot[7] から引用した菌類データを適用し，細胞内局在部位予測の精度向上とタンパク質のデータの解析の観点から，階層型多クラス分類手法の有効性を検証する．以下では，対象問題とパラメータ，実験結果とその考察について示す．

5.1 対象問題

実験の対象問題としては，Swiss Prot[7] から引用したタンパク質の菌類データ（クラス数 14，データ数 2158，特徴数 56）を用いる．

5.2 パラメータについて

階層型多クラス分類手法で使用しているパラメータは，以下の 6 つが存在する．

- 重複許容率：グループ分けの時に複数のグループに重複するクラスの割合．
- データ含有率 α：円を作成する時に円に含むデータの割合．
- グループ判定 β：グループ分けの時に同一グループとする分離度の値．
- ビーム幅：特徴選択の時に保持する特徴の組み合わせの数．
- 交差検定数：特徴選択の時の精度計算に使用する交差検定の分割数．
- k 値：kNN 法における投票数．

6 つのパラメータの値は，前もって調べた予測精度が最も高くなった時の値を使用する．本実験で使用する各パラメータの値を表 1 に示す．

表 1: 使用したパラメータ

パラメータ	値
重複許容率	0.2
データ含有率 α	0.8
グループ判定 β	0.8
ビーム幅	3
交差検定数	10
k 値	10

5.3 実験結果

本実験により得られた識別結果を図 5 に示す．この図では，各階層のグループに含まれる学習用データの数とグループ分けによって所属するべきクラスの id（クラスの id と局在部位の関係は表 2 を参照）を示す．また，より詳細な結果について分析するため，クラス毎のデータ数と予測精度の結果を表 2 に，分岐 1 と分岐 4 で選択された特徴とその重みを表 3 に示す．

5.3.1 予測精度の考察

階層型多クラス分類手法のデータ全体の予測精度は WoLF PSORT の予測精度を若干上回ったが，大きな違いは見られなかった．

クラス毎の予測精度について見た場合，データ数の少ないクラスのほとんどに精度の向上が見られた．特に細胞骨格の予測精度には大幅な上昇を見る事ができる．しかし，データ数の多いクラスは予測精度が減少する傾向にある事が分かった．

5.3.2 データ解析

得られた分類器を生物学的な見地で分析すると，図 5 の分岐 1 では大体分泌経路に局在するクラスとそれ以外のクラスに大別されている事が分かる．さらに，分岐 4 ではミトコンドリアに局在するクラスとそれ以外のクラスが分類されている様子が読み取れる．

また，分岐毎に選択された特徴を見ると，分岐 4 ではミトコンドリアの局在化シグナルの存在に対する特

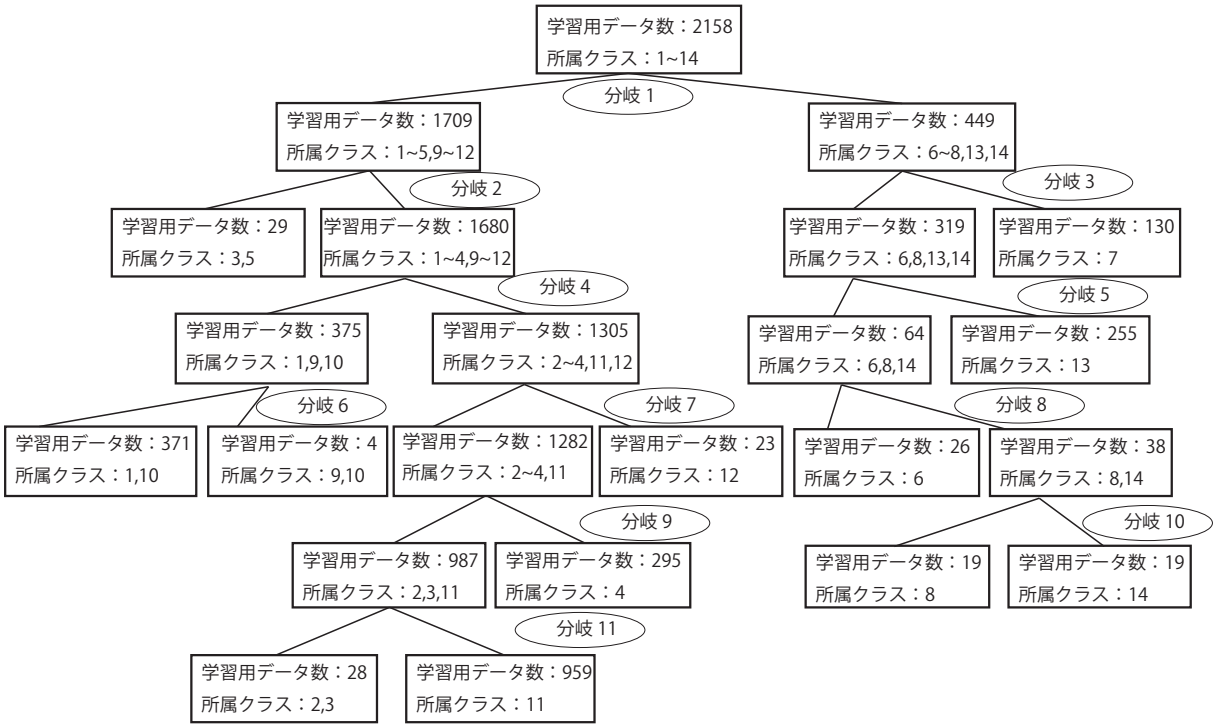

図 5: 識別結果の階層図

表 2: 階層型多クラス分類手法のクラス毎の予測精度

id	局在部位	データの数	WoLF PSORT	階層型分類
1	細胞質とミトコンドリア	8	0.00 %	0.00 %
2	細胞質と核	91	19.78 %	14.29 %
3	細胞質とペルオキシソーム	2	0.00 %	0.00 %
4	細胞質	354	69.21 %	57.06 %
5	細胞骨格	34	35.29 %	**85.29 %**
6	小胞体	66	10.61 %	**36.36 %**
7	細胞外	140	96.43 %	87.14 %
8	ゴルジ体	38	10.53 %	**26.32 %**
9	ミトコンドリアと核	4	0.00 %	0.00 %
10	ミトコンドリア	435	81.84 %	75.63 %
11	核	666	86.19 %	**92.94 %**
12	ペルオキシソーム	77	6.49 %	**23.38 %**
13	細胞膜	220	95.91 %	92.27 %
14	液胞膜	23	0.00 %	**26.09 %**
	総合	2158	72.61 %	72.98 %

徴の mit が選択され，分岐 7 ではペルオキシソームの局在化シグナルの存在に対する特徴の pox が選択されており，分類結果もそれらを基準としたものになっているため，大まかに納得のいく特徴選択が行われている事が分かった．

以上の事から，階層型多クラス分類手法は生物学的に意味がある分類が行われていると期待できる．そのため，階層化によりデータ解析のために比較的有意義な情報が得られると思われる．

6 おわりに

本研究では，データ分布に基づく階層型多クラス分類を提案し，タンパク質の細胞内局在部位予測を行う WoLF PSORT に対して提案手法を導入した．この手法の有効性を検証するために WoLF PSORT で使用されている菌類のタンパク質のアミノ酸配列データに対して実験を行った. その結果，生物学的に意味のある分類が行われている分岐が見られたため，クラス間や各クラスと属性の関係の明確化，内部構造の分析に役立つと期待できる結果が得られた．しかし，識別精度では提案手法により僅かな精度向上が見られたが大きな改善は見られなかった．

タンパク質データの解析においては，階層化によって分類の回数が増えるため，WoLF PSORT より多くの

情報が得られる．さらに，分岐毎の選択された特徴やクラス間の近接度合い等の有意義な情報が得られると思われる．

今後は菌類以外の動物，植物のタンパク質データに対しても実験を行う必要がある．

表 3: 分割に使用した特徴（分岐 1，4）

分岐 1		分岐 4	
特徴	重み	特徴	重み
Acont	1	*Dcont	2
Rcont	1	*Econt	2
Dcont	1	Hcont	1
*Econt	2	Kcont	1
Gcont	1	Scont	1
Icont	1	Wcont	1
*Kcont	2	act	1
Ycont	1	alm	1
Vcont	1	caa	1
*length	2	erl	1
*alm	2	*mH0_29_12	2
bac	1	*mit	2
*dna	2	*nuc	2
erl	1		
*gvh	2		
leu	2		
*mH0_29_12	2		
*mit	2		
myr	1		
pox	1		
*psg	2		
rib	1		
tms	1		

参考文献

[1] Paul Horton and et al. Wolf psort: Protein localization predictor. *Nucleic Acids Research*, Vol. 35(Web Server issue), pp. W585–W587, 2007.

[2] Paul Horton, Keun-Joon Park, Takeshi Obayashi, and Kenta Nakai. Protein subcellular localization prediction with WoLF PSORT. In Tao Jiang, Ueng-Cheng Yang, and Yi-Ping Phoebe Chen, editors, *Proceedings of the 4th Annual Asia Pacific Bioinformatics Conference, APBC06*, pp. 39–48. Imperial College Press, London, 2006.

[3] Masahiro Gomi, Masashi Sonoyama, and Shigeki Mitaku. High performance system for signal peptide prediction: Sosuisignal. *Chem-Bio Informatics Journal*, Vol. 4, No. 4, pp. 142–147, 2004.

[4] Kuo-Chen Chou and David W.Elrod. Protein subcellular location prediction. *Protein Engineering*, Vol. 12, No. 2, pp. 107–118, 1999.

[5] Annette Höglund, Pierre Dönnes, Torsten Blum, Hans-Werner Adolph, and Oliver Kohlbacher. Multiloc: prediction of protein subcellular localization using n-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, Vol. 22, No. 10, pp. 1158–1165, 2006.

[6] Wenfeng Qian and Jianzhi Zhang. Protein subcellular relocalization in the evolution of yeast singleton and duplicate genes. *Genome Biology and Ebolution*, pp. 198–204, 2009.

[7] Boeckmann B, Bairoch A, Apweiler R, and et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *NUCLEIC ACIDS RESEARCH*, Vol. 31, No. 1, pp. 365–370, 2004.

[8] 琢吉岡, 信石井. サポートベクトルマシンによるタンパク質局在部位の予測. 電子情報通信学会技術研究報告. NC, ニューロコンピューティング, Vol. 101, No. 364, pp. 63–70, 20011011.

[9] 金久實. ポストゲノム情報への招待. 共立出版, 2001.

[10] McGeoch. On the predictive recognition of signal peptide sequences, 1985.

[11] Nakai and Kanehisa. Refinement of the prediction methods of signal peptides for the genome analyses of saccharomyces cerevisiae and bacillus subtilis, 1996.

[12] Nakai K and Kanehisa M. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, Vol. 14, No. 4, pp. 897–911, 1992.

[13] H. Bannai, Y. Tamada, O. Maruyama, K. Nakai, and S. Miyano. Extensive feature detection of n-terminal protein sorting signals. *Bioinformatics*, Vol. 18, No. 2, pp. 298–305, 2002.