ゲノム機能解析のためのタンパク質細胞内局在の自動認識*

3U－5

蕪山典子 後藤敏行 影井清一郎（横浜国立大学）†、立野玲子（東京都臨床医学総合研究所）‡、
菅野純夫 富樫卓志（東京大学医科学研究所）§、清水哲男（富士通株式会社）¶

1．はじめに

ヒトゲノムの研究はゲノムの解読からゲノムの機能解析へと移行しつつあり、その第一段階としてゲノムとタンパク質が局在する細胞内小器官の関係が注目されている。ゲノム機能解析では、多数の cDNA[1]を網羅的に処理することが要求されており、これまでの人手や目視による判定から計算機を用いた高速処理が望まれている。

筆者らはこのような背景から、画像処理を利用したタンパク質局在の自動識別アルゴリズムの検討を進めている。

2．タンパク質局在

図1が今回対象とした画像の一例であり、小胞体(ER)、ゴルジ体(GOL)、細胞膜(MEM)、ミトコンドリア(MITO)、核(NUC)、ペルオキシソーム(PER)の6つの細胞内小器官に蛍光タンパク質が局在した顕微鏡画像である[2]。蛍光タンパク質顕微鏡画像は、視野内に複数の細胞が存在する場合、焦点に合致した明瞭な細胞像と焦点から外れた不明瞭な像が混在するという問題がある。さらに、生きた細胞を対象とするため、細胞毎にタンパク質の生成の進み方が異なる。この結果、図2に示すように、同じ小器官に局在する場合でも、画像に差異を生じるという問題もある。

3．認識アルゴリズム

図3に本手法の処理フローを示す。本手法では、焦点外れやタンパク質生成の進み方による不明瞭サンプルの存在を考慮して、各細胞内小器官ごとに複数クラスを割り振り、部分空間法[3]をベースとした認識手法を採っている。

部分空間法は、統計的認識手法の1つであり、最初に学習により特徴空間上に各クラスごとの部分空間を生成する。次に認識において、サンプルの特徴ベクトルと部分空間との距離を判定することによりクラスに類別する。部分空間作成のための特徴ベクトルは、今回は注視点（局在領域の中

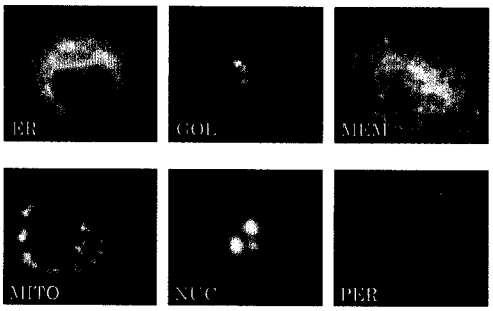

図1．細胞内小器官へのタンパク質局在画像

図2．不明瞭サンプルの一例

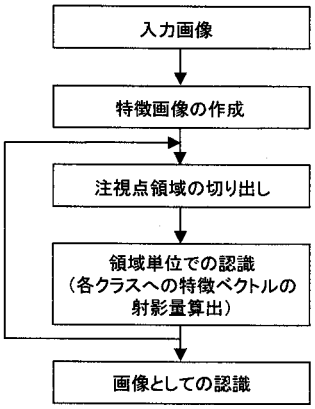

図3．認識処理の流れ

*Pattern recognition of localization of proteins for functional genomics
† Noriko Kabuyama, Toshiyuki Gotoh, Seiichiro Kagei (Yokohama National University)
‡ Reiko Tachino (The Tokyo Metropolitan Institute of Medical Science)
§ Sumio Sugano, Takushi Togashi (The Institute of Medical Science, The University of Tokyo)
¶ Tetsuo Shimizu (Fujitsu Limited)

心点）を中心とした矩形領域の画素濃度値を使用した。また、計算時間を軽減のために原画像の小領域（$2^n \times 2^n$）単位で特徴抽出を行っている。認識のクラスは、不明瞭な局在やタンパク質生成の進度による変化にも対応できるよう６つの局在に加えそれぞれ不明瞭な局在のクラスを用意した。さらに、タンパク質生成の進度による変化が顕著なペルオキシソームは生成進度の異なるクラスを別に用意し計 13 クラスとした。

今回の培養条件では、同一細胞に 1 種類の cDNA を注入したものであることから、画像内では同一の細胞内局在を示すことが想定されている。本手法では、複数ある領域について領域単位での認識を行った後に、その結果を総合判定して画像としての認識結果を推定している。

４．実験結果

認識実験は、局在部位が既知の６つの cDNA をヒトのガン細胞由来の HeLa 細胞に注入し、約 12 時間の培養後に撮影を行った。６つの局在それぞれ 160 枚（計 960 枚）撮影し、そのうち１局在当り 75 枚（計 450 枚）の画像を学習サンプルに使用した。残りの 510 枚は未知の画像として認識実験を行った。

入力画像から 4×4 画素領域の濃度平均値及びエッジの平均強度をそれぞれ１画素に置き換えた 2 枚の特徴画像を作成した。注視点を中心とした 24×24 画素領域の 2 枚の特徴画像の画素濃度値を特徴データとして、24×24×2=1152 次元の特徴空間内で 13 のクラスの 16 次元の部分空間への射影量をそれぞれ求め比較し、その領域の認識結果を出力した。1 枚の画像内の複数ある領域で異なった局在の認識結果が出た場合、今回の実験ではそれらの中でもっとも射影（ベクトルのノルムで正規化）が大きい認識クラスを最も信頼性があると判断し、その画像の認識結果としている。

表１および表２に認識実験の結果を示す。表１は、細胞内小器官に対応した６種のクラスを用いて実験した結果であり、GOL、NUC、MITO は 90%近く認識できているが、ER、MEM、PER に関してはこれらのグループ内では誤認が多いことが分かる。表 2 は不明瞭サンプルを考慮した 13 クラスを用いて実験した結果であり、全般的に誤認識の確率が減少している。

表１．認識結果（６クラス）

対象＼結果	ER	GOL	MEM	MITO	NUC	PER	認識率 A (%)
ER	72(枚)	2	4	0	0	7	85
GOL	0	82	0	0	3	0	96
MEM	19	3	44	0	0	17	53
MITO	3	0	0	72	3	7	85
NUC	0	2	0	0	82	0	98
PER	22	0	17	1	2	40	49
認識率 B (%)	62	92	68	99	91	56	

表２．認識結果（１３クラス）

対象＼結果	ER	GOL	MEM	MITO	NUC	PER	認識率 A (%)
ER	77(枚)	2	3	0	0	3	91
GOL	0	84	0	0	1	0	99
MEM	22	1	51	0	0	9	61
MITO	2	1	0	76	4	2	89
NUC	0	3	0	0	81	0	96
PER	26	0	17	0	2	37	45
認識率 B (%)	61	92	72	100	92	73	

５．まとめ

細胞内小器官へのタンパク質局在画像を対象とした認識手法の検討を行い、GOL、NUC、MITO への局在に対して 90%程度の認識が可能であることを確認した。今後の課題は、今回の手法で誤認が多かった ER、MEM、PER に対する認識手法の検討、今回は未検討であった画像データからの局在領域の注視点検出法の検討などである。

謝 辞

本研究は、NEDO 委託研究「ゲノム機能解明のための細胞画像自動解析システムの研究開発」によるものであり、関係各位に深く感謝する。

参考文献

1) 金久：遺伝子とゲノムの情報処理，情報処理，Vol.35, No.11, pp.983-990, Nov.1994.
2) J.C.Simpson, et.al. :Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing, EMBO Report, Vol.1, No.3, pp.287-292, 2000.
3) オヤ：パターン認識と部分空間法，産業図書，1986.