SOMによる蛋白質情報を用いた癌細胞の分類手法

北風裕教*　松岡秀実**　池田信彦***　松野浩嗣****

Classification of cancer cell using protein information by SOM

Hironori KITAKAZE, Hidemi MATSUOKA, Nobuhiko IKEDA and Hiroshi MATSUNO

Abstract

The analysis of cancer cells at the level has been studied. However, it was difficult to clarify the characteristics of cancer cells such as the speed of becoming worse, the efficiency of the medicine on the cell level. Recently, Laser Scanning Cytometer(LSC) which measures the data of cancer cells has been developed. Dr.Furuya et al. tried to analyze the characteristics of cancer cells using to the amount and the cohesiveness of protein. However, they could not find the efficient method for classification of cancer cells.

In the paper, we try to classify cancer cells using the amount and the cohesiveness of protein by Self-Organizing Map (SOM). The data used in the classification is the image data created from the amount and the cohesiveness of protein of cancer cell extracted from the patient by LSC. The result shows the probability that SOM would be able to classify cancer cells by those protein information.

Key Words:　cancer cells, LSC, amount of protein, cohesiveness of protein, SOM, classification

1.　緒言

癌の検査，治療方針の決定，さらには治療法および治療薬開発のために，患者から癌組織および細胞を摘出し，組織および細胞レベルで分析することが医学，薬学，生物学の分野で行われている．癌はその形態の違いからある程度は分類することが可能であるが（組織型の決定），同一部位から摘出した同じ組織型の癌であっても，進行の速さ，治療薬に対する反応が一様ではないことがしばしばあり，形態学的特徴により決定される組織型分類よりさらに詳細な分類が医療分野では求められている．

近年になって癌細胞に一次抗体を反応させた染色癌細胞に，特殊なレーザを照射し，反射された一次抗体の蛍光量を測定する Laser Scanning Cytometer(LSC)が開発され（図 1.1），多種多様な癌細胞のデータ取得が行われている（図 1.2，図 1.3）．古屋・佐々木らは，LSC により抽出した多種の癌細胞のデータから蛋白質の量とその凝集度に着目した分析を行い，顕微鏡では分類が困難な癌細胞について分類[1]を試みている．しかし，未だに有効となる分類法を見出すまでには至っていない．松岡らは，これまでに，古屋らの手法から得られた蛋白質情報から画像処理により特徴付けを施し，最大エントロピー法（MEM）手法を用いてスペクトルによる分類を試みている．しかし目標とする結果までには至らなかった[2].

そこで本研究では分析手法として自己組織化マップ（SOM）による分類を試みた．SOM とは，T.Kohonen が提案したニューラルネットワークにおける教師なし学習モデルの一つである[3,4]．これは学習により入力データの類似度を自動的に見出し，似た入力同士をマップ層の近くに配置するトポロジカルマップを形成するものである．本研究では，このトポロジカルマッピングにおいて，同種の癌細胞データでも,蛋白質情報による差異を抽出できるのではないかと考え，LSC により抽出した癌細胞データから，蛋白質の量とその凝集度の関係を表した画像データを作成し，その距離情報を学習によりマップ化して分類を試みた．その結果，同種の癌細胞においても蛋白質の違いによってマップ上で分類を行うことができる可能性を示すことができたので報告する．

*商船学科 **みい眼科(医療事務) ***徳山工業高等専門学校 ****山口大学大学院
2008 年 9 月 30 日受付

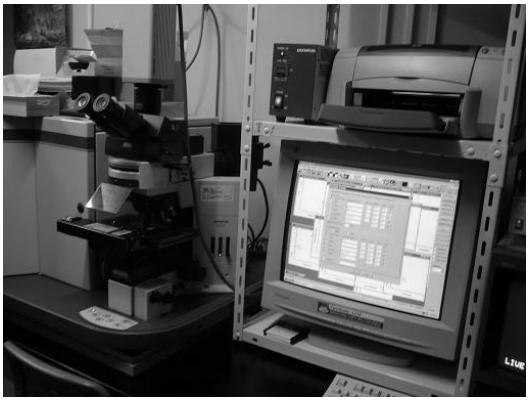

図 1.1　Laser Scanning Cytometer（LSC）システム

```
laser 8=1
laser 9=1
mark multiple cells=yes
mark partial cells=no
maximum sample count=2000
mirror synch pulse delay=950
name=DNA
number of channels=2
peak channel=1
pixels=768
pmt voltage 10=0
pmt voltage 11=0
pmt voltage 1=19
pmt voltage 2=10
pmt voltage 3=20
pmt voltage 4=20
pmt voltage 5=0
pmt voltage 6=24
pmt voltage 7=20
pmt voltage 8=20
pmt voltage 9=20
rate=96
sample id=1
scatter title=Fwd Scatter
specimen data= 04GI No.1 CEA
step hz=100
use cosine=yes
version=301
x move hz=50
x segmentation slop=4
y step=368
                                                                    |--
#,      X,      Y, Y Pixel, Area, Perim,    DI, Mult,         Annotation,
1,  58750,   6122,     671,  326,   157,  0.00,   No,           A cell,
2,  58678,   6250,     172,   41,    49,  0.00,   No,           A cell,
3,  58698,   6209,      86,   72,    80,  0.00,   No,           A cell,
4,  58364,   6371,     425,   66,    64,  0.00,   No,           A cell,
5,  58374,   6378,     440,  117,    84,  0.00,  Yes,           A cell,
6,  58586,   6365,     412,   44,    54,  0.00,   No,           A cell,
7,  58694,   6380,     443,  954,   691,  0.00,  Yes,           A cell,
8,  58722,   6359,     399,   53,    62,  0.00,   No,           A cell,
```

図 1.2　Laser Scanning Cytometer（LSC）から抽出された癌細胞のデータ

	A	B	C	D	E	F	G	H	I	J	K
1	#	X	Y	YPixel	Area	Perim	DI	Mult	Annotation	Integral(G)	MaxPixel(G
2	1	31812	5923	260	127	108	0.52	No	AcelI	985844	2520
3	2	31919	6069	560	122	108	0.58	No	AcelI	1139509	3700
4	3	32089	6036	492	78	79	0.34	No	AcelI	629009	2105
5	4	32483	6117	663	42	55	0.19	No	AcelI	342233	1700
6	5	32551	6045	511	54	68	0.23	No	AcelI	424662	1788
7	6	30964	6210	90	56	62	0.23	No	AcelI	452920	2153
8	7	30809	6412	510	142	93	0.86	No	AcelI	125882	430
9	8	32508	6463	614	53	64	0.23	No	AcelI	397994	2126
10	9	32536	6603	141	45	63	0.19	No	AcelI	354567	1945
11	10	32580	6686	314	57	68	0.25	No	AcelI	460509	2279
12	11	32780	6778	507	50	59	0.22	No	AcelI	393781	2075
13	12	32855	6771	493	44	60	0.22	No	AcelI	377426	2175
14	13	30951	7020	243	194	134	1.18	No	AcelI	222293	704
15	14	31044	7056	317	136	143	0.54	No	AcelI	614685	1320
16	15	32414	7034	271	118	101	0.45	No	AcelI	797629	1683

図 1.3　癌細胞データから必要箇所を抽出したデータ

2. 実験手順と方法

図 2.1 にシステム全体の流れ図を示す．これらについて以下に詳細を示す．

2.1. 分析用の癌細胞データ

2.1.1. 利用する癌細胞

本研究は，専門業者から購入され，特殊な培養方法で培養された癌細胞のデータを利用している．表 2.1 に本研究で利用する癌細胞を示し，癌細胞と反応させる一次抗体を表 2.2 に示す．一次抗体は癌細胞と反応させることによって，特定の蛋白質と反応する特性がある．従って，一次抗体と癌細胞を反応させることによって蛋白質の特性を調べることができる．

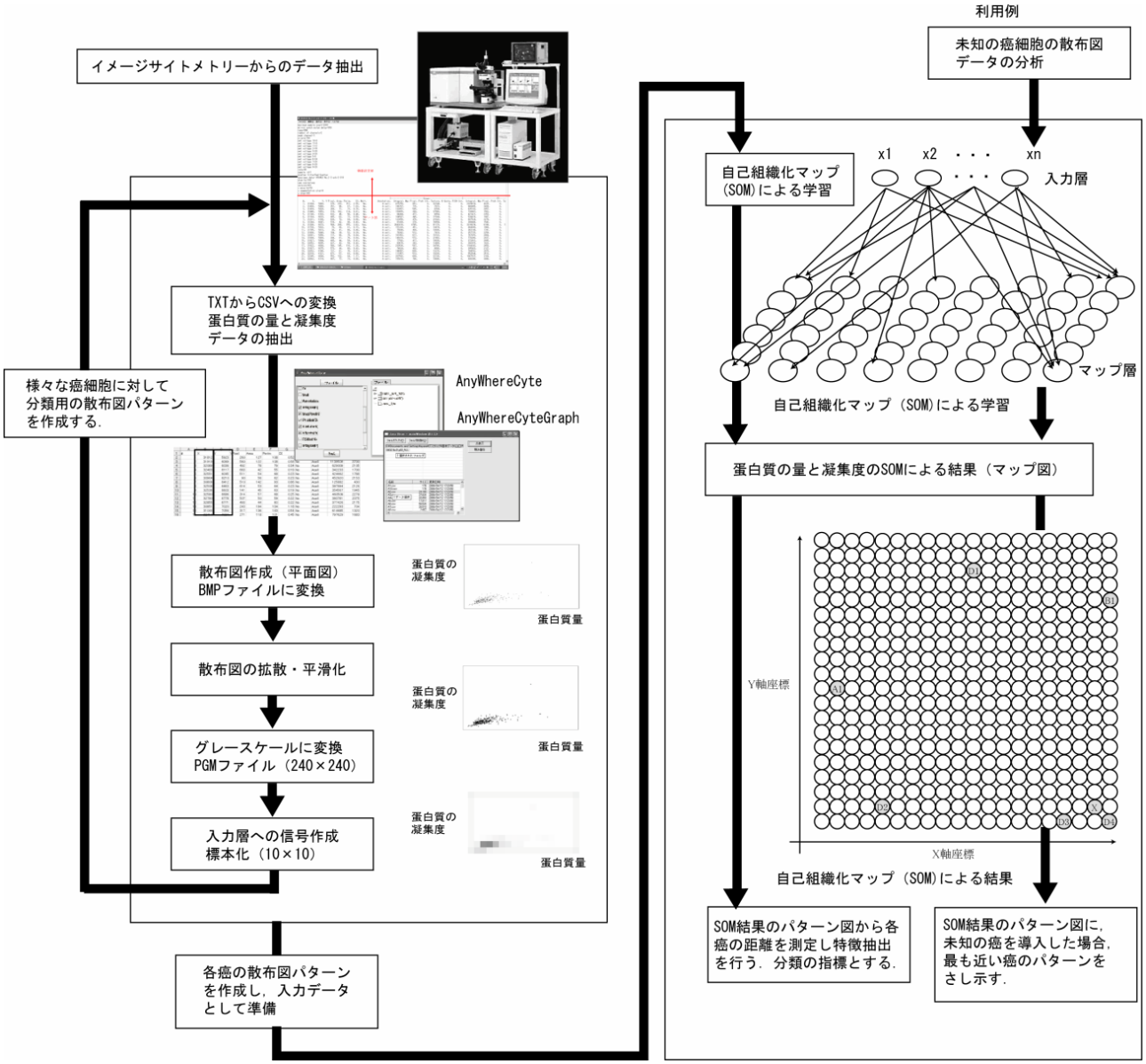

図 2.1　システム全体の流れ図

表 2.1　使用する癌細胞の名称

癌細胞名	部位癌
MKN1	胃癌　(線扁平上皮癌)
MS-1	肺癌　(小細胞癌)
KURAMOCHI	卵巣癌　(未分化癌)
SW-13	副腎癌　(線癌)
WiDr	大腸癌　(線癌)

表 2.2　使用する一次抗体の名称

一次抗体
RET
L-26
p27
6LCA

2.1.2.　LSC について

図 1.1 に LSC(laser scanning cytometer)を示す．LSC は 1991 年に，Kamentsky らによって開発されたサイトメータである．サイトメータとは，レーザを照射することによって細胞から発せられる散乱光や蛍光を検出する装置であり，散乱光として細胞の大きさ，内部構造の複雑さを知ることができる装置である．また検出された蛍光により生物学的特徴(DNA 量・細胞表面抗原など)を解析することができる．また，蛍光物質の量，面積などを測定していく際，同時に位置情報も記録する．また，LSC では蛍光量を小さなピクセル単位で測定しており，１つの細胞として認識された領域内の全ピクセルの蛍光量を合算して１つの細胞における蛍光量（Integral）とし，その中で最も高い蛍光量を示したピクセルを max pixel とする．この max pixel は，換言すれば蛍光物質の凝集度，すなわち測定する細胞内物質の凝集度を表す指標である．これらの Integral と max pixel の値のよって，これまで識別不可能であった細胞周期を簡単に識別することができる．

2.1.3 Integral と Max Pixcel について

本研究では，癌細胞の分類において利用するデータとして Integral と MaxPixel に焦点を当てて行う．Integral は蛋白質の量，MaxPixel は蛋白質の凝集度を表している．蛋白質の量とは，癌細胞内に存在する蛋白質の量のことで，この蛋白質が癌細胞内にどのように存在しているかを表すのが凝集度である．従来のサイトメーターでは，蛋白質の量である Integral しか測定できなかったが，LSC が開発されたことにより，凝集度である MaxPixel も測定することが可能となった．本研究では，蛋白質の重要な要素として Integral と MaxPixel 間の散布図に癌を分類する特徴が出ると仮定し研究を行った．

2.1.4 散布図の作成方法

独自に開発したプログラム Main window を用いて，切り出した値からデータとなる散布図を作成する．この切り出しプログラムは，プレパラートのスポットごとにデータを抽出し，散布図を作成するツールである．このツールを用いることで，容易に大量のデータから指定したデータを自動的に抽出し，その散布図を自動生成することが可能となる．特定の蛋白質について Integral と MaxPixel をパラメータとした癌細胞の散布図の例を図 2.2～図 2.4 に示す．図中の各点は癌細胞のデータを表している．

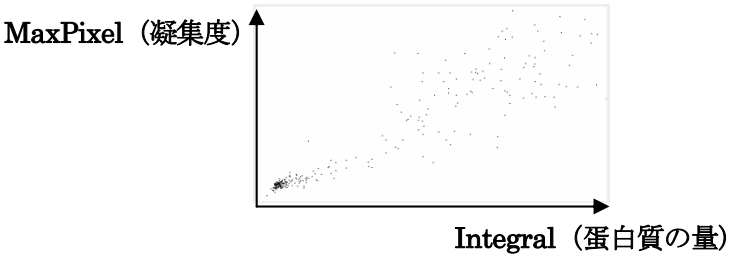

図 2.2　散布図データ(蛋白質：L-26 癌細胞：KURAMOCHI)

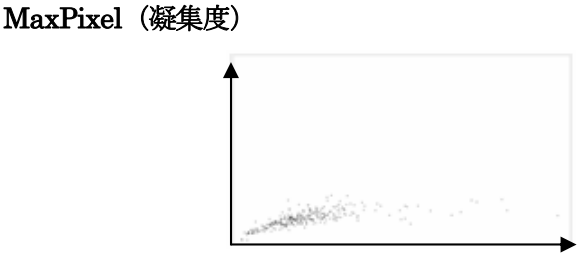

図 2.3　散布図データ(蛋白質：RET 癌細胞：KURAMOCHI)

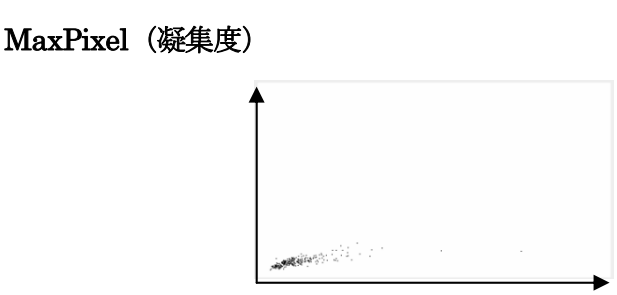

図 2.4　散布図データ（蛋白質：LCA 癌細胞：KURAMOCHI)

図 2.2～図 2.4 に示す散布図のデータを見比べると，各々の点の集まりの違い，傾斜の違い・密度の違いなどにおいて特徴が見られる．しかし，異種の蛋白質でも，その形状が似ているものや，同種の癌細胞で同種の蛋白質でも異なった形状を持つものなど様々な散布図データが得られた．

そのような多様なデータに対して，人間の目で散布図の特徴や類似性を見分けることは容易ではない．そこで，類似性に基づき，データを分類するのに有用な SOM を用いる方法を試みる．

2.2 自己組織化マップ(SOM)

2.2.1. 構造

図 2.5 に SOM ネットワークの構造を示す．ネットワークは入力層とマップ層の 2 層からなり，層内での結合はない．入力層ニューロンとマップ層ニューロンは全結合である．マップ層は出力を視覚的に見るため，通常２次元に配列されている．

時刻 t において入力層に入力ベクトル

$$x(t) = [x_1(t), \cdots, x_i(t), \cdots, x_n(t)]$$

が与えられると，マップ層のニューロンは結合荷重

$$w_j(t) = [w_{j1}(t), \cdots, w_{ji}(t), \cdots, w_{jn}(t)]$$

を介して入力層からの入力を受け，後述の学習アルゴリズムに従って学習を繰り返す．その結果，似た入力に対しては，マップ層の互いに近くのニューロンが反応するようになる．すなわちネットワークは，トポロジカルマッピングを実現する．ここで，w_{ji} は入力層 i 番目のニューロンと，マップ層 j 番目のニューロン間の荷重である．

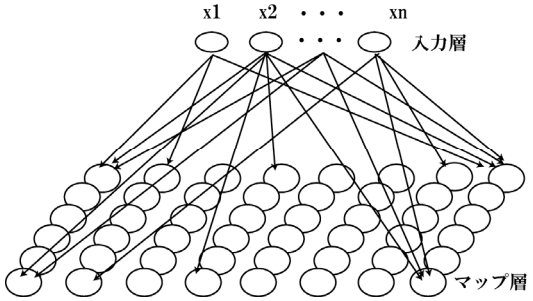

図 2.5　SOM の構造

2.2.2.　アルゴリズム

入力層とマップ層のニューロン間の結合荷重を学習することにより，トポロジカルマッピングを実現する．以下にアルゴリズムを示す．

（1）　ネットワークの初期化

入力層とマップ層間の荷重の初期値を乱数によって設定する．

（2）　入力ベクトルの入力

入力層に入力ベクトル $x=(x_1,\cdots,x_i,\cdots,x_n)$ を入力する．

（3）　入力ベクトルと荷重ベクトルの距離計算

入力ベクトルとマップ層の各ニューロンの荷重ベクトルの距離を計算する．入力ベクトルとマップ層の j 番目のニューロンの荷重ベクトルの距離 d_j は次式で与えられる．

$$d_j=\sqrt{\sum_{i=1}^{n}(x_i-w_{ji})^2} \qquad (2.2.1)$$

（4）　勝者ニューロンの決定

距離 d_j が最小となるニューロン，すなわち入力ベクトルに最も近い荷重ベクトルを持つマップ層でニューロンを選択する．このニューロンを勝者ニューロンと呼び，j^* と書くことにする．

（5）　結合荷重と各パラメータの更新

勝者ニューロンとその近傍領域内の全てのニューロンの荷重を，式(2.2.2)に基づいて更新する．近傍関数は式(2.2.3)で定義され，その領域は学習の経過と共に狭まるようにする．

$$\Delta w_{ji}=\alpha h(j,j^*)(x-w_{ji}) \quad (2.2.2)$$

$$h(j,j^*)=\exp\left((\frac{-|j-j^*|}{\sigma^2}\right) \qquad (2.2.3)$$

ここで，α は学習率係数であり，学習の経過と共に減少させる．同様に，σ も学習の経過と共に減少させる．

（6） （2）へ戻る
（2）から（5）を繰り返す.

このアルゴリズムにより，入力の類似度によるパターン分類が可能になる.

2.2.3. 画像処理

蛋白質量(Integral)と凝集度(MaxPixel)の散布図に対して SOM を用いて学習を行う．しかし単純にこのままでは, 細胞となる各点の情報が非常に小さいために, SOM の学習に時間がかかる問題点が残った．そこで，この散布図を 1 枚の画像として考え，画像処理を行う事で各点を大きなある程度塊のある情報として捉える事とする．その処理には，最初に拡散を２回，次に平滑化を１回施し，これを２回繰り返す方法を用いた．拡散処理は点を中心に８近傍を塗りつぶす方法で行った．平滑化処理は８近傍では上下左右にしかフィルタをかけることができないので，斜めも含めて広い範囲で点をぼかすことができる 25 近傍で行った．その結果を図 2.6 に示す．これからわかるように
一つ一つの点が拡大されて，ある程度の固まりとして捕らえることができる（図 2.6 右）.

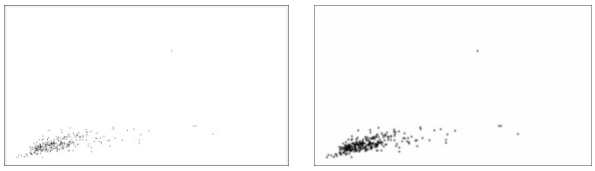

図 2.6 画像処理前（左），画像処理後（右）

2.3. 実験手順

図 2.7 に SOM を用いた癌細胞データの散布図による分類システムの構成を示す. LSC により得られた癌細胞のデータを散布図にし，SOM により学習を行う．その結果，得られた SOM マップによって分類が行えたか検討をする．マッピングに用いるデータは癌細胞 5 種，蛋白質 4 種を用意した．同種の癌細胞に対して同種の蛋白質のデータをそれぞれ 6 枚ずつ用意し，学習を行った．SOM の学習に用いる散布図は，データ量を減らすため，特徴を保持したまま低次元化する必要がある．そこで，低次元化の手法として標本化を行う．標本化では 10×10＝100 次元までダウンマッピングすることで，SOM への入力ベクトルを作成する．入力ベクトルの各要素は，標本化後の散布図の画像の 1 ブロックに対応しており，0～255 の値を持つ．マップ層は，20×20＝400 ニューロンの二次元マップとする．マップ層の各ニューロンは 100 次元の荷重ベクトルを持つ．これは標本化後の散布図の画像全体に対応している．このマップを，全ての入力ベクトルを用いて自己組織化させる．学習は，1 種類の蛋白質当り 6 枚の散布図の画像を 15 回学習し（これを 1 サイクルとする），これを 30 サイクル行う．ここで，1 種類の蛋白質当たりの学習をサブ学習と呼び, その回数をサブ学習回数と呼ぶことにする．学習における各パラメータと学習回数は表 2.4 に示す.

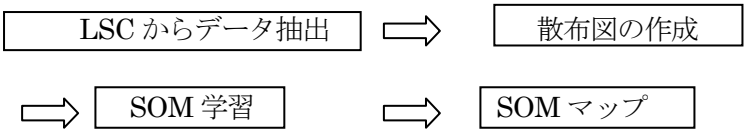

図 2.7 システム構成

表 2.4　学習におけるパラメータと学習回数

蛋白質の種類（種）	α の初期値	σ の初期値	サブ学習回数（回）	サイクル数(回)	学習用画像（枚）
4	0.01	8.0	15	30	6

3.　実験結果および考察

癌細胞は，KURAMOCHI，MS-1，MKN1，SW-13，WiDr の 5 種を用い，各癌細胞の蛋白質である L-26, LCA，RET，p27 に注目した．学習終了後の結果を図 3.1～図 3.5 に示す．１種類の癌細胞に対する４種の蛋白質のデータを各々の勝者ニューロンとして表示している．

４種類の蛋白質 L-26，LCA，RET，p27 をそれぞれ A，B，C，D で表し，1 種類の蛋白質に対する 6 枚のデータに，それぞれ 1～6 の番号を対応付けている．すなわち，図中 A1 は，蛋白質 L-26 に関する 1 番目のデータの勝者ニューロンを示している．X は 2 種類以上の蛋白質に対する勝者ニューロンを示している．マッピングした結果の全てのマップで X が見られる．これは，2 種類以上の蛋白質においてその散布図のデータが類似していることを示している．しかし，この結果より特徴的な蛋白質が存在することが分かった．例えば，図 3.1 において，A1～A4 や D1・D2 は X とは異なる場所にマッピングされている．これは，同じ種類の癌細胞のデータであっても，散布図に違いがあるということを示している．癌細胞のデータを取得する際，細胞の周期や蛋白質のデータを取得する実験にかかる時間が異なる．従って，A1～A4 が示す L-26 と D1･D2 が示す p27 は癌細胞 KURAMOCHI において時間的に変化があった蛋白質データではないかと推測される．すなわち，L-26 と p27 は，癌細胞 KURAMOCHI において他とは区別される特徴的な蛋白質ではないかと推測される．図 3.1~図 3.5 の結果より，各癌細胞において時間的に変化する特徴的なデータであると推測される蛋白質の種類を表 3.1 に示す．

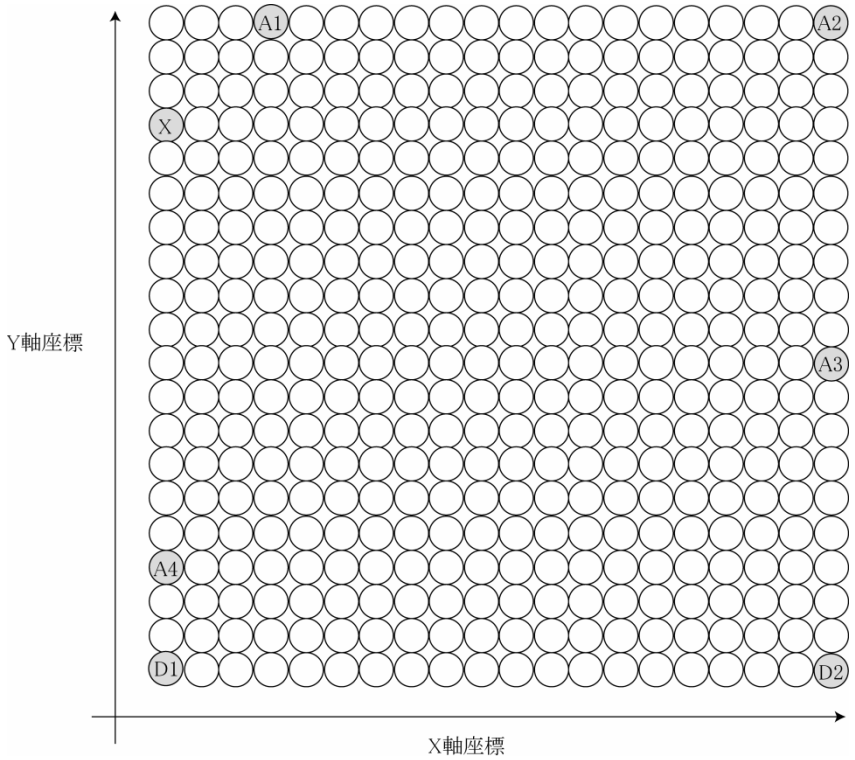

図 3.1　癌細胞 KURAMOCHI の蛋白質によるマッピング

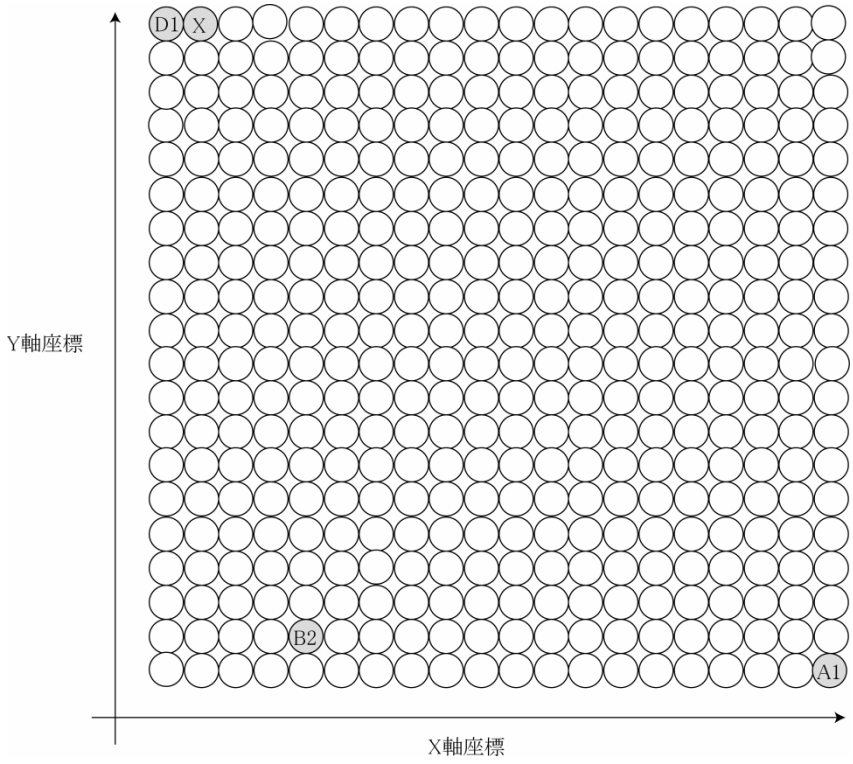

図 3.2　癌細胞 MKN 1 の蛋白質によるマッピング

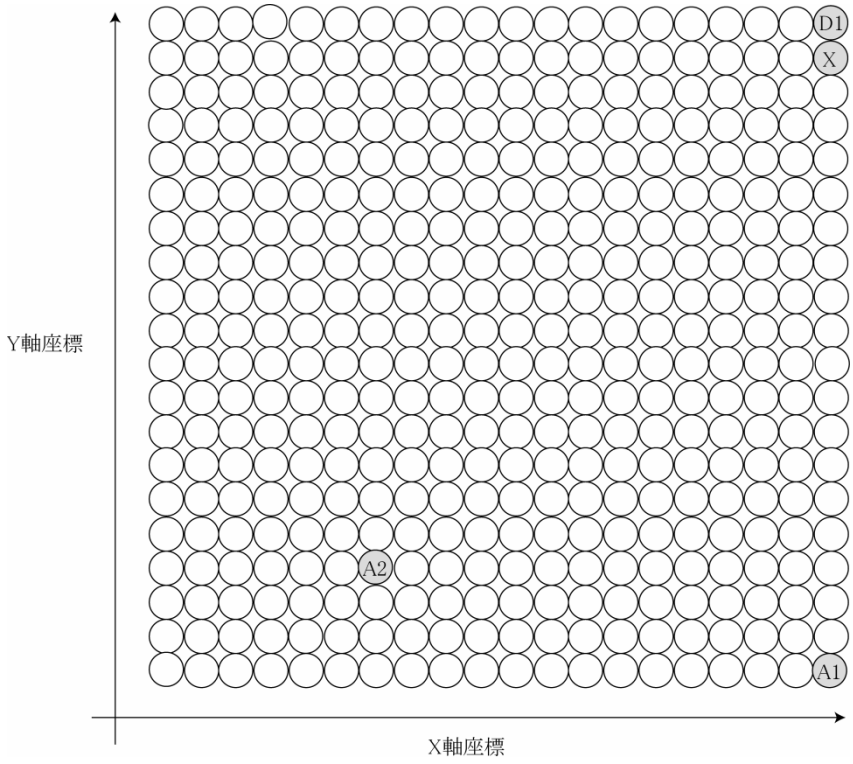

図 3.3　癌細胞 MS1 の蛋白質によるマッピング

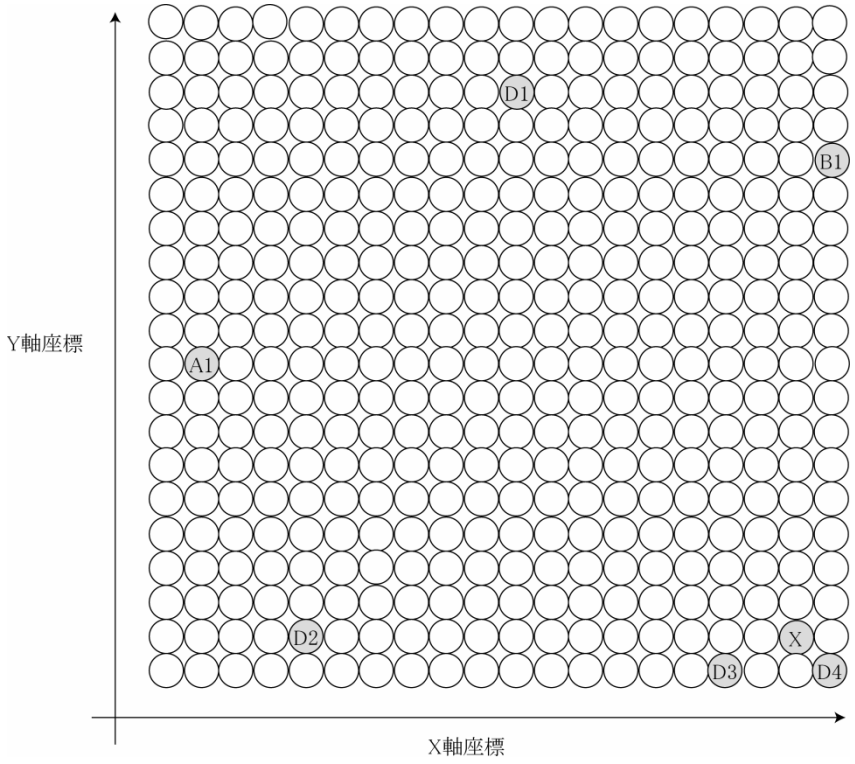

図 3.4　癌細胞 SW-13 の蛋白質によるマッピング

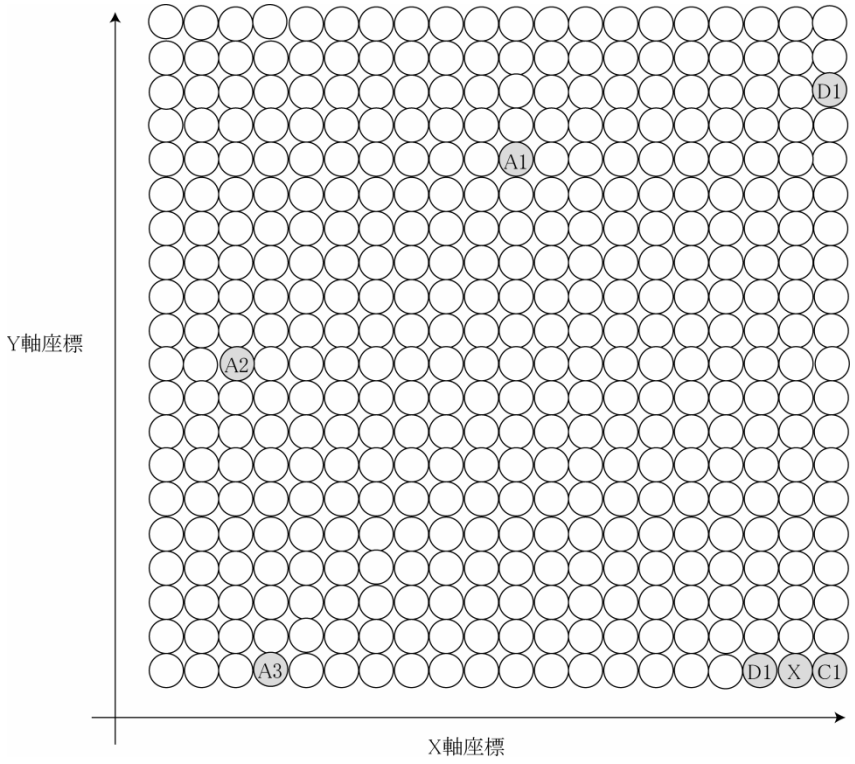

図 3.5　癌細胞 WiDr の蛋白質によるマッピング

表 3.1　癌細胞の蛋白質の特徴

癌細胞名	部位癌	時間的変化が推測される蛋白質
KURAMOCHI	卵巣癌　(未分化癌)	L-26・p27
MS-1	肺癌　(小細胞癌)	L-26
MKN1	胃癌　(線扁平上皮癌)	L-26・LCA
SW-13	副腎癌　(線癌)	L-26・LCA・p27
WiDr	大腸癌　(線癌)	L-26・p27

4.　結言

LSC により抽出された癌細胞データの分類に SOM が有効的であるか検討した．ここでは，分類を行うためのデータとして癌細胞の蛋白質に着目する．まず，LSC より抽出した癌細胞データを，扱いやすいようにテキスト形式に変換し，必要なデータのみを取得して csv 形式に変換した．その csv 形式にしたデータから蛋白質の量を示す Integral と蛋白質の凝集度を示す MaxPixel 切り出し，散布図を作成した．更に散布図のデータに標本化を施し，得られた画像データを SOM の入力として分類を試みた．その結果，癌細胞ごとに，他とは区別されるような特徴的な蛋白質が存在し，時間的変化の可能性がある蛋白質を推測できる可能性があることが分かった．これにより，同種の癌細胞を蛋白質によって分類できる可能性があることが示された．今後，癌細胞の種類を増やし，多くの蛋白質で更なる詳しい分析を行う必要がある．

謝辞

ご指導をして頂いた徳山高専の池田信彦教授，山口大学理学部の松野博嗣教授，癌細胞データを提供して頂いた山口大学医学部の古屋智子助教に深く感謝致します．

参考文献

[1]古屋智子，古賀厚徳，佐々木功典，セルアレイシステムの開発と応用，Bio ベンチャー，Vol.4，7-8，羊土社, 2004.
[2]松岡秀実，池田信彦，北風裕教，藤田重隆，松野博嗣，LSC を用いた癌細胞の蛋白質による分類の一考察，日本機械学会 中国四国学生会 第 37 回学生会員卒業研究発表講演会　講演前刷集, 9-9，2007.
[3]徳高平蔵，藤村喜久郎，山川烈，自己組織化マップ応用事例集，海文堂，2002.
[4]徳高平蔵，岸田悟，藤村喜久郎，自己組織化マップ応用，海文堂，1999.