情報学基礎 6－3
（1987. 9. 21）

生物学における同定支援システムの試作

鵜川義弘、舘野義男、菅原秀明、工藤卓二、清野昭雄（理化学研究所）
小野幹雄（都立大学）

分類と同定は、生物学の基本となる知的活動である。例えば、環境から新たに分離された細菌は、既存の分類体系に基づいて既知の細菌と比較され、既知の一員として分類されるか、新種として命名される。また、従来の分類体系が変更を受ければ、既存の細菌の学名も変更されることがある。こうした知的活動を支援する計算機プログラムは、エキスパートシステムが喧伝される以前から実用的に工夫され、利用されてきた。今回は、分類・同定の手法を述べた後、当室でこれまでに試作した、日本産樹木同定システム及び放線菌画像データベースを紹介する。

A TRIAL SUPPORT SYSTEM OF IDENTIFICATION IN BIOLOGY

Yoshihiro UGAWA, Yoshio TATENO, Hideaki SUGAWARA, Takuji KUDO, Akio SEINO
The Institute of Physical and Chemical Research (RIKEN)
2-1, Hirosawa, Wako, 351-01 Saitama
and
Mikio ONO
Makino Herbarium, Tokyo Metropolitan University
2-2-1, Fukasawa, Setagaya, 158 Tokyo

Classification and identification of living things are the basis of biology. We have developed identification systems to support biologists and researchers in other fields. The systems can transfer experiences and knowledge to novices though they are not considered to be an authentic Expert system. We introduce a theory of probabilistic identification, Identification System of Japanese Woody Plants (JUMOKU) and Image Database System of Actinomycetes (ACTINOBASE).

1．生物の分類と同定

古くから、人類は生物を観察したり利用したりしてきた。したがって日常生活において、コミュニケーションのために、種々の生物に、名前を与える必要があった。学問的には、１８世紀にリンネによって２命名法を用いて生物を分類し名前を与えることが始められた。生物の分類学においては、生物の形質の特徴を明確にし（記載）、他の生物との類縁関係を分析することで分類群を確立し（分類）、個々の分類群に対して学名を決定する（命名）。この分類体系を基に所属すべき分類群を決める（同定）。

最近の生物学の発展により分類、同定のために扱うデータの種類も変遷してきている。形態のみのデータから、生理的性質、生化学的性質、遺伝的性質（分子データ）まで幅広いデータが扱われるようになってきた。また、これらのデータの形は、文字や数値、２値（＋、－）、さらにグラフ、写真などがあり多様である。コンピュータの発達により、これらのデータが多変量解析などの諸法を用いて有効に活用できるようになり、新しく分類・同定方法の一つとして認められるようになってきた。

2．数値分類と確率的同定

１９７０年代から、コンピュータの普及とあいまって、数値分類と確率的同定の手法が、形態や生化学的性質データに基づいて、客観的かつ安定な分類体系を築き、同定を行う手段として導入されてきた。

1）数値分類

分類は、似たもの同志からなる集合を見つけ出し、しかも他の集合と区別をつける作業である。この類似性を決める指標（類似度(Similarity)）には以下のようなものがある。

Simple :	$S_{SM} = (a + d) / p$	p:全項目数
Jaccard's:	$S_j = a / (p - d)$	a:＋で一致する項目の数　R:値域
Quantitative:	$S_q = 1 - \lvert x_i - x_j \rvert / R$	d:－で一致する項目の数　x:測定値

このとき、項目の選び方には注意が必要である（項目の重み付けの問題）。特定の項目を振り分けのキーとする２分法（後述）と異なり、数値分類では各項目が同じ重みで取り扱われる。しかし、項目の選び方によって重み付けが忍び込むことがあるので注意しなければならない。例えば、生化学的に明らかな相関がある項目を多数測定した場合には、他の項目とのバランスも考えて、そのうちの１項目を選択するなどして、意図しない重み付けを避ける必要がある。

個々の菌株の類似性が判れば、類似性の高い菌株同志をまとめ分類体系を作り上げていくことができる。集合と集合の類似度を決める場合も次のようないくつかの方法がある。

菌株間の類似度の平均値:average linkage
最も似ている菌株の類似度で代用:single linkage
最も似ていない菌株の類似度で代用:complete linkage

2）確率的同定

同定の方法には２種類ある。一つは、同定のキーとして決定された形質について順次比較し、階層的な、分類体系をたどる２分法である。この場合は、先に比較される項目ほど、同定に対してより決定的な役割を果たし、また、データに欠損がある形質がキーになっていると、同定作業を進めることが難しい。もう一つは、測定した形質のデータを同時に平行的に比較する確率的同定である。確率的同定では、各項目の陽性率を菌種ごとに整理した陽性率表を基に相似度を計算する。

陽性率表

陽性率表は、測定項目と属や種の分類群の２元表であり、データが陽性（＋）になる確率を示す。陽性率は多数の菌株の、測定結果や文献中のデータを勘案して決める。

相似度(likelihood)

分離された菌Ｘのデータが表1の中段のようになっていたとしよう相似度が高い種ほど、菌Ｘである可能性が高いことと、'－'のデータに注意して、菌Ｘの各菌種に対する相似度は表2のように計算される。一方、表1の中の Citrobacter amalonaticus に属している菌株αの測定データは表1の下段のようになる可能性が高い。従って、菌株αの C. amalonaticus に対する相似度は0.83687になる（1.00000にはならない）。

相似度は同定の指標となるが、表2の相似度を基準化することによって、相対相似度と同定スコアという指標も定義できる。相対相似度は相似度を相似度の和で割ったものである。同定スコアは、菌Ｘの相似度を、それぞれの菌種に属する菌株の相似度で割ったものである。従って、菌Ｘの相対相似度と同定スコアは表3のようになる。

表3から、菌XはC. diversusに属する可能性が非常に高いと考えられる。相対相似度から、他の菌種よりも遙かにC. diversusに近く、しかも同定スコアが1であることから、菌XがC. diversusの典型的な性質を持っていることも明らかである。

表1　陽性率表と分離された菌のデータ

	H2S	IND	CIT	ARG	ORN	MAL
C. amalonaticus	0.01	0.99	0.99	0.88	0.99	0.01
C. freumdii	0.82	0.04	0.92	0.39	0.13	0.16
C. diversus	0.01	0.99	0.99	0.60	0.99	0.89
分離された菌X	−	+	+	+	+	+
C. amalonaticus に属する菌株α	−	+	+	+	+	−

表2　分離された菌Xの相似度の計算

C. amalonaticusに対して　(1-0.01)x0.99x0.99x0.88x0.99x0.01=0.00845

C. freumdiiに対して　(1-0.82)x0.04x0.92x0.39x0.13x0.16=0.00005

C. diversusに対して　(1-0.01)x0.99x0.99x0.60x0.99x0.89=0.51296

相似度の和　0.00845+0.00005+0.51296=0.52146

菌株αのC. amalonaticusに対する相似度　(1-0.01)x0.99x0.99x0.88x0.99x(1-0.01)=0.83687

表3　分離された菌Xの相対相似度と同定スコア

	相対相似度	同定スコア
C. amalonaticus	0.00845/0.52146=0.01620	0.00845/0.83687=0.01010
C. freumdii	0.00005/0.52146=0.00010	0.00005/0.32285=0.00015
C. diversus	0.51296/0.52146=0.98370	0.51296/0.51296=1.00000

3．事例1　日本産樹木同定システム(JUMOKU)

現在のわが国に自生する高等植物は、ほぼ調べつくされていて、ほとんどの日本産植物を網羅した植物図鑑も何種類か刊行されている。しかし、これらの図鑑などが種の同定を目的として作られおらず、また分類学的順序に従って並んでいたりして、一般の人々が植物の名前を知ろうとしても、簡単かつ正確に同定する方法がないのが現状であった。図鑑で用いられる二分法は、前述のように、ある項目に対するデータが判らないと検索が途中でストップする（キーとなる形質に茎、葉、花、果実などがあり、四季を通じて観察する必要がある）ほか、1ヶ所間違うだけで正しい答が出ない、検索表の引き方に一定の順序があるなど、同定作業に不慣れな場合ほど同定が難しいという欠点がある。

本システムは、日本産樹木約５００種について４３の検索項目にわたるマトリックスを基にして、入力された検索データに最も近い樹木を検索するもので、検索結果は入力データに一致しない項目数［最悪の場合４３］の少ない順に樹木名が並び替えられて出力される。従って、検索用データは、全ての項目を入力する必要はなく、観察できた項目についてのみでよい。検索用項目は学問的に吟味された分類体系に準拠し、少ない項目で絞りこめるよう、評価した項目を選択した。選択した項目の一覧を表4に示す。

表4　日本産樹木検索に用いた検索項目

1　常緑樹/落葉樹
2　高木/低木/蔓(つる状)
3　自生地：琉球/九州/四国/中国・近畿/中部・関東/東北/北海道
4　生育場所：海岸/高山帯・亜高山帯/植栽木/山地・平地・その他
5　若枝の毛：あり/なし
6　幹・枝のその他の特徴：樹皮が斑紋状にはげる/刺あり/乳汁あり/若枝・前年枝に稜・翼あり/前年枝が緑色/中空髄/階段状髄/枝・葉に芳香・臭気あり
7　葉序：互生/対生/輪生/束生(枝先、短枝に輪生状に集合する)
8　(単葉の場合に限る)：単純葉/掌状裂/羽状裂
9　葉身(複葉の場合は小葉身)の裂片数：１(単純葉)/２/３/４/５/６/７/８/９/１０/１１以上
10　三出複葉/掌状複葉/奇数羽状複葉/偶数羽状複葉/二、三回羽状複葉
11　小葉数(複葉の場合に限る)：５以下/６－１０/１１－１５/１６－２０/２１－２５/２６以上
12　葉身(複葉の場合は小葉身)の長さ：２ｃｍ以下/２－４ｃｍ/４－８ｃｍ/８－１２ｃｍ/１２－２０ｃｍ/２０－３０ｃｍ/３０ｃｍ以上
13　葉柄の長さ(単葉の場合に限る)：０．５ｃｍ以下/０．５－１ｃｍ/１－２ｃｍ/２－５ｃｍ/５－１０ｃｍ/１０ｃｍ以上
14　葉の全形：線形・披針形/楕円型・卵形/円形/三角形/ひし形/多角形(五角形以上)
15　葉の先端の形：凹形/鈍形(まるい)/鋭形(尖る)/尾状(尾のように伸びる)

16　葉の基部の形：心形(ハート形)/鈍形(まるい)/鋭形(尖る)/切形/盾型
17　葉縁：全縁/波状縁(波うつ)/単鋸縁あり/重鋸縁あり/欠刻あり
18　葉脈：羽状脈/掌状脈/三主脈/一主脈で側脈不明
19　葉質：薄い/厚い(硬い)
20　成葉の毛の分布の特徴：葉裏に密生/葉柄に毛あり/葉身は表裏とも無毛
21　成葉の毛の特徴：星状毛/腺毛/鱗毛(うろこ状の毛)/以上の特徴なし
22　葉のその他の特徴：托葉あり/葉柄・葉軸に翼あり/葉柄・葉軸に腺体あり/葉身に腺体あり/明点・腺点あり/葉裏が白味を帯びる/葉形が左右不同(単葉の場合に限る)/葉縁が上下に波打つ/葉がザラつく
23　花期：１月/２月/３月/４月/５月/６月/７月/８月/９月/１０月/１１月/１２月
24　花・花序の特徴：葉前性/芳香・臭気あり/下向きに咲く/装飾花あり
25　花序：単生(花序を作らない単一の花)/枝分かれを繰り返さず主軸が伸びる花序/枝分かれを繰り返す花序/上部が平面状になる花序/束状・球状の花序/頭状花序・つぼ状花序
26　花序の位置：頂生/腋生/葉の上に乗る/葉と対生する
27　花被の構成：花弁状のものとがく片状のものの両方がある/花弁状のもののみ/がく片状のもののみ/花弁状のものもがく片状のものもない
28　花弁状のものは互に離生する/多少とも癒合する
29　花の色：褐色/緑/白/黄/橙/赤・紅/紫・青
30　花弁状のものの数：０/１/２/３/４/５/６/７/８/９/１０/１１以上
31　がく片状のものの数(花弁状のものがない場合に限る)：１/２/３/４/５/６/７/８/９/１０/１１以上
32　放射相称花/左右相称花
33　おしべの数：１/２/３/４/５/６/７/８/９/１０/１１以上
34　おしべの特徴：葯は黄色でない/おしべは花弁状のものに癒着/花糸は互いに癒合する
35　めしべの数：１/２/３/４以上
36　めしべ１本あたりの花柱の数：０(柱頭無柄)/１/２/３/４/５/６/７以上
37　柱頭の裂片数：１/２/３/４/５/６/７以上
38　子房上位/子房下位
39　液果/裂開しない乾果/裂開する乾果
40　実の特徴：翼あり/殻斗あり/核あり/トゲ・イガあり/サヤあり/球果状(マツカサ状)/ナシ状/イチヂク状/クワ・キイチゴ状/ミカン状/以上の特徴なし
41　熟果の色：褐色/緑/白/黄/橙・赤・紅/紫/黒/青
42　種子数：１/２/３/４/５/６/７以上
43　種子の特徴：長毛あり/翼あり/以上の特徴なし

このシステムは、パソコンでの普及を考え、検索ソフトを含めた全体のデータ量を７００ＫＢ－１ＭＢに抑え、フロッピーディスクに収められるようにした。また、検索に使用できるメモリー容量も１２８ＫＢ以下と少ないためビット列による検索用データの圧縮を行い、全体で３０ＫＢ以下にした。例えば、項目番号１から４のハマボウと、フヨウの検索用データは次のようになっている。

項目番号	1		2			3							4			
選択肢	落葉	常緑	高木	低木	草状	琉球	九州	四国	中国近畿	中部関東	東北	北海道	海岸	高山帯	植栽木	山地平地
ハマボウ	0	1	0	1	0	1	1	1	1	1	0	0	1	0	1	0
フヨウ	0	1	0	1	0	1	1	0	0	0	0	0	0	0	1	1

これらは、１６進数表記でそれぞれ、ハマボウ　５７ＣＡ(H)、フヨウ　５６０３(H)と書ける。

パソコンは演算速度も遅いので検索結果のソートにはクイックソートのアルゴリズムを用いた。さらに、ユーザーインターフェイスを考え、検索用項目のうち難解なものには挿絵を表示し、データ入力を助けるようした。観測データ入力と検索結果表示の画面の例を図1に示す。

４．事例２　放線菌の画像データベース(ACTINOBASE)

放線菌は、古くから、抗生物質やビタミンなどの生理活性物質を生産することで知られており、今後も遺伝子工学を応用して利用の拡大や、研究の発展が望まれている微生物である。しかしながら放線菌の同定は、対象とする種の数が多く、個々の形態に関する情報が重要で比較対照が難しいために、専門家に頼らざるをえないのが現状である。

そこで、放線菌を形態の特徴のみならず、他の生理、生化学的及び化学分類学的特徴との組み合せで検索し、検索結果の画像を瞬時に出力するシステムを作ることを計画した。本システムは、基本となる放線菌（約１２００株）の走査型電子顕微鏡写真（約９０００枚）の画像を追記型レーザーディスクに記録しパソコンと連動させ同定の支援システムを実現しようというものである。レーザーディスクには１枚当り静止画が２４０００枚記録でき、パソコンとレーザーディスク再生装置を合わせて１２０万円程度でシステム構成ができる予定である。

現在のところ、画像と菌株番号を関連づけるデータベース構造が完成し、学名または菌株番号を入力すれば対応する画像を瞬時に出力することができる。これからの作業としては検索・同定に用いる菌株の分類学的特徴

データの抽出、データ入力、検索プログラムの開発などを予定している。

検索プログラムに関しては単純確率法を基本にし、従来の専門家による同定手順が織り込めるものとしたい。例えば、形態、生理試験、色などは、データ項目や、項目内選択肢の間で種を同定する場合の重要度が一様でなく、検索項目の重み付けがシステムの出来不出来を決定する重要な鍵となると思われる。予定されているデータ項目の一覧を表5に示す。

表5　放線菌画像データベース検索項目例

A. Japan Collection of Microorganisms (JCM) Catalogue data
  1) JCM number
  2) Scientific name
  3) Type strain or not
  4) Mol% G+C of DNA
  5) Menaquinone
  6) Cell wall composition (containing whole-cell sugar pattern and cell wall acyl type)
  9) Antibiotic production

B. International Streptomyces Project (ISP) description:
  1) Spore chain morphology: 1.Rectiflexibiles 2.Retinaculiaperti 3.Spirales 4.Verticillati 5.Others 6.Comments
  2) Number of spore: 1.less than 10 2.10-50 3.more than 50 4.Others 5.Comments
  3) Spore surface: 1.Smooth 2.Warty 3.Spiny 4.Hairy 5.Knobby 6.Rugose 7.Others 8.Comments
  4) Special morphology: 1.Hygroscopic 2.Sporangiumlike (subglobose bodies composed of masses of spores, moist masses of spores) 3.Coremea 4.Fragmentation of substrate mycelium 5.Substrate conidium 6.Sclerotium 7.Pycnidium 8.Others 9.Comments
  5) Aerial mass color: 1.White 2.Gray 3.Blue 4.Green 5.Red 6.Yellow 7.Violet (or Purple) 8.Others 9.Comments
  6) Reverse side of colony: 1.No distinctive pigment (Yellow or Brown) 2.Red 3.Gray 4.Blue 5.Violet (or Purple) 6.Orange 7.White 8.Green 9.Others 10.Comments
  7) Melanoid pigment (+, -, d で入力) : 1.on ISP no.6 2.on ISP no.7 3.on ISP no.1
  8) Other soluble pigment: 1.Present 2.No pigment 3.Red 4.Green 5.Yellow 6.Orange 7.Blue 8.Violet (or Purple) 9.Others 10.Comments 11.Change by HCl or NaCl 12.No change by HCl or NaCl
  9) Utilization of carbohydrates: 1.D-Glucose 2.L-Arabinose 3.Sucrose 4.D-Xylose 5.i-Inositol 6.D-Mannitol 7.D-Fructose 8.Rhamnose 9.Raffinose 10.Cellulose (11.Salicin) 12.Others 13.Comments

C. Bergey's manual systematic bacteriology

C.1 Generic criteria
  1) Marked fragmentation of mycelium: 1.+ 2.- 3.d
  2) Aerial mycelium production: 1.+ 2 - 3.d
  3) Conidia formation: 1.+ 2 - 3.d 4.Short chain((20) 5.Long chain()20)
  4) Motile elements production: 1.+ 2 - 3.d
  5) Facultative anaerobe: 1.+ 2 - 3.d
  6) Cell wall type: 1.I 2.II 3.III 4.IV 5.V 6.VI 7.VII 8.VIII 9.IX 10.X
  7) Whole-cell sugar pattern: 1.A 2.B 3.C 4.D
  8) Mycolic acid presence: 1.+ 2.- 3.total carbon number (幅を持たせた数字で入力)
  9) Phospholipid type: 1.PI 2.PII 3.PIII 4.PIV 5.Phosphatidylethanolamine
  10) Menaquinone A と B の組み合せが複数あることが多い。(3-4とおりぐらいまで)
    A.Number of isoprene units: 1.MK-7 2.MK-8 3.MK-9 4.MK-10 5.MK-11 6.MK-12 7.Others
    B.Number of Hydrogen atoms saturating the side chain: 1.OH 2.2H 3.4H 4.6H 5.8H 6.10H 7.Others
  11) Mol% G+C of DNA (幅を持たせた数字で入力) :
  12) Special morphology: 1.Ovoid vesicle 2.Growth by budding 3.Others
  13) Fatty acid: 1.Unsaturated 2.Tuberculostearic (10 methyl-19) 3.Iso and anteiso

C.2 Specific criteria (+, -, d で入力)
  (例) Rhodococcus
  1) Morphogenetic sequence: 1.Elementary branching-rod-coccus growth cycle (EB-R-C) 2.Rod-coccus growth cycle (R-C) 3.Hypha-rod-coccus growth cycle (H-R-C)
  2) Decomposition of: 1.Adenine 2.Tyrosine 3.Urea
  3) Growth on sole carbon sources: 1.Ethanol (1.0%) 2.Glycerol (1.0%) 3.Inositol (1.0%) 4.Maltose (1.0%) 5.Mannitol (1.0%) 6.Rhamnose (1.0%) 7.Sorbitol (1.0%) 8.Sucrose (1.0%) 9.p-Cresol (0.1%) 10.m-Hydroxybenzoic acid (0.1%) 11.p-Hydroxybenzoic acid (0.1%) 12.Pimelic acid (0.1%) 13.Sebacic acid (0.1%) 14.Sodium adipate (0.1%) 15.Sodium benzoate (0.1%) 16.Sodium citrate (0.1%) 17.Sodium fumalate (0.1%) 18.Sodium gluconate (0.1%) 19.Sodium lactate (0.1%) 20.Sodium malate (0.1%) 21.Sodium pyruvate (0.1%) 22.Sodium succinate (0.1%) 23.Testosterone (0.1%) 24.L-Tyrosine (0.1%)
  4) Growth on sole carbon and nitrogen source: 1.Acetamide 2.Serine 3.Trimethylenediamine
  5) Growth at: 1.10℃ 2.40℃ 3.45℃
  6) Growth in the presence of: 1.Crystal violet (0.001%) 2.Crystal violet (0.0001%) 3.Phenol

(0.1%) 4.Phenyl ethanol (0.3%, v/v) 5.Sodium azide (0.01%) 6.Sodium azide (0.02%) 7.Sodium chloride (5.0%) 8.Sodium chloride (7.0%)

7) Mycolic acids (number of carbon) (幅を持たせた数字で入力)
8) Menaquinone A と B の組み合せが複数あることが多い。(3-4とおりぐらいまで)
A.Number of isoprene units: 1.MK-7 2.MK-8 3.MK-9 4.MK-10 5.MK- 11 6.MK-12 7.Others
B.Number of Hydrogen atoms saturating the side chain: 1.0H 2.2H 3.4H 4.6H 5.8H 6.10H 7.Others
9) Mol% G+C of DNA (幅を持たせた数字で入力)

次に、分類のキーとされている形態上の特徴を紹介する。例として検索項目の 表5 B. 3) Spore surfase の 1.Smooth 2.Warty 3.Spiny 4.Hairy を 図2 に示す。

将来は、走査型電子顕微鏡から画像を直接レーザーディスクやパソコンに取り込み、形態の特徴のパターン認識により同定の自動化を実現したい。

5．これからの生物データベース

現在、分類・同定に用いられるデータは、２値、数値、文字から画像へと急速にその裾野を広げている。また、ＮＭＲのスペクトラム、ガスクロマトグラムなどのパターン認識による同定など、専門家による作業は複雑を極めている。これらのパターン解析をコンピュータ化することにより同定をより早く確実に行うことが可能になる。今急速に発展してきている遺伝子の塩基配列による同定は、もとより人間による作業は困難であり、計算機による同定支援システムはますます必要になると考えられる。また、検索時間を飛躍的に短縮することでセルソータなど高速判定にも用いることができるデータベースも計算機が担うべき重要な課題である。

生物はそれ自体、情報の塊である。遺伝子情報をはじめとして、生命システムが持つ他種多様な情報を生かすことのできるデータベースの開発が待たれている。

図1 日本産樹木検索－検索項目番号１７の葉縁の挿絵と検索結果表示画面

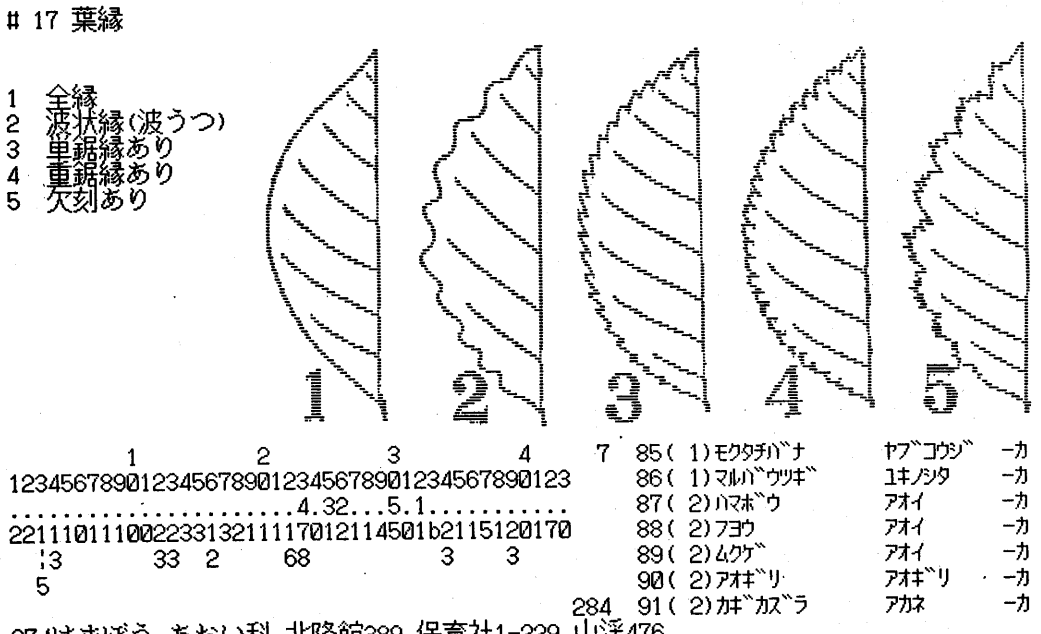

図2　放線菌画像データベース　表5 B. 3) Spore surface の　1.Smooth 2.Warty 3.Spiny 4.Hairyの例
（左上Smooth、　右上Warty、　左下Spiny、　右下Hairy）