

The Construction by Computer of a Diagnostic Key to the Genera of Yeasts and Other Such Groups of Taxa

By R. W. PAYNE,^{1*} D. YARROW² AND J. A. BARNETT³

¹ Rothamsted Experimental Station, Harpenden, Hertfordshire AL5 2JQ, U.K.

² Centraalbureau voor Schimmelcultures, Yeast Division, Laboratory for Microbiology, Julianalaan 67A, 2628 BC Delft, The Netherlands

³ School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, U.K.

(Received 3 February 1982)

Groups of taxa such as genera, or groups derived from some forms of cluster analysis, may have insufficient test results that are constant within the groups to allow diagnostic keys and tables to be constructed in the usual way. This paper describes how the usual methods can be adapted to allow construction based on information about the individual group members, instead of on the overall group information. A new key to the genera of yeasts is constructed by these modified methods.

INTRODUCTION

This paper reports new methods for computing diagnostic keys and tables to identify the group of taxa to which a specimen belongs, rather than the taxon itself. These methods are of general interest for microbial taxonomy and are illustrated below by the construction of a key to the genera of yeasts.

Barnett & Pankhurst (1974) and Barnett *et al.* (1979) devised new keys to yeast species. These keys were constructed by computer programs which required, for data, a species \times test table, each entry indicating the results that can occur when a test is used with a particular species. An entry was coded (i) as positive, if all yeasts of that species always give a positive response to the test, (ii) as negative, if all yeasts of the species always give a negative response, or (iii) as variable (or equivocal), if the response might differ between strains of a species, or be particularly dependent on the precise method of testing, so that sometimes the result might appear positive and sometimes negative.

It might seem that a key to genera could be constructed similarly, by forming a genus \times test table of results, and then using the same methods as those for the keys to the species. However, in such a table of results for the genera of yeasts, 45% of the entries were variable, and 133 pairs of genera could not be distinguished by this overall summary. The results are unsuitable for distinguishing between the genera, although suitable for the species, for two main reasons. First, most of these results are from nutritional tests, whilst the genera are mostly not classified with respect to their nutritional characteristics. Secondly, many characteristics important for classifying species into genera are inconvenient to use for routine identification and, hence, were not used in the battery of features on which the key to the species was based (Barnett *et al.*, 1979). For example, *Debaryomyces* is distinguished from *Pichia* by the former producing ascospores with warty walls and the latter producing ascospores with smooth walls. However, some species of *Pichia* have ascospores with warty walls; but, unlike those of *Debaryomyces*, electron microscopy has revealed that these warts are formed exclusively by the outer layer of the ascospore wall (Kreger-van Rij, 1970*a, b*).

Thus, the difficulties in constructing keys to yeast genera arise because of the need to use different characteristics for identifying from those used for classifying the genera. However,

Table 1. Results for taxa A to I with tests 1 to 6

Taxon	Group	Tests						
		1	2	3	4	5	6	
A	I	-	+	-	-	-	+	} Results for each taxon
B	I	+	-	-	-	v	+	
C	I	+	+	-	+	-	+	
D	II	-	+	-	+	+	+	
E	II	+	+	+	+	+	v	
F	II	-	v	+	+	+	+	
G	III	-	-	-	+	-	+	
H	III	-	-	-	v	+	-	
I	III	-	-	+	+	-	-	
	I	v	v	-	v	v	+	} Summarized results for each group of taxa
	II	v	v	v	+	+	v	
	III	-	-	v	v	v	v	

problems can occur even when the same characteristics are used for both. For example, many numerical methods of classification obtain a measure of the similarity of each pair of taxa and then form groups, or clusters, by merging similar taxa. Because the similarity measure is a single figure, based on all the characteristics of the pair of taxa concerned, there is no guarantee that any of the characteristics will be other than variable for the groups formed.

Maximal predictive classification (Gower, 1973, 1974; Barnett *et al.*, 1975) is not based on pairwise similarity, but aims to construct groups such that knowing the group to which a taxon belongs enables the maximum number of correct predictions to be made about that taxon. However, even maximal predictive groups may sometimes have few non-variable characteristics. As an example, Table 1 shows a set of hypothetical characteristics for taxa A to I and binary tests 1 to 6; the second column of the table contains the group number of each taxon for a maximal predictive classification into three groups; the group characteristics are summarized in the last three lines of the table.

This paper shows how these two difficulties, namely, (i) the disassociation of identifying and classifying and (ii) the presence of many variable results, may be overcome.

METHODS

Irredundant test sets. To identify yeasts it is usually impracticable to use tests sequentially, that is, to do each test only after interpreting the results of previous tests. Accordingly, Barnett *et al.* (1979) assumed that all the tests required for a given key would be done simultaneously. As the full set of tests in the key will be done for any identification, that set should be minimal. Hence it should contain no *redundant* tests: for example, test 5 in Table 1 is redundant, as it can be omitted without making any taxon unidentifiable. However, the set of tests 12346 is termed *irredundant* (with regard to the identification of individual taxa), since if any further tests are omitted there will be at least one pair of taxa that can no longer be distinguished. For example, if test 1 is also omitted, taxa C and D cannot be distinguished.

Barnett *et al.* (1979) used a method (reviewed by Payne & Preece, 1980) that determined all irredundant test sets available to identify the yeast species in a particular key. Appendix 1 describes how to adapt this method to form sets to identify groups instead of individual taxa, and shows that there are five irredundant sets of tests to identify the groups in Table 1, namely, 12346, 1456, 2456, 1235 and 1256. Each set can distinguish between all pairs of taxa belonging to different groups, but not necessarily between pairs of taxa in the same group. For example, there is no test in set 1456 to distinguish taxon D from taxon F, both of which are in group II.

Diagnostic keys and tables. To enable the specimens to be identified, given their results for the tests in the chosen irredundant set, a diagnostic table may be printed. Table 2 shows a diagnostic table for identifying the groups of taxa in Table 1, based on the second irredundant set, tests 1456.

An alternative means of identification is the diagnostic (or identification) key. This is most commonly used in situations when tests are done sequentially. However, as described above, it is equally possible to use a key with

Table 2. Diagnostic table for the groups of taxa in Table 1, based on tests 1, 4, 5 and 6

Characters				Group (Taxon)
1	4	5	6	
+	+	+	+/-	II (E)
+	+	-	+	I (C)
+	-	+/-	+	I (B)
-	+	+	+	II (D or F)
-	+	+	-	III (H)
-	+	-	+/-	III (G or I)
-	-	+	-	III (H)
-	-	-	+	I (A)

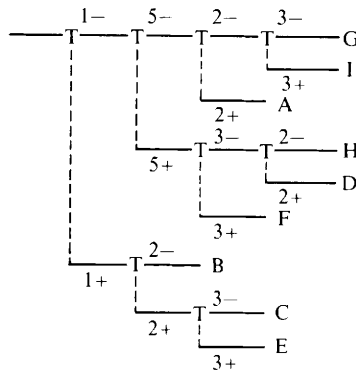


Fig. 1. Key to the individual taxa A to I in Table 1, based on tests (T) 1, 2, 3 and 5.

non-sequential testing and, if many test results are variable, such a key will be more compact than the corresponding diagnostic table (Payne & Preece, 1980).

An example, for identifying the individual taxa in Table 1, is shown diagrammatically in Fig. 1. In this key, test 1 is first, followed by test 5 for a negative result, test 2 for a positive result and so on. Thus, specimens of taxon B would be identified by test 1 +, test 2 -. Such a sequence of test results is termed a *branch* of the key.

Several computer programs have been devised for constructing keys to taxa (e.g. Pankhurst, 1970; Morse, 1971; Dallwitz, 1974; Payne, 1975). These all operate similarly, selecting first the test that best divides the taxa into subsets. For a binary test, there are two subsets: the first containing the taxa not eliminated if a negative result is observed with the test (i.e. the taxa with either negative or variable results), and the second subset containing the taxa not eliminated by a positive result. The programs then select the best test to use with each subset, continuing until the subsets each contain only one taxon. The best test is usually taken as that with minimum value of some *selection criterion function*, a function involving the following: m_i , the number of possible results to test i ; p_{ik} , the proportion of taxa in the current subset that always give result k to test i ; and r_i , the proportion of taxa in the current subset with variable results to test i . The proportions are often weighted by the frequencies with which the taxa are expected to occur.

Keys to identify groups of taxa. It is usually wasteful to use a key like that in Fig. 1 and identify the group by first identifying the individual taxon. For example, after test 1 -, test 5 - and test 2 - in Fig. 1, the taxa not eliminated, G and I, both belong to group III; so test 3 is unnecessary. This suggests one modification to the usual method of key construction, namely, that each branch should now terminate when the taxa are all from the same group.

Selection criterion functions, designed to select tests to identify individual taxa, are usually unsuitable for selecting tests to identify groups of taxa. They require modification to take account of the fact that there is no need to separate taxa belonging to the same group. Appendix 2 explains how this can be done and derives two new functions, G_v and G_c .

Table 3. *Irredundant test sets to identify yeast genera*

These are sets of tests that separate the genera as far as possible, and are minimal in the sense that, if any test is deleted from the set, there will be some pair of species in different genera that can no longer be distinguished. The tests are taken from those listed by Barnett *et al.* (1979).

Six sets are presented below, formed by taking all 43 tests in group A and one test from each of groups B and C.

<i>Group A</i>		<i>Group B</i>
1 D-Glucose fermentation	39 D- δ -Gluconolactone growth	23 Cellobiose growth
3 Maltose fermentation	40 2-Ketogluconate growth	24 Salicin growth
6 Trehalose fermentation	44 Citrate growth	25 Arbutin growth
12 D-Galactose growth	47 Ethylamine growth	
13 L-Sorbose growth	49 Nitrate growth	
14 D-Ribose growth	54 Growth without biotin	
18 L-Rhamnose growth	55 Growth without thiamin	
19 Sucrose growth	56 Growth without pyridoxine	
20 Maltose growth	57 Growth without niacin	
21 Trehalose growth	60 Growth in 50% D-glucose	<i>Group C</i>
26 Melibiose growth	61 Growth in 60% D-glucose	42 DL-Lactate growth
27 Lactose growth	63 Growth with 0.01% cycloheximide	43 Succinate growth
28 Raffinose growth	65 Starch formation	
29 Melezitose growth	66 Pink colonies	
30 Inulin growth	67 Budding cells	
31 Starch growth	68 Splitting cells	
32 Glycerol growth	69 Apical budding	
33 Erythritol growth	72 Filamentous	
35 Galactitol growth	74 Septate hyphae	
36 D-Mannitol growth	75 Ascosporeogenous	
37 D-Glucitol growth	76 Ascospores spherical, oval or reniform	
38 <i>myo</i> -Inositol growth		

RESULTS AND DISCUSSION

The methods described above have been incorporated into the computer program Genkey (Payne, 1975, 1978) and applied to the data of Barnett *et al.* (1979). Table 3 lists the six smallest irredundant test sets for identifying the genera of yeasts, each set containing 45 tests. There were 716 sets in all to choose from, containing up to 52 tests.

A key to the genera based on the irredundant set containing all 43 tests in group A of Table 3, together with tests 23 and 42, was constructed using selection criterion G_e and is printed in Table 4 is one of the compact forms of Payne *et al.* (1974). This key includes the results of physiological tests, some morphological tests and tests for ascospores. If the results of testing for ascospores are omitted there are six irredundant sets, each containing 54 tests, and a key based on one of these sets, constructed similarly to that in Table 4, took 581 lines (compared to the 307 lines in Table 4). Also, the number of pairs of species in different genera that cannot be distinguished rises from 32 to 140.

The key in Table 4 thus provides a compromise between including (i) solely physiological characteristics, as in the key of Barnett & Pankhurst (1974) and some of those of Barnett *et al.* (1979), and (ii) the varied features used by Lodder (1970) in her key to the genera. Lodder (1970) included, amongst others, the formation of ballistospores and their shape, and the presence of clamp connections, as well as some more complex features that are difficult to observe, excessively subjective or awkward to interpret. These are exemplified by the following: (i) 'cells often "ogival"; strong acetic acid production from glucose; characteristic aroma; cells on malt agar short-lived'; (ii) 'ascospores hat- or helmet-shaped, or apparently globose with an indistinct ledge, not conjugating in pairs'.

In order to diagnose a yeast's genus, Lodder's key to the genera uses about the same number of steps as that in Table 4. Although her key involves only about 12 kinds of examination in the laboratory, the scrutiny of asci, ascospores and filaments is in much

Table 4. Key to the genera of yeasts

The genera are those accepted by Barnett *et al.* (1979).

	Negative	Positive
1	Ascosporogenous	2 140
2	Pink colonies	3 111
3	Filamentous	4 44
4	myo-Inositol growth	5 34
5	Budding cells	Sterigmatomyces spp. 6
6	Maltose growth	7 20
7	Lactose growth	Sporobolomyces sp. 8
8	Apical budding	9 19
9	L-Sorbose growth	10 17
10	D-Glucose fermentation	Torulopsis spp. 11
11	Nitrate growth	12 Torulopsis spp.
12	Cellobiose growth	13 Torulopsis spp.
13	Growth without biotin	14 Torulopsis spp.
14	Sucrose growth	15 16
15	D-Mannitol growth	Torulopsis spp. Candida sp.
16	D-Mannitol growth	Candida sp. Torulopsis sp.
17	D,L-Lactate growth	Torulopsis spp. 18
18	D-Glucose fermentation	Trigonopsis sp. Torulopsis spp.
19	Cellobiose growth	Schizoblastosporion sp. Torulopsis sp.
20	Starch formation	21 33
21	Erythritol growth	22 30
22	Lactose growth	23 28
23	Cellobiose growth	Torulopsis spp. 24
24	Melezitose growth	Torulopsis spp. 25
25	Melibiose growth	26 Torulopsis spp.
26	Starch growth	27 Torulopsis spp.
27	D-Mannitol growth	Torulopsis sp. Candida sp.
28	Raffinose growth	29 Torulopsis spp.
29	D-Glucose fermentation	Bullera sp. Torulopsis sp.
30	Nitrate growth	31 32
31	D,L-Lactate growth	Candida sp. Selenozyma sp.
32	D-Galactose growth	Torulopsis spp. Rhodotorula sp.
33	Raffinose growth	Aessosporon sp. Cryptococcus sp.
34	Budding cells	Sterigmatomyces spp. 35
35	Maltose growth	Cryptococcus spp. 36
36	Melezitose growth	37 38
37	Glycerol growth	Cryptococcus spp. Filobasidium sp.
38	Nitrate growth	39 41
39	Lactose growth	Cryptococcus spp. 40
40	Raffinose growth	Cryptococcus spp. Bullera sp.
41	Starch formation	Bullera sp. 42
42	Sucrose growth	Cryptococcus sp. 43
43	Melibiose growth	Cryptococcus sp. Cryptococcus spp.
44	Splitting cells	45 106
45	Apical budding	46 102
46	myo-Inositol growth	47 87
47	Budding cells	Sterigmatomyces spp. 48
48	Erythritol growth	49 81
49	D-Mannitol growth	50 60
50	Maltose fermentation	51 Brettanomyces spp.
51	Nitrate growth	52 59
52	Trehalose growth	53 56

Table 4—cont.

	Negative	Positive
53 Growth without pyridoxine	54	Candida spp.
54 D-Glucose fermentation	55	Candida spp.
55 Cellobiose growth	Torulopsis sp.	Candida sp.
56 L-Sorbose growth	57	58
57 Cellobiose growth	Brettanomyces sp.	Candida sp.
58 D-Galactose growth	Candida spp.	Torulopsis sp.
59 D-Galactose growth	Candida spp.	Brettanomyces spp.
60 Lactose growth	61	80
61 Nitrate growth	62	76
62 Cellobiose growth	Candida spp.	63
63 Sucrose growth	64	67
64 Galactitol growth	65	Torulopsis sp.
65 Citrate growth	66	Candida spp.
66 Maltose fermentation	Brettanomyces sp.	Candida sp.
67 Septate hyphae	68	75
68 D,L-Lactate growth	69	Candida spp.
69 Melibiose growth	70	74
70 Starch growth	Candida spp.	71
71 D-Galactose growth	Candida sp.	72
72 Maltose fermentation	73	Candida spp.
73 Raffinose growth	Torulopsis sp.	Candida sp.
74 L-Sorbose growth	Torulopsis sp.	Candida sp.
75 Maltose fermentation	Aciculoconidium sp.	Candida spp.
76 Trehalose growth	77	78
77 Raffinose growth	Torulopsis spp.	Candida sp.
78 Cellobiose growth	79	Candida spp.
79 L-Rhamnose growth	Candida sp.	Torulopsis sp.
80 Sucrose growth	Sporobolomyces sp.	Candida spp.
81 Septate hyphae	82	85
82 Sucrose growth	83	Candida spp.
83 Citrate growth	Candida spp.	84
84 Growth without biotin	Candida sp.	Torulopsis sp.
85 Maltose fermentation	Candida spp.	86
86 L-Rhamnose growth	Trichosporon spp.	Candida sp.
87 Budding cells	Sterigmatomyces sp.	88
88 Nitrate growth	89	97
89 Melezitose growth	90	92
90 Erythritol growth	91	Candida spp.
91 Growth without biotin	Candida spp.	Filobasidium sp.
92 D-Glucose fermentation	93	Candida spp.
93 Melibiose growth	94	95
94 Lactose growth	Candida sp.	Candida spp.
	Cryptococcus sp.	
95 Lactose growth	Candida spp.	96
96 Erythritol growth	Candida sp.	Candida sp.
		Cryptococcus sp.
		Filobasidium sp.
97 Starch formation	98	
98 Erythritol growth	99	100
99 Starch growth	Candida spp.	Bullera sp.
100 L-Rhamnose growth	Candida spp.	101
101 Septate hyphae	Sterigmatomyces sp.	Candida sp.
102 Nitrate growth	103	105
103 Cellobiose growth	Schizoblastosporion sp.	104
104 Erythritol growth	Torulopsis sp.	Symptodiomyces sp.
105 Septate hyphae	Candida sp.	Leucosporidium spp.
106 Budding cells	Geotrichum spp.	107
107 Nitrate growth	108	110
108 Lactose growth	Trichosporon spp.	109

Table 4—cont.

	Negative	Positive
109 Erythritol growth	Trichosporon spp.	Sarcinosporon sp. Trichosporon sp.
110 D-Galactose growth	Sporobolomyces sp.	Trichosporon spp.
111 myo-Inositol growth	112	138
112 Starch formation	113	137
113 Nitrate growth	114	123
114 Trehalose growth	115	116
115 L-Rhamnose growth	Torulopsis sp.	Rhodotorula sp.
116 Galactitol growth	117	Rhodotorula spp.
117 Starch growth	118	Sporobolomyces sp.
118 Filamentous	119	122
119 2-Ketogluconate growth	120	Rhodotorula spp.
120 Maltose growth	121	Rhodotorula sp. Sporobolomyces sp.
121 D-Galactose growth	Rhodotorula sp. Sporobolomyces sp.	Rhodotorula spp.
122 Raffinose growth	Rhodospiridium sp.	Rhodotorula sp.
123 Erythritol growth	124	Rhodotorula sp.
124 2-Ketogluconate growth	125	134
125 Sucrose growth	126	127
126 Filamentous	Rhodotorula sp.	Sporobolomyces spp.
127 Melezitose growth	128	132
128 Growth without thiamin	Rhodospiridium spp.	129
129 Galactitol growth	130	Sporidibolus sp.
130 Maltose growth	131	Sporobolomyces sp.
131 Septate hyphae	Sporobolomyces sp.	Aessosporon sp. Sporobolomyces sp.
132 Septate hyphae	133	Sporidibolus sp. Sporobolomyces sp.
133 Starch growth	Rhodotorula sp.	Sporobolomyces sp.
134 Melezitose growth	135	Rhodotorula spp.
135 Sucrose growth	Rhodotorula sp.	136
136 Growth without thiamin	Rhodospiridium sp.	Rhodotorula sp.
137 L-Sorbose growth	Phaffia sp.	Sporobolomyces sp.
138 Erythritol growth	139	Cryptococcus spp.
139 Nitrate growth	Cryptococcus sp.	Rhodospiridium spp.
140 Ascospores spherical, oval or reniform	141	185
141 Nitrate growth	142	177
142 Apical budding	143	175
143 L-Rhamnose growth	144	168
144 Sucrose growth	145	149
145 Filamentous	Pichia spp.	146
146 Septate hyphae	Pichia spp.	147
147 D-Glucitol growth	148	Saccharomycopsis spp.
148 Trehalose growth	Saccharomycopsis sp.	Nematospora sp.
149 Erythritol growth	150	164
150 D-Mannitol growth	151	155
151 Filamentous	Pichia sp.	152
152 Septate hyphae	Dekkera spp.	153
153 D-Glucitol growth	154	Saccharomycopsis spp.
154 Cellobiose growth	Nematospora sp.	Saccharomycopsis sp.
155 Starch growth	156	162
156 Citrate growth	157	160
157 D-Galactose growth	158	Metschnikowia spp.
158 Melezitose growth	Saccharomycopsis sp.	159
159 Growth without pyridoxine	Pichia sp.	Metschnikowia spp.
160 D-Galactose growth	Pichia spp.	161
161 Raffinose growth	Metschnikowia spp.	Pichia spp.

Table 4—cont.

	Negative	Positive
162 Septate hyphae	163	Saccharomycopsis sp.
163 D,L-Lactate growth	Schwanniomyces sp.	Pichia sp.
164 Starch growth	165	166
165 Septate hyphae	Pichia spp.	Ambrosiozyma spp.
166 D-Galactose growth	Saccharomycopsis sp.	167
167 Raffinose growth	Pichia sp.	Hyphopichia sp.
168 myo-Inositol growth	169	172
169 Erythritol growth	Pichia spp.	170
170 D-Galactose growth	171	Pichia spp.
171 Sucrose growth	Pichia sp.	Ambrosiozyma sp.
172 L-Sorbose growth	173	174
173 D-Galactose growth	Botryosascus sp.	Pichia sp.
174 Sucrose growth	Hansenula sp.	Stephanosascus sp.
175 Cellobiose growth	176	Hanseniaspora spp.
176 D-Glucose fermentation	Arthroascus sp.	Wickerhamia sp.
177 D-Mannitol growth	178	179
178 Maltose fermentation	Hansenula spp.	Dekkera spp.
179 Sucrose growth	180	182
180 D-Galactose growth	Hansenula spp.	181
181 L-Rhamnose growth	Pachysolen sp.	Hansenula sp.
182 Erythritol growth	Hansenula spp.	183
183 Raffinose growth	184	Hansenula spp.
184 D,L-Lactate growth	Hansenula spp.	Hormoascus sp.
185 Apical budding	186	260
186 Budding cells	Schizosaccharomyces spp.	187
187 Erythritol growth	188	252
188 Nitrate growth	189	251
189 L-Rhamnose growth	190	250
190 Raffinose growth	191	224
191 Cellobiose growth	192	222
192 Growth in 60% D-glucose	193	217
193 L-Sorbose growth	194	214
194 D-Galactose growth	195	201
195 Citrate growth	196	Pichia spp.
196 Ethylamine growth	197	Pichia spp.
197 Growth with 0.01% cycloheximide	198	200
198 D-Mannitol growth	199	Pichia sp.
		Saccharomyces sp.
199 Glycerol growth	Pichia sp.	Pichia spp.
	Saccharomyces spp.	Saccharomyces sp.
200 D,L-Lactate growth	Kluyveromyces sp.	Pichia spp.
	Pichia sp.	
201 D-Glucose fermentation	202	203
202 Septate hyphae	Pichia sp.	Guilliermondella sp.
203 D,L-Lactate growth	204	213
204 Ethylamine growth	205	Saccharomyces sp.
205 Growth without niacin	206	208
206 Growth with 0.01% cycloheximide	207	Saccharomyces spp.
207 D-Ribose growth	Saccharomyces spp.	Kluyveromyces sp.
		Saccharomyces sp.
208 D-Ribose growth	209	Kluyveromyces sp.
209 Growth without pyridoxine	210	211
210 Growth without biotin	Pachytichospora sp.	Kluyveromyces sp.
	Saccharomyces sp.	Saccharomyces sp.
211 Glycerol growth	Saccharomyces spp.	212
212 D-delta-Gluconolactone growth	Saccharomyces sp.	Kluyveromyces sp.

Table 4—cont.

	Negative	Positive
213 Septate hyphae	Saccharomyces spp.	Guilliermondella sp.
214 Septate hyphae	215	Zendera sp.
215 Sucrose growth	Pichia spp.	216
216 Starch growth	Lodderomyces sp.	Pichia sp.
217 D-Glucose fermentation	Pichia spp.	218
218 Growth without pyridoxine	Kluyveromyces sp. Pichia spp.	219
219 D,L-Lactate growth	220	221
220 D-Glucitol growth	Pichia sp. Torulaspora sp. Zygosaccharomyces spp.	Torulaspora sp. Zygosaccharomyces spp.
221 Filamentous	Pichia sp. Torulaspora sp.	Pichia spp.
222 Citrate growth	223	Pichia spp.
223 D-Glucose fermentation	Lipomyces sp.	Kluyveromyces spp.
224 D-Glucose fermentation	225	228
225 Starch growth	Debaryomyces spp.	226
226 Growth without thiamin	Pichia sp.	227
227 D, L-Lactate growth	Lipomyces sp.	Debaryomyces sp.
228 Melibiose growth	229	243
229 Cellobiose growth	230	241
230 Growth without niacin	231	233
231 Ethylamine growth	232	Kluyveromyces spp.
232 L-Sorbose growth	Saccharomyces sp.	Kluyveromyces sp.
233 L-Sorbose growth	234	240
234 2-Ketogluconate growth	235	Torulaspora spp.
235 Growth in 60% D-glucose	236	Torulaspora sp.
236 D-Galactose growth	237	238
237 Growth with 0.01% cycloheximide	Saccharomyces sp.	Torulaspora sp.
238 Glycerol growth	Saccharomyces spp.	239
239 D-delta-Gluconolactone growth	Saccharomyces sp.	Kluyveromyces spp.
240 Filamentous	Torulaspora sp.	Kluyveromyces sp.
241 Growth without niacin	Kluyveromyces spp.	242
242 Starch growth	Zygosaccharomyces sp.	Debaryomyces sp.
243 Cellobiose growth	244	248
244 Growth in 60% D-glucose	245	Torulaspora sp.
245 Growth with 0.01% cycloheximide	246	Zygosaccharomyces spp.
246 D-delta-Gluconolactone growth	Saccharomyces spp.	247
247 Growth in 50% D-glucose	Saccharomyces sp.	Zygosaccharomyces sp.
248 Lactose growth	249	Debaryomyces spp.
249 Starch growth	Saccharomyces sp.	Debaryomyces sp.
250 Sucrose growth	Pichia sp.	Debaryomyces spp.
251 D-Glucose fermentation	Wickerhamiella sp.	Citeromyces sp.
252 Starch formation	253	Lipomyces spp.
253 Maltose growth	Pichia sp.	254
254 Inulin growth	255	259
255 L-Rhamnose growth	Debaryomyces spp.	256
256 Melibiose growth	257	258
257 Trehalose fermentation	Debaryomyces sp.	Wingea sp.
258 Lactose growth	Debaryomyces spp.	Debaryomyces sp. Pichia sp.
259 Growth in 50% D-glucose	Debaryomyces spp. Lipomyces sp.	Debaryomyces spp.
260 Cellobiose growth	Nadsonia spp.	261
261 Raffinose growth	Hanseniaspora spp.	Saccharomycodes sp.

greater detail than that required for the key in Table 4, needing more expert knowledge and experience. The physiological tests which constitute the greatest part of our key are also less time-consuming than microscopical examinations and can be carried out by a less experienced operator. Moreover, the key in Table 4 does not require examination of the sexual cycle of the basidiomycetous species.

However, in the same way that a key for identifying species cannot identify new species not in the key, the key in Table 4 cannot be applied to species other than those listed by Barnett *et al.* (1979). Any new species will be assigned to the genus of a species whose test results in the key are identical to those of the new species, but this may not give the correct genus as the genera are not classified in terms of the characteristics used in the key. However, the key is still only marginally less useful than that of Lodder (1970), in view of the many recent changes in generic characteristics and the newly invented genera, to which Lodder's key also cannot apply.

APPENDIX I

Irredundant test sets to identify groups of taxa

Barnett *et al.* (1979) used a method (reviewed by Payne & Preece, 1980) that determined all irredundant test sets available to identify the yeast species in a particular key. This method can be adapted to form sets to identify groups instead of individual taxa. In the first stage a triangular array is formed and in the (i, j) th position ($i < j$) are listed the tests that can distinguish the pair of taxa, i and j . To identify groups instead of individual taxa, array entries are omitted where each member of the pair of taxa is from the same group. Table 5 contains the appropriate array for the groups in Table 1; taxa C and D have different results with tests 1 and 5, thus the (3,4)th entry is '1,5'. Any entries that contain other entries are deleted. For example, as '4,5' is an entry in Table 5, '1,3,4,5', '3,4,5' etc. are deleted. [Such entries can be omitted since, in order to distinguish between the pair of taxa A and D, corresponding to entry '4,5' in Table 5, either test 4 or test 5 must be in the irredundant set. These tests (4 and 5) will also distinguish between the pairs of taxa corresponding to entries '1,3,4,5' etc.] The deleted entries in Table 5 are enclosed in brackets.

The surviving entries are then expressed as a sum and are multiplied together according to the Boolean rules $ii = i$ and, for example, $ij + ij = ij$.

For example, from Table 5,

$$\begin{aligned} & (4 + 5)(2 + 4)(1 + 4)(1 + 6)(1 + 5)(3 + 5)(1 + 2)(2 + 5)(2 + 6)(3 + 6)(5 + 6) \\ &= (24 + 44 + 25 + 45)(1 + 4)(1 + 6) \dots \\ &= (4 + 25)(1 + 4)(1 + 6) \dots \end{aligned}$$

Each term in the first bracket is a set of tests that can distinguish between taxa A and D and taxa A and G. (This indicates the rationale of the rules. '44' becomes '4' because there is no need to record a test more than once in a set. The fact that '4' is a set in its own right implies that '24' and '45' contain redundant tests 2 and 5, respectively; if these are deleted two more instances of set '4' would be obtained; there is no need to record a set more than once so '24' and '45' can be deleted.) Once all the brackets have been multiplied together the sets each contain tests to distinguish all the pairs of taxa belonging to different groups, and if any test in one of the sets is deleted there will be some pair of taxa in different groups that can no longer be distinguished.

Continuing the multiplication from Table 5,

$$\begin{aligned} & (4 + 25)(1 + 4)(1 + 6)(1 + 5)(3 + 5)(1 + 2)(2 + 5)(2 + 6)(3 + 6)(5 + 6) \\ &= (4 + 125)(1 + 6) \dots \\ &= (14 + 46 + 125)(1 + 5) \dots \\ &= (14 + 456 + 125)(3 + 5) \dots \\ &= (134 + 145 + 456 + 125)(1 + 2) \dots \\ &= (134 + 145 + 2456 + 125)(2 + 5) \dots \\ &= (1234 + 145 + 2456 + 125)(2 + 6) \dots \\ &= (1234 + 1456 + 2456 + 125)(3 + 6) \dots \\ &= (1234 + 1456 + 2456 + 1235 + 1256)(5 + 6) \\ &= (12346 + 1456 + 2456 + 1235 + 1256) \end{aligned}$$

There are five possible irredundant sets of tests, namely, 12346, 1456, 2456, 1235 and 1256. These sets cannot be obtained from the group data in Table 1 as there is no single test that can distinguish any of the pairs of groups.

Table 5. Lists of tests that can distinguish the taxa of Table 1 that are in different groups

Taxon	Group	Taxon								
		A	B	C	D	E	F	G	H	I
A	I	-	-	-	4,5	(1,3,4,5)	(3,4,5)	2,4	(2,5,6)	(2,3,4,6)
B	I		-	-	(1,2,4)	(2,3,4)	(1,3,4)	1,4	1,6	(1,3,4,6)
C	I			-	1,5	3,5	(1,3,5)	1,2	(1,2,5,6)	(1,2,3,6)
D	II				-	-	-	2,5	2,6	(2,3,5,6)
E	II					-	-	(1,2,3,5)	(1,2,3)	(1,2,5)
F	II						-	(3,5)	3,6	5,6
G	III							-	-	-
H	III								-	-
I	III									-

See Appendix 1 for meaning of entries in parentheses.

When only a single irredundant set is required, it may be formed sequentially by choosing tests one at a time. Payne & Preece (1980) reviewed criteria for deciding which test to include at each stage. Many of these criteria involve the separation coefficient of Gyllenberg (1964). This is the number of pairs of taxa (i, j) that are distinguished either by the current test, or by some test already in the set. To select tests to distinguish groups instead of individual taxa, only pairs of taxa belonging to different groups need to be considered.

APPENDIX 2

Criterion functions for selecting tests to identify groups of taxa

Selection criterion functions, designed to select tests to identify individual taxa, require modification to select tests to identify groups of taxa. An exception is the function of Dallwitz (1974), whose program allows intra-taxon variability to be expressed by specifying more than one 'item' for a taxon. Thus, each taxon may itself be a group of several 'items'. Most taxa will, however, consist of a single 'item' so, at any point in the key, there will be few (if any) tests that are variable for all the groups of 'items' which occur there; Dallwitz's function does not satisfactorily distinguish between tests with this much variability. For yeast genera, however, so many results are variable that such tests can be expected to occur at many points of the key. The two functions derived below can cope with such tests.

The first function is obtained by modifying the function M_v of Payne (1981):

$$(M_v)_i = - \sum_{k=1}^{m_i} \{ (p_{ik} + r_i/m_i) (1 - r_i - p_{ik}) \}$$

where m_i is the number of possible results to test i , p_{ik} is the proportion of taxa in the current subset that always give result k to test i , and r_i is the proportion of taxa in the current subset with variable results to test i . This is an extension, to tests with more than two possible results, of the function DV derived by Morse (1971) from Gyllenberg's Separation Coefficient. The term in the first bracket is the proportion of specimens that give result k to test i assuming that, for each variable taxon, there is an equal probability ($1/m_i$) of obtaining results, 1, 2, ..., m_i . The second term is the proportion of specimens belonging to taxa that cannot give result k . Thus $(M_v)_i$ is minus the proportion of pairs of taxa separated (either wholly or partially) by test i . For identifying groups, this function should become minus the proportion of pairs of taxa in different groups separated by the test; that is

$$\begin{aligned} (G_v)_i &= - \sum_{k=1}^{m_i} \sum_{j=1}^n \left[(q_{ijk} + s_{ij}/m_i) \times \sum_{1 \neq h} \left\{ \left(\sum_{h=1}^{m_i} q_{ilh} \right) - q_{ilk} \right\} \right] \\ &= - \sum_{k=1}^{m_i} \sum_{j=1}^n q_{ijk}^2 + \sum_{k=1}^{m_i} \left(\sum_{j=1}^n q_{ijk} \right)^2 + \sum_{j=1}^n \left(\sum_{k=1}^{m_i} q_{ijk} \right)^2 - \left(\sum_{k=1}^{m_i} \sum_{j=1}^n q_{ijk} \right)^2 \\ &\quad + (1 - 1/m_i) \left\{ \sum_{j=1}^n \left(s_{ij} \times \sum_{k=1}^{m_i} q_{ijk} \right) - \left(\sum_{j=1}^n s_{ij} \right) \times \left(\sum_{k=1}^{m_i} q_{ijk} \right) \right\} \end{aligned}$$

where n is the number of groups, q_{ijk} the proportion of taxa in the current subset that are from group j and that always give result k to test i , and s_{ij} is the proportion of taxa from group j that have variable results.

The second function, which was used to construct the key to yeast genera in Table 4, is derived from the function M_e of Brown (1977). This selects the test for which the expected entropy of the posterior probabilities of the taxa, given the result of the test, is minimum. Thus, the aim when selecting each test is to make the probabilities that the specimen belongs to each taxon as different as possible. This will be achieved when, for each result, all the probabilities except one are zero; that is, when the subsets formed by the test all contain only one taxon. The derivation, like that of M_v , assumes that equal proportions of variable taxa give each result.

$$(M_e)_i = \sum_{k=1}^{m_i} \{(p_{ik} + r_i/m_i) \log (p_{ik} + r_i/m_i)\} + r_i \log m_i$$

[This is the negative of the function of Brown (1977), who defined the best test to be that with maximum function value.] Under the same assumptions, the expected entropy of the posterior probabilities of the groups is given by

$$(G_e)_i = - \sum_{k=1}^{m_i} \left[t_{ik} \times \sum_{j=1}^n \left[\{(q_{ijk} + s_{ij}/m_i)/t_{ik}\} \log \{(q_{ijk} + s_{ij}/m_i)/t_{ik}\} \right] \right]$$

where $t_{ik} = \sum_{j=1}^n (q_{ijk} + s_{ij}/m_i)$. Thus,

$$(G_e)_i = \sum_{k=1}^{m_i} \left\{ \sum_{j=1}^n (q_{ijk} + s_{ij}/m_i) \right\} \log \left\{ \sum_{j=1}^n (q_{ijk} + s_{ij}/m_i) \right\} - \sum_{k=1}^{m_i} \sum_{j=1}^n \left\{ (q_{ijk} + s_{ij}/m_i) \log (q_{ijk} + s_{ij}/m_i) \right\}$$

An alternative justification for M_e (Payne & Preece, 1980; Payne, 1981) uses the noiseless coding theorem of Shannon (1948) to relate $(M_e)_i$ to the expected number of tests required to complete the key after test i , assuming that this is done optimally. This enables M_e to be extended to tests with different costs. However, this would be less convincing for G_e because to complete a key to the groups in an optimal way requires tests to be available that have constant results within each group.

The assumption that equal proportions of variable taxa give each result, which greatly simplifies the algebraic form of M_v , G_v , M_e and G_e , is not crucial. If it were badly wrong, the test selected might not be the best available and the resulting key might be less efficient; however, the identifications obtained would still be correct. In most situations the assumption will be reasonable – either because the probabilities are known to be nearly equal, or because (as with the yeasts) there is not sufficient information to contradict it. However, if estimates of the probabilities are available, the functions can easily be modified. For example, the full form of M_e is given in equations (4) and (5) of Payne (1981).

Use of the expected entropy of the posterior probabilities of the taxa to select test for probabilistic identification, has been discussed by, for example, Good (1970), Moiseeva & Usov (1969), Taylor (1970), Knill-Jones *et al.* (1973) and Payne (1975). Other functions used for this purpose can be adapted similarly.

REFERENCES

- BARNETT, J. A. & PANKHURST, R. J. (1974). *A New Key to the Yeasts*. Amsterdam: North-Holland Publishing Co.
- BARNETT, J. A., BASCOMB, S. & GOWER, J. C. (1975). A maximal predictive classification of Klebsiellae and of the yeasts. *Journal of General Microbiology* **86**, 93–102.
- BARNETT, J. A., PAYNE, R. W. & YARROW, D. (1979). *A Guide to Identifying and Classifying Yeasts*. Cambridge: Cambridge University Press.
- BROWN, P. J. (1977). Functions for selecting tests in diagnostic key construction. *Biometrika* **64**, 589–596.
- DALLWITZ, M. J. (1974). A flexible computer program for generating identification keys. *Systematic Zoology* **27**, 50–57.
- GOOD, I. J. (1970). Some statistical methods in machine-intelligence research. *Mathematical Biosciences* **6**, 185–208.
- GOWER, J. C. (1973). Classification problems. *Bulletin of the International Statistical Institute* **44**, 296–301.
- GOWER, J. C. (1974). Maximal predictive classification. *Biometrics* **30**, 643–654.
- GYLLENBERG, H. A. (1964). An approach to numerical description of biological populations. *Annales*

- Academiae scientiarum fennicae, Series A IV, Biological Part* 81, 1–23.
- KNILL-JONES, R. P., STERN, R. B., GIRMES, D. H., MAXWELL, J. D., THOMPSON, R. P. H. & WILLIAMS, R. (1973). Use of sequential Bayesian model in diagnosis of jaundice by computer. *British Medical Journal* 1, 530–534.
- KREGER-VAN RIJ, N. J. W. (1970*a*). *Debaryomyces Lodder et Kreger-van Rij nom. conserv.* In *The Yeasts. A Taxonomic Study*, pp. 129–156. Edited by J. Lodder. Amsterdam: North-Holland Publishing Co.
- KREGER-VAN RIJ, N. J. W. (1970*b*). *Pichia Hansen.* In *The Yeasts. A Taxonomic Study*, pp. 455–554. Edited by J. Lodder. Amsterdam: North-Holland Publishing Co.
- LODDER, J. (1970). Introduction to the chapters IV, V, VI and VII, and key to the genera. In *The Yeasts. A Taxonomic Study*, pp. 114–120. Edited by J. Lodder. Amsterdam: North-Holland Publishing Co.
- MOISEVA, N. I. & USOV, V. V. (1969). Some medical and mathematical aspects of computer diagnosis. *Proceedings of the Institution of Electrical and Electronics Engineers* 57, 1919–1925.
- MORSE, L. E. (1971). Specimen identification and key construction with time-sharing computers. *Taxon* 20, 269–282.
- PANKHURST, R. J. (1970). A computer program for generating diagnostic keys. *Computer Journal* 13, 145–151.
- PAYNE, R. W. (1975). Genkey: a program for constructing diagnostic keys. In *Biological Identification with Computers*, pp. 65–72. Edited by R. J. Pankhurst. London: Academic Press.
- PAYNE, R. W. (1978). *Genkey: a Program for Constructing and Printing Identification Keys and Diagnostic Tables.* Harpenden: Rothamsted Experimental Station.
- PAYNE, R. W. (1981). Selection criteria for the construction of efficient diagnostic keys. *Journal of Statistical Planning and Inference* 5, 27–36.
- PAYNE, R. W. & PREECE, D. A. (1980). Identification keys and diagnostic tables: a review (with discussion). *Journal of the Royal Statistical Society, Series A* 143, 253–292.
- PAYNE, R. W., WALTON, E. & BARNETT, J. A. (1974). A new way of representing diagnostic keys. *Journal of General Microbiology* 83, 413–414.
- SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423, 623–656.
- TAYLOR, T. R. (1970). Computer-guided diagnosis. *Journal of the Royal College of Physicians of London* 4, 188–194.