

Algorithms for Identification Key Generation and Optimization with Application to Yeast Identification

Alan P. Reynolds¹, Jo L. Dicks², Ian N. Roberts³, Jan-Jap Wesselink¹,
Beatriz de la Iglesia¹, Vincent Robert⁴, Teun Boekhout⁴, and
Victor J. Rayward-Smith¹

¹ School of Information Systems, University of East Anglia, Norwich, UK

² Computational Biology Group, John Innes Centre, Norwich, UK

³ National Collection of Yeast Cultures, Institute of Food Research, Norwich, UK

⁴ Centraalbureau voor Schimmelcultures, Utrecht, The Netherlands

Abstract. Algorithms for the automated creation of low cost identification keys are described and theoretical and empirical justifications are provided. The algorithms are shown to handle differing test costs, prior probabilities for each potential diagnosis and tests that produce uncertain results. The approach is then extended to cover situations where more than one measure of cost is of importance, by allowing tests to be performed in batches. Experiments are performed on a real-world case study involving the identification of yeasts.

1 Introduction

Identification problems occur in a wide range of disciplines. Biological applications include the identification of plant or animal species from observable specimen characteristics [1], the diagnosis of disease in plants, animals and humans and the identification of bacteria from the results of laboratory tests [2]. Each situation is different, and so a range of identification tools are used.

One such tool is the identification (or diagnostic) key. Given a database of test results for known species, an identification key provides a sequence of tests to be performed, with the choice of test depending upon the results of previous tests. This paper describes the generation and optimization of such keys. The algorithms described handle both variable test costs and uncertain and unknown data, creating keys where tests are performed individually or in batches.

The construction of identification keys is introduced in sections 2 and 3. Section 4 describes and compares greedy key construction heuristics, with results in section 6. Section 7 describes the randomization of the greedy algorithm, to reduce further the cost of the keys produced. Finally, section 8 describes the adaptation of these algorithms to handle batches of tests.

2 Testing Approaches and Key Evaluation

When performing tests to identify specimens, one of three different approaches may be appropriate.

Perform all tests at once. This approach is suitable if the speed of identification is more important than the material costs, provided tests may be performed in parallel. In this case, it is desirable to find a minimal cost subset of the tests that maximizes the likelihood of identification [3].

Perform tests individually. If tests cannot be performed in parallel, or the material costs of the tests are high, tests should be performed individually. Each test is chosen after analysing the results of the previous tests.

Perform tests in batches. In some situations, it may be appropriate to perform the tests in batches. For example, if specimens must be sent off to a distant lab for the tests, performing all tests at once may be expensive in terms of the costs of the tests, but performing tests individually will certainly be expensive in terms of time.

This paper focuses on the production of identification keys for the cases where tests are either performed individually or in batches.

A number of different measures may be used to evaluate a key. Some considerations are as follows:

- The expected total test cost should preferably be minimized.
- It may be desirable to complete the identification quickly.
- It may be important to discover specimens of a certain kind quickly. For example, if medical samples are being tested, it may be important to discover dangerous, fast acting but rare diseases quickly.

The expected total test cost is the most commonly used measure of key efficiency [4,5,6], and is used throughout this paper, with the exception of section 8.

3 A Simple Example

Figure 1 shows test data for four yeast species, and the corresponding optimum key, assuming that each species is equally likely. Two tests are needed, regardless of the species the test sample happens to be, so the expected test cost is 20.

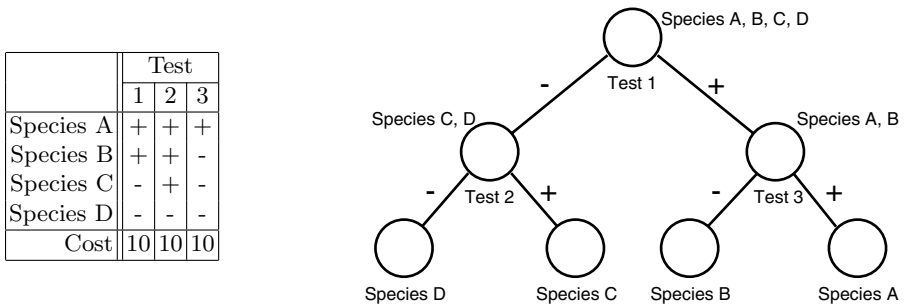


Fig. 1. Simple species data with the corresponding optimum identification key.

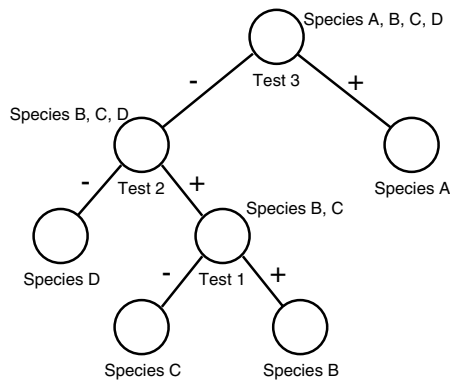


Fig. 2. Alternative identification key

An alternative key is shown in figure 2. Here species A is identified with just one test, but if the sample is equally likely to be any of the species, the expected test cost is now $(10 + 20 + 30 + 30)/4 = 22.5$. However, if species A is more common than the other species, the resulting expected test cost may be less than 20, making the resulting key preferable to that shown in figure 1.

Suppose an extra test is added, that produces a positive result for species A, C and D, but produces an uncertain result for species B. Let each outcome for species B be equally likely. If the test is cheap, it may be present in the optimum key, despite the fact that a positive result does not eliminate any of the potential yeast species. This is shown in figure 3. The cost of identification, if the sample

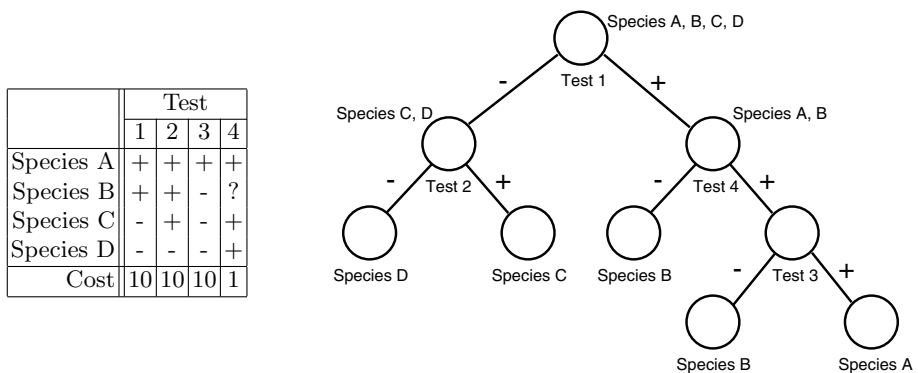


Fig. 3. The optimum key, upon the addition of an extra test.

is species A has increased to 21, but the cost of identification of species B has changed to either 11 or 21, depending on the result obtained for test 4. This leads to an expected test cost of $(21 + (11 + 21)/2 + 20 + 20)/4 = 19.25$, provided

each species is equally likely. Notice that, after obtaining a positive result to test 4, the sample is twice as likely to be species A than species B.

4 A Greedy Algorithm for Key Construction

4.1 Key Construction

As illustrated in figures 1 and 3, each node in an identification key has a set of species and, if the node is not a leaf, a test to be performed. A child node contains each species of the parent node that produces a particular result to the parent's test. A leaf node occurs only if identification is complete or no useful tests remain.

A greedy key construction heuristic starts at the top node. At each node, the heuristic either determines that the node is a leaf node, or finds the 'best' test, according to some measure, and creates child nodes associated with each test outcome. The algorithm stops when all nodes have been examined. The three greedy heuristics discussed in this section differ only in the measure used to select tests.

```

Create node top and place a reference to it on empty heap H;
S(top) := set of all yeast species in database, with associated probabilities;
T(top) := set of all tests;
while (H not empty)
  Remove a node, N, from the heap;
  if  $|S(N)| > 1$  and there exists a useful test in T(N) then
    Find the 'best' test  $t_{best} \in T(N)$ ;
    for each test outcome r produced by at least one species in S(N)
      Create child node C and place a reference to it on heap H;
       $S(C) := \{s \in S(N) : s \text{ may produce outcome } r \text{ for test } t_{best}\}$ ;
      Calculate the new probabilities associated with these species;
       $T(C) := T(N) - \{t_{best}\}$ ;
    endfor
  endif
endwhile

```

Fig. 4. Basic pseudo-code for greedy key construction

4.2 Test Selection

Payne and Preece's review of identification tables and diagnostic keys [7] reports several test selection criteria. Those that are suitable for use with tests with unequal costs are compared in a later paper by Payne and Thompson [4].

Define:

$S = \{(s_1, p_1), \dots, (s_n, p_n)\}$: Set of potential species and associated probabilities of occurrence;
S_{ij}	: Similar set containing potential species after test i produces result j ;
$P(i, j)$: Probability that test i produces result j ;
$E(S)$: Estimate of the cost of completing identification from set S ;
c_i	: Cost of test i ;
m_i	: Number of possible outcomes of test i .

The functions used in Payne and Thompson [4] as selection criteria are all of the form

$$PT_i = c_i + \sum_{j=1}^{m_i} P(i, j)E(S_{ij}).$$

Here PT_i is the cost of test i plus an estimate of the expected remaining test costs after test i is performed, and is therefore an estimate of the total test costs. The test that minimizes this function is considered to be the best test.

A problem with this approach is that of finding a good estimate, $E(S)$. Payne and Thompson suggest a number of possibilities. However, each of the estimates suggested can be shown to be either optimistic or pessimistic, depending on the number and quality of the available tests.

A second problem is that such criteria tend to favour poorly performing cheap tests over more useful expensive tests. Consider the following example. A sample specimen may be one of only two species, each equally likely. $E(S)$ is the estimate of the remaining test costs. Test one has a cost of one, but produces the same result for both species. Test two has cost five and completes identification. Then $PT_1 = 1 + E(S)$ and $PT_2 = 5$. If $E(S)$ equals three, the average test cost, then the useless cheap test is chosen. This choice of test can only increase the expected test costs.

The approach used in this paper is to calculate a measure of the expected amount of work performed by the test. If the work performed, upon application of a test, is defined to be the reduction in the estimate of the remaining test costs, then the expected amount of work performed is given by

$$W_i = E(S) - \sum_{j=1}^{m_i} P(i, j)E(S_{ij}).$$

This may then be divided by the test cost, to give an estimate of the work performed per unit cost. The test that performs the most work per unit cost is selected. This selection criterion has neither of the aforementioned problems.

Of the selection criteria discussed in Payne and Thompson [4], those that use Shannon entropy [8] are of particular interest. The entropy of set S is given by

$$H(S) = - \sum_{k=1}^n p_k \log_2 p_k.$$

$H(S)/\log_2 m$ can be shown to be a very good approximation to the number of tests required for identification if all conceivable tests that produce m results are available. After test i is performed, the expected entropy is given by

$$H(S, i) = \sum_{j=1}^{m_i} P(i, j) H(S_{ij}).$$

Payne and Thompson [4] give the following two functions that may be used as test selection criterion.

$$(PT1)_i = c_i + (c_{min}/\log_2 \bar{m})H(S, i)$$

$$(PT2)_i = c_i + (\bar{c}/\log_2 \bar{m})H(S, i)$$

Here c_{min} is the cost of the cheapest test, \bar{c} is the mean test cost and \bar{m} is the mean number of outcomes of a test.

In the approach used in this paper, Shannon entropy is used as a measure of how much work is required to complete identification. Setting the estimate function $E(S)$ to be equal to the entropy $H(S)$, the expected amount of work performed per unit cost, upon application of test i , is given by the information gain ratio

$$GR_i = \frac{W_i}{c_i} = \frac{H(S) - H(S, i)}{c_i}.$$

The test selected is the one that maximizes this value.

Note that Shannon entropy is commonly used in the closely related research area of decision tree induction in machine learning and data mining [9].

5 The Case Studies

5.1 The Data Used

Data from the Centraalbureau voor Schimmeltcultures (<http://www.cbs.knaw.nl>) was used for the experiments. Results of 96 different tests on 742 yeast species are provided. The possible responses are as shown in table 1. Test costs and the prior probabilities of each specimen are not given.

The data was manipulated in a number of ways in order to provide more than one case study.

Table 1. Possible responses

Meaning	Symbol	Meaning	Symbol
Positive	+	Negative, weak and/or delayed	-,w,d
Negative	-	Positive, weak and/or delayed	+,w,d
Weak	w	Negative and/or positive	-,+
Delayed	d	Unknown	-,+,w,d

Simplification of the results range. Either the full results range of table 1 was used, or this range was simplified. In the latter case, any response that was not ‘positive’ or ‘negative’ was replaced by the response ‘other’.

Certain and uncertain results. The reported test outcomes were interpreted as either certain or uncertain outcomes. In the first case, the 8 different result values in table 1 are interpreted as distinct test outcomes. In the second, more realistic case, a test produces one of four different outcomes: ‘negative’, ‘weak’, ‘positive’ or ‘delayed’. The other four possibilities in table 1 are interpreted as being uncertain outcomes, with each outcome in the symbol list occurring with equal probability. If the simplified results range is used, the result ‘other’ is interpreted as an uncertain outcome.

Test costs. Tests were either assigned unit costs or random integer costs from 1 to 5.

Prior probabilities. Each species was either assumed to be equally likely, or assigned a random prior probability.

The resulting 16 case studies provided the test bed for the algorithms of this paper. The randomly generated test costs and prior species probabilities are listed at <http://www.sys.uea.ac.uk/biodm/>.

6 Results: Greedy Heuristics

When each test has unit cost, each of the selection criteria discussed produce the same result, since each selects the test that minimizes the Shannon entropy after the test is performed. The results in this case are shown in table 2.

By using the randomly generated test costs, it is possible to compare the three selection criteria. Table 3 shows the results of this comparison. When creating the test trees, useless tests were discarded, preventing selection criteria *PT1* and *PT2* from choosing cheap but useless tests.

When the tests have uncertain outcomes, full identification cannot be guaranteed, as it is possible for two or more species to produce the same result for each of the tests. In this case, tests were applied until no test remains that can possibly produce an identification.

Notice that in all but one of the experiments, selection criterion *GR* produces keys with lower expected test costs than both *PT1* and *PT2*. Keys produced with selection criterion *GR* also tend to be smaller, although this is only really an issue if the key needs to be produced in printed form. Since the keys produced when there are uncertain results and varying test costs are large, the rest of this paper concentrates on the case where test outcomes are certain.

7 Randomizing the Algorithm

When solving an optimization problem, a metaheuristic can often be expected to outperform a greedy algorithm in terms of solution quality. Unfortunately, in this case it is difficult to find a suitable neighbourhood structure for use in local search techniques. However, a simplified GRASP algorithm [10] may be applied

Table 2. Unit test costs

Outcome type	Results range	Species probabilities	Expected #tests	#Nodes
Certain	Simplified	Equal	6.248	1194
		Random	6.176	1209
	Full	Equal	4.624	1073
		Random	4.554	1071
Uncertain	Simplified	Equal	12.887	30223
		Random	12.912	27289
	Full	Equal	8.027	182750
		Random	8.048	231858

Table 3. Random test costs

Outcome type	Results range	Species prob.	Expected cost			#nodes		
			<i>PT1</i>	<i>PT2</i>	<i>GR</i>	<i>PT1</i>	<i>PT2</i>	<i>GR</i>
Certain	Simplified	Equal	8.222	8.011	8.015	1311	1245	1300
		Random	10.132	8.136	7.863	1707	1332	1294
	Full	Equal	6.453	6.282	6.240	1200	1170	1165
		Random	8.386	7.542	6.137	1582	1498	1155
Uncertain	Simplified	Equal	37.353	33.824	29.076	893811	535953	205169
		Random	36.986	33.900	29.229	912325	571107	213185
	Full	Equal	20.883	19.078	16.958	5717075	3692803	2165164
		Random	21.059	19.380	17.083	5630820	4167501	2086172

with ease. Instead of choosing the best test, according to selection criterion *GR*, one of the best three tests is selected at random. Many trees are created and the best is selected.

7.1 Results

Experiments were performed to determine the best selection probabilities for the tests. The selection probability for the best test varied between 0.5 and 0.9. The selection probability for the second best test varied from zero to 0.5, provided that the probability of selecting one of the best two tests did not exceed one. Whenever neither of the best two tests was selected, the third was chosen. Each run created 1000 trees. Ten runs were performed for each case.

Figure 5 shows the mean result of the ten runs for each of the probability settings. In this case the simplified results range was used, test costs were equal to one and each species was assumed to be equally likely. Notice that the best results obtained were found by selecting only the best two tests and setting the probability of choosing the best test to its highest value (0.9). Results for the other case studies were remarkably similar, suggesting that the information gain ratio is a remarkably good choice of test selection criterion.

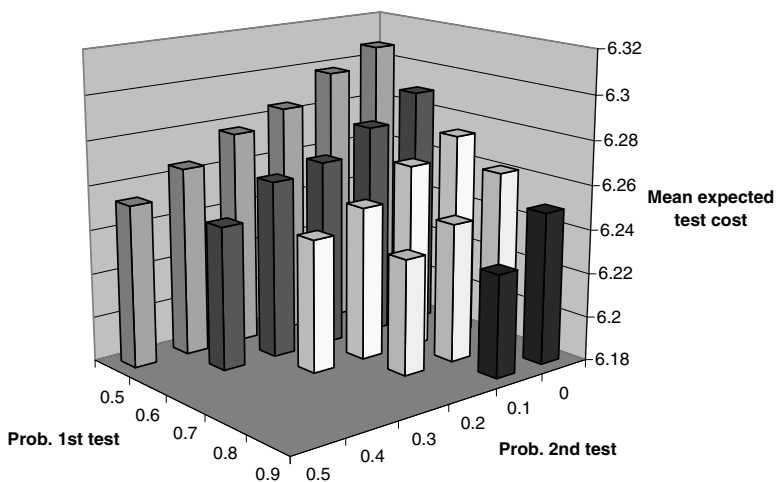


Fig. 5. Mean expected test costs for ten runs

The mean and best results of the ten runs, with test selection probabilities set to 0.9, 0.1 and 0.0 are shown in table 4. The results obtained are not a great deal better than those obtained using the greedy algorithm alone, further suggesting that information gain ratio is a highly effective test selection criterion.

8 Time and Cost Minimization: Performing Tests in Batches

As already discussed, the three approaches to testing — performing all tests at once, performing tests individually and performing tests in batches — are each suited to different situations. When it is important to perform the identification quickly and the material costs are irrelevant, all tests are performed in parallel. When material costs are the overriding factor, each test should be performed individually. It is reasonable to suspect that, when both test duration and material costs are important, testing should be performed in batches.

Table 4. Expected test costs with selection probabilities of 0.9, 0.1, 0.0.

Results	Test outcomes	Equal probabilities		Random probabilities	
		Unit costs	Random costs	Unit costs	Random costs
Mean	Simplified	6.227	8.007	6.152	7.835
	Original	4.606	6.221	4.514	6.109
Best	Simplified	6.221	8.000	6.148	7.825
	Original	4.597	6.214	4.506	6.099

The greedy algorithm must be modified in order to handle batches of tests and objective functions involving both expected test costs and duration.

Objective Function. Functions used for key evaluation are weighted sums, $W = \lambda C + (1 - \lambda)T$, where C is the expected material cost of identification and T is the expected amount of time required.

Batch Quality. To measure the quality of a batch, the information gain is divided by the weighted cost, w , of the batch. This weighted cost is set to $w = \lambda c + (1 - \lambda)t$, where c is the material cost of the batch and t is its duration. Here λ takes the same value as in the evaluation of the key.

The material cost of a batch is simply the sum of the material costs of the individual tests. However, as tests in a batch may be performed in parallel, the duration of a batch is the duration of the longest test.

Batch Selection. It is not feasible to find the best batch of tests by evaluating all possible batches. However, a number of techniques may be applied to this optimization problem. Expending a lot of effort in each optimization is likely to be counter-productive; the best batch according to the selection criterion need not be the best batch for the creation of the key. A simple search algorithm, such as a hillclimber, is likely to be sufficient.

8.1 Results

The following results were obtained by running a stochastic first-found hillclimber at each node of the tree. Neighbourhood moves either added a test to the batch, removed a test, or did both, allowing batch sizes to be determined automatically. The material costs weight, λ , took values between 0.1 and 1. Ten runs were performed for each value of λ and mean results are reported.

Figure 6 shows results for the case study with the full results range, unit material test costs and durations and equal species probabilities. The expected material cost of identification, the expected testing duration and the weighted cost are plotted against the material cost weight. As material test costs become less significant and testing duration becomes more so, the material cost of the key produced by the algorithm increases and the testing duration decreases. Figure 7 illustrates how the average batch size increases as material test costs become less significant compared to testing duration.

9 Further Research

The research outlined in this paper may be extended in a number of ways.

Improved Estimates of Work Performed. Entropy provides a good measure of the work remaining when each species is equally likely, but figure 8 shows that as one species becomes more likely, entropy performs less well.

A possible solution is to use the Huffman algorithm to create optimal trees as shown in Payne and Preece [7], under the assumption that all conceivable binary tests are available. The expected test cost of the trees created may then be used as an alternative to Shannon entropy.

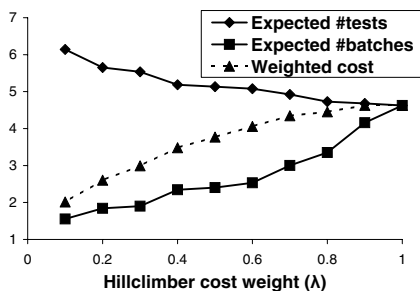


Fig. 6. Expected material costs and time required for identification, with the weighted costs of the keys.

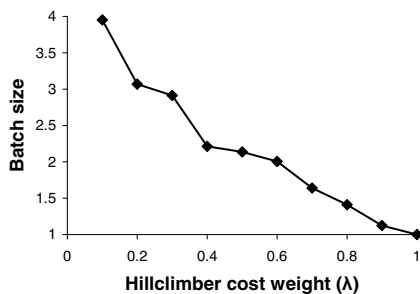


Fig. 7. Batch size.

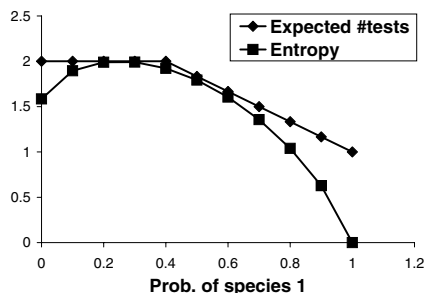


Fig. 8. Expected test costs, provided all conceivable binary tests are available, and entropy for four species, with three equally likely.

Identification of Yeast Class. Sometimes it may only be necessary to determine the class of species to which a specimen belongs, e.g. whether a yeast specimen is one of a set of known food spoilage yeasts. In this case, entropy is not always a reliable measure of the work required to complete identification. Further work to produce a reliable test selection criterion is required.

Although the greedy algorithm described in this paper is less reliable in this situation, the simplified GRASP algorithm has been used to produce useful keys for determining if a yeast specimen is one of the food spoilage species.

Non-independence of Tests. The algorithms described in this paper work under the assumption that tests are independent of each other. In reality this may not be the case. Further work is necessary to account for this.

Practical Application of Greedy Key Construction. In order to *evaluate* a key construction algorithm, it is necessary to create complete identification keys in order to calculate the expected test costs. If a greedy algorithm is used, its *application* to specimen identification does not require the creation of the full key. Instead, the selection criterion is used to determine the tests to be performed as required. The result is that only one branch of the tree is

created per identification. Since trees with an average branch length of just ten may contain over a million nodes, this results in computational savings. This allows more complex methods to be used for test selection. Simulated annealing or tabu search could be applied to the selection of test batches. Test evaluation could be improved by also considering subsequent tests.

10 Conclusions

A simple identification key construction algorithm has been described and shown to perform well on a real-world identification problem. This algorithm can handle variable test costs, uncertain test outcomes and species with differing prior probabilities. Furthermore, it has been shown that the algorithm can be extended to handle situations where both material costs and identification time are important, by allowing tests to be performed in batches.

References

1. M. Edwards and D.R. Morse. The potential for computer-aided identification in biodiversity research. *Trends in Ecology and Evolution*, 10(4):153–158, 1995.
2. T. Wijtzes, M.R. Bruggeman, M.J.R. Nout, and M.H. Zwietering. A computerised system for the identification of lactic acid bacteria. *International Journal of Food Microbiology*, pages 65–70, 1997.
3. B. De la Iglesia, V.J. Rayward-Smith, and J.J. Wesseling. Classification/identification on biological databases. Proc MIC2001, 4th International Metaheuristics Conference, ed. J.P. de Souza, Porto, Portugal, 2001.
4. R.W. Payne and C.J. Thompson. A study of criteria for constructing identification keys containing tests with unequal costs. *Comp. Stats. Quarterly*, 1:43–52, 1989.
5. R.W. Payne and T.J. Dixon. A study of selection criteria for constructing identification keys. In T. Havranek, Z. Sidak, and M. Novak, editors, *COMPSTAT 1984: Proceedings in Computational Statistics*, pages 148–153. Physica-Verlag, 1984.
6. R.W. Payne. Genkey: A program for constructing and printing identification keys and diagnostic tables. Technical Report m00/42529, Rothamsted Experimental Station, Harpenden, Hertfordshire, 1993.
7. R.W. Payne and D.A. Preece. Identification keys and diagnostic tables: a review. *Journal of the Royal Statistical Society, Series A*, 143(3):253–292, 1980.
8. C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423 and 623–656, 1949.
9. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
10. T. A. Feo and M. G. C. Resende. Greedy randomized adaptive search procedures. *Journal of Global Optimization*, 6:109–133, 1995.