Classification of Enterobacteriaceae by minimization of stochastic complexity

H. G. Gyllenberg, M. Gyllenberg, T. Koski, T. Lund, J. Schindler and M. Verlaan

A new method for classifying bacteria is presented and applied to a large set of biochemical data for the Enterobacteriaceae. The method minimizes the bits needed to encode the classes and the items or, equivalently, maximizes the information content of the classification. The resulting taxonomy of Enterobacteriaceae corresponds well to the general structure of earlier classifications. Minimization of stochastic complexity can be considered as a useful tool to create bacterial classifications that are optimal from the point of view of information theory.

Keywords: classification, Enterobacteriaceae, information theory, stochastic complexity, taxonomy

INTRODUCTION

According to Rissanen (1989) the best theory (or model) to explain a given set of data is the one which minimizes the sum of (1) the length in bits of the description of the theory, and (2) the length in bits of the description of the data within the theory or model. This statement can be viewed as a formalization, programmable on a computer, of Occam's razor, the principle that tells us not to introduce more concepts than necessary to explain observed facts. Classifying a collection of items according to some specified method (classification model) can be viewed as a means for encoding information about the data. Following the above-mentioned principle of Rissanen, the best classification is therefore the one which requires the least number of bits to code the classification with respect to the model chosen and to code the items within the classification. The relevant mathematical quantity describing the minimum number of bits is that of stochastic complexity (SC) (Rissanen, 1989).

Gyllenberg et al. (1994a) (see also Gyllenberg et al., 1993 and Gyllenberg & Koski, 1996) gave a precise description of the mathematical model corresponding to probabilistic numerical identification in phenetic bacterial taxonomy (Dybowski & Franklin, 1968; Lapage et al., 1973; Willcox et al., 1980). The 'best' taxonomy according to Rissanen's principle is therefore the one that minimizes SC with respect to this model.

A good classification should have an information content as large as possible (Pankhurst, 1991; Sneath, 1995a, b). Gyllenberg et al. (1994b) showed that minimizing SC amounts to maximizing the information content of the classification. Thus increasing SC implies loss of information whereas decreasing SC indicates gain in information content. Hence SC provides a means to compare different classifications of a given collection of items and to select the best (from an information theory point of view) classification among different alternatives. Since the total number of possible classifications is finite, albeit extremely large, the SC has an

Abbreviations: SC, stochastic complexity; CFARM, classification of Farmer et al. (1985); HMO, hypothetical mean organism; ENTE, Enterobacteriaceae material; ESSY, escherichias, salmonellas, shigellas and yersinias from ENTE; COLI, E. coli strains from ENTE; SCENTE, SCESY, SCCOLI, SC-minimizing classification of ENTE, ESSY, COLI.
absolute number of classes given beforehand, nor is there any a priori given level of similarity that two items have to exceed in order to be placed into the same group. The number of classes is determined solely from the requirement of minimizing the SC.

The purpose of this paper is to demonstrate the applicability of SC minimization as a classification method in microbiological taxonomy by comparing its outcome with a generally applied classification based on other principles. The family Enterobacteriaceae was chosen for demonstration for two reasons: first, because Enterobacteriaceae constitute a particularly well-studied group of bacteria the taxonomy of which can be considered to be rather stable; and second, because the abundant availability of data on Enterobacteriaceae allowed the utilization of a relevantly large set of test material.

The Enterobacteriaceae classification of Farmer et al. (1985) was chosen as the main reference. Both genotypic and phenotypic models were applied in the work of Farmer et al. (1985), e.g. probabilistic numerical identification in the spirit of Lapage et al. (1973). All species and genus names used in this paper follow the nomenclature of Farmer et al. (1985). The discussions of Enterobacteriaceae (Brenner, 1992) in Bergey’s Manual (Krieg & Holt, 1984) and The Prokaryotes (Balows et al., 1992) are in accordance with the conclusions of Farmer et al. (1985). The comparability of the results presented in this paper and those of Farmer et al. (1985) was secured by the fact that the present authors used the binary codes of isolates from the files of Farmer et al. (1985). The classification of Farmer et al. (1985) is referred to as CFARM in subsequent sections of this paper.

METHODS

Material. Altogether data of 5313 strains of Enterobacteriaceae representing 104 species or corresponding biogroups were included in the study. The source of the material (which is listed in Table 1) was the database of Enterobacteriaceae and Vibrionaceae compiled from 1972 to 1989 by the Enteric Bacteriology Laboratories, CDC, Atlanta, GA, USA. There were 67 un-named strains of Shigella in the material. We decided to include them in the runs, and in the files these strains were designated as Shigella provisionally.

The data consisted of 47 binary characters (biochemical reactions) for each specimen (strain, isolate); the characters were the same as those described by Farmer et al. (1985) and in the binary codes they are presented in the same order as by Farmer et al. (1985). Since details of the characters are irrelevant for the purposes of the present paper, readers are referred to Farmer et al. (1985) for this information. Some data were missing in the specimen codes (total frequency: 1.55%). Missing data occurred mainly in isolates from the early 1970s and in character 46 (yellow pigment) for 2803 specimens. Missing bits were replaced by either 0 or 1 determined by coin tossing. It could be concluded that this did not affect the final results (when character 46 and the oldest isolates were omitted the material contained almost no missing data and the classification method gave rise to outcomes almost identical with those obtained using the complete material).

Description of classes. The centroid of a class is by definition the vector giving the frequencies of 1s for the different attributes. Rounding off each component of the centroid to the nearest integer (0 or 1) one obtains the hypothetical mean organism (HMO) of that class (Gower, 1974; Sneath, 1979). As a measure of the heterogeneity of a class we chose its distortion, which is the mean number of bits by which the members of the class differ from the HMO.

Mathematical and computational methods. We describe mathematically a classification of strains with d binary (0 or 1) features into k classes by the numbers

\[ \lambda_1, \lambda_2, \ldots, \lambda_k \quad (\lambda_1 + \ldots + \lambda_k = 1) \]

and

\[ \theta_{ij}; i = 1, \ldots, d; j = 1, \ldots, k, \quad (0 \leq \theta_{ij} \leq 1), \]

where \( \lambda_i \) is the relative frequency of strains in the \( j^{th} \) class and \( \theta_{ij} \) is the relative frequency of 1s in the \( i^{th} \) position in the \( j^{th} \) class. The centroid of the \( j^{th} \) class is the vector \( \theta_{1j}, \theta_{2j}, \ldots, \theta_{dj} \). The distribution of feature vectors \( x = (x_1, x_2, \ldots, x_d) \), \( (x_i = 1 \text{ or } 0) \) of strains in class \( j \) is given by

\[ p_j(x) = \prod_{i=1}^{d} (1 - \theta_{ij})^{1-x_i} \theta_{ij}^{x_i} \]

(Dybowski & Franklin, 1968; Wilcox et al., 1980). As a statistical model of the classification we therefore choose the distribution

\[ p(x) = \sum_{j=1}^{k} \lambda_j p_j(x) \]  

(2)

with the numbers \( k, \lambda_j \) and \( \theta_{ij} \) being the parameters of the model. We emphasize that this statistical representation is simply a mathematically convenient way of defining the classification model and it does not imply any randomness in the data.

It was shown by Gyllenberg et al. (1994b) that the stochastic complexity SC of a set of \( i \) strains with respect to the above model is

\[ \text{SC} = \log \frac{k(k+1)\cdots(t+k-1)}{t_1t_2\cdots t_k} + \sum_{j=1}^{k} \sum_{i=1}^{d} \log \frac{t_j!}{t_j(t_j-t_i)!} \]

(3)

where \( t_i \) is the number of strains in class \( i \) and \( t_j \) is the number of strains in class \( j \) with the \( p^{th} \) feature equal to 1 (log denotes the logarithm to the base 2). The first term in equation (3) describes the complexity of the classification and the second term the complexity of the strains with respect to the classification.

Gyllenberg et al. (1994b) also showed that minimizing the SC with respect to the model (2) amounts to maximizing the information content of the classification.

Even if SC has a well-defined minimum it is practically impossible to find the classification at which it is attained, since an exhaustive enumeration of all classifications is computationally prohibitive even for small data sets. Gyllenberg et al. (1994b) therefore developed an algorithm which finds an approximation of the minimum value of SC for
a given value \( k \) of the number of classes. Repeating the algorithm for different values of \( k \) and choosing the one giving the smallest SC one obtains a classification that should be close to the true SC-minimizing classification.

The algorithm consists of initialization and three steps of re-estimation which are repeated as long as the classification is improving.

1. **Initialization.** \( k \) binary feature vectors are chosen at random. They represent the hypothetical mean organisms (HMOs) of \( k \) classes.

2. **Identification.** The strains in the data are identified, that is, each strain is associated to the nearest HMO. Here the distance is interpreted in terms of codelength, that is, the number of bits needed to represent a strain in relation to the HMO.

3. **Parameter estimation.** The parameters \( \lambda \) and \( \theta_0 \) of the classification obtained in step 2 are calculated.

4. **Redefining the HMOs.** The HMOs of the classification obtained in step 2 are calculated.

The procedure is repeated starting from step 2 using the HMOs found in step 4 at the previous iteration until the HMOs do not change.

The above algorithm was implemented in a computer program written in ANSI-C, almost 6500 lines of code. The compilers Borland C++ 4.51 and GNU-CC 2.7.1 were used. Since the initialization (step 1) is random, the algorithm does not always yield the same result. For each value of \( k \) the algorithm was therefore repeated at least 20 times and the least value of SC was considered as the best value for that particular \( k \). The standard deviation of SC was typically about 0.15 and the classifications corresponding to SC values close to the least value all showed the same general structure.

**RESULTS**

**Minimization of SC**

Fig. 1 shows how the minimum of SC changed as the *Enterobacteriaceae* material (ENTE) was divided into an increasing number of classes. With two classes the minimum of SC was approximately 33, but the minimum decreased rather steeply as the number of classes rose and attained its least value of 21.421 at 69 classes. We denote this SC-minimizing classification SCENTE. With further increase in the number of classes the minimum of SC increased, but slightly. ENTE included 104 traditional species (or corresponding biogroups) and the SC of this traditional classification (CFARM) was 23.124, compared with an SC of 21.888 when ENTE was classified into 104 classes by the method of SC minimization.

What do these SC values indicate? The difference 23.124 – 21.421 = 1.703 in SC between SCENTE and CFARM means that 1.703 more bits per item are needed to encode the classification and the data using CFARM than using the SC-minimizing classification SCENTE (69 classes). In other words SCENTE is \( 2^{1.703} = 3.27 \) times ‘better’ than CFARM. The SC classification into 104 classes, on the other hand, is 2.34 times ‘better’ (from an information theory point of view) than CFARM (2\(^{23.124} - 2^{21.888} = 2.339 \)).

**SC classification into 104 classes**

Before turning to the SC-minimizing classification of ENTE its SC classification into 104 classes will be briefly compared with CFARM, which also contains 104 classes (Table 1). Although the SC classification confirms the general structure of CFARM, differences are obvious. Only 17 of the 104 species (or biogroups) of CFARM were clearly distinguishable in the SC classification. These were *Cedecea davisae*, *Citrobacter diversus*, *C. amalonaticus* 1, *Enterobacter georgeviae*, *Escherichia fergusonii*, *E. hermanii*, *E. vulneris*, *Leclercia adecarboxylata*, *Moellerella wisconsensis*, *Morganella morganii*, *Serratia ficaria*, *S. fonticola*, *S. odorifera* 1, *S. odorifera* 2, *S. plyruthica*, *S. rubidae* and *Tatumella ptyseos*. Other species were divided into two or more classes (particularly *Escherichia coli*, which produced some 20 classes). Certain genera were collected together without separation of species (e.g. *Kluyvera* and *Edwardsiella*) whereas other genera (e.g. *Enterobacter*, *Klebsiella*, *Proteus*, *Providencia*) were split into several classes but not according to the species division of CFARM.

**SC-minimizing classification of Enterobacteriaceae (SCENTE)**

The outcome of the SC-minimizing classification of ENTE, and a comparison with CFARM, are given in Table 2. It gives rise to the following conclusions. (1) Only 11 of the 104 species in CFARM are strictly confirmed (i.e. they formed their own classes, possibly
The size of a class is the number of specimens in that class. The distortion denotes the items' average distance from the HMO measured in number of differing bits.

<table>
<thead>
<tr>
<th>Species</th>
<th>Size</th>
<th>Distortion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Budvicia aquatica</td>
<td>6</td>
<td>2.83</td>
</tr>
<tr>
<td>Buttiauxella agrestis</td>
<td>5</td>
<td>1.20</td>
</tr>
<tr>
<td>Cedecea daviseae</td>
<td>31</td>
<td>2.45</td>
</tr>
<tr>
<td>Cedecea lapagei</td>
<td>10</td>
<td>2.70</td>
</tr>
<tr>
<td>Cedecea neteri</td>
<td>4</td>
<td>0.75</td>
</tr>
<tr>
<td>Cedecea species 3</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>Cedecea species 5</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>Citrobacter amalonaticus</td>
<td>31</td>
<td>2.48</td>
</tr>
<tr>
<td>Citrobacter amalonaticus 1</td>
<td>40</td>
<td>2.58</td>
</tr>
<tr>
<td>Citrobacter diversus</td>
<td>54</td>
<td>3.00</td>
</tr>
<tr>
<td>Citrobacter freundii</td>
<td>187</td>
<td>4.23</td>
</tr>
<tr>
<td>Edwardsiella hoshinae</td>
<td>9</td>
<td>1.11</td>
</tr>
<tr>
<td>Edwardsiella icthuri</td>
<td>13</td>
<td>1.46</td>
</tr>
<tr>
<td>Edwardsiella tarda</td>
<td>35</td>
<td>0.97</td>
</tr>
<tr>
<td>Edwardsiella tarda 1</td>
<td>7</td>
<td>0.14</td>
</tr>
<tr>
<td>Enterobacter aerogenes</td>
<td>26</td>
<td>1.89</td>
</tr>
<tr>
<td>Enterobacter amnigenes 1</td>
<td>14</td>
<td>1.50</td>
</tr>
<tr>
<td>Enterobacter amnigenes 2</td>
<td>6</td>
<td>1.67</td>
</tr>
<tr>
<td>Enterobacter asburiae</td>
<td>45</td>
<td>2.42</td>
</tr>
<tr>
<td>Enterobacter cloacae</td>
<td>156</td>
<td>3.75</td>
</tr>
<tr>
<td>Enterobacter dissolvens</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>Enterobacter gergoviae</td>
<td>72</td>
<td>0.86</td>
</tr>
<tr>
<td>Enterobacter intermediun</td>
<td>10</td>
<td>0.60</td>
</tr>
<tr>
<td>Enterobacter nimpersurals</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>Enterobacter sakazakii</td>
<td>90</td>
<td>2.17</td>
</tr>
<tr>
<td>Enterobacter taylorae</td>
<td>98</td>
<td>1.16</td>
</tr>
<tr>
<td>Erucina ananas</td>
<td>5</td>
<td>3.00</td>
</tr>
<tr>
<td>Erucina chrysanthemi</td>
<td>2</td>
<td>2.00</td>
</tr>
<tr>
<td>Erucina cyprippedi</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>Erucina herbicola</td>
<td>15</td>
<td>5.67</td>
</tr>
<tr>
<td>Erucina milletiae</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>Erucina rhapontici</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>Erucina rubrifaciens</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>Erucina stevartii</td>
<td>3</td>
<td>1.00</td>
</tr>
<tr>
<td>Erucina uredovora</td>
<td>2</td>
<td>3.00</td>
</tr>
<tr>
<td>Escherichia blattae</td>
<td>3</td>
<td>1.33</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>1708</td>
<td>7.07</td>
</tr>
<tr>
<td>Escherichia fergusonii</td>
<td>40</td>
<td>3.80</td>
</tr>
<tr>
<td>Escherichia hermannii</td>
<td>68</td>
<td>3.62</td>
</tr>
<tr>
<td>Escherichia vulneris</td>
<td>77</td>
<td>3.51</td>
</tr>
<tr>
<td>Ewingella americana</td>
<td>87</td>
<td>5.35</td>
</tr>
<tr>
<td>Hafnia alvei</td>
<td>99</td>
<td>4.89</td>
</tr>
<tr>
<td>Hafnia alvei 1</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>Klebsiella oxytoca</td>
<td>90</td>
<td>1.00</td>
</tr>
<tr>
<td>Klebsiella ozaena</td>
<td>53</td>
<td>5.68</td>
</tr>
<tr>
<td>Klebsiella planticola</td>
<td>17</td>
<td>0.59</td>
</tr>
<tr>
<td>Klebsiella pneumonia</td>
<td>167</td>
<td>4.01</td>
</tr>
<tr>
<td>Klebsiella pneumonia 1</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>Klebsiella rhinoscleromatis</td>
<td>12</td>
<td>1.58</td>
</tr>
<tr>
<td>Klebsiella terrigena</td>
<td>9</td>
<td>1.56</td>
</tr>
</tbody>
</table>

Table 1. The CFARM classification of ENTE

Table 1. (cont.)

<table>
<thead>
<tr>
<th>Species</th>
<th>Size</th>
<th>Distortion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kluyvera cryocrescens</td>
<td>35</td>
<td>1.86</td>
</tr>
<tr>
<td>Kluyvera species 3</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>Koserella trabulsi</td>
<td>12</td>
<td>1.75</td>
</tr>
<tr>
<td>Leclercia adecarboxylata</td>
<td>26</td>
<td>2.85</td>
</tr>
<tr>
<td>Leminorella grimontii</td>
<td>4</td>
<td>1.75</td>
</tr>
<tr>
<td>Moellerella wisconsensis</td>
<td>17</td>
<td>2.59</td>
</tr>
<tr>
<td>Morganella morganii</td>
<td>66</td>
<td>3.23</td>
</tr>
<tr>
<td>Morganella morganii 1</td>
<td>12</td>
<td>2.33</td>
</tr>
<tr>
<td>Obesumbacterium proteus 1</td>
<td>3</td>
<td>2.33</td>
</tr>
<tr>
<td>Obesumbacterium proteus 2</td>
<td>3</td>
<td>0.67</td>
</tr>
<tr>
<td>Proteus mirabilis</td>
<td>41</td>
<td>4.20</td>
</tr>
<tr>
<td>Proteus myxofaciens</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>Proteus penneri</td>
<td>51</td>
<td>3.14</td>
</tr>
<tr>
<td>Proteus vulgaris</td>
<td>29</td>
<td>2.99</td>
</tr>
<tr>
<td>Providencia alcalifaciens</td>
<td>17</td>
<td>1.71</td>
</tr>
<tr>
<td>Providencia rettgeri</td>
<td>28</td>
<td>5.79</td>
</tr>
<tr>
<td>Providencia rustigianii</td>
<td>29</td>
<td>2.90</td>
</tr>
<tr>
<td>Providencia stuartii</td>
<td>67</td>
<td>4.34</td>
</tr>
<tr>
<td>Rahmella aquatilis</td>
<td>19</td>
<td>2.95</td>
</tr>
<tr>
<td>Salmonella arizonae</td>
<td>2</td>
<td>6.00</td>
</tr>
<tr>
<td>Salmonella choleraeis</td>
<td>22</td>
<td>2.96</td>
</tr>
<tr>
<td>Salmonella enteritidis</td>
<td>99</td>
<td>6.56</td>
</tr>
<tr>
<td>Salmonella pullorum</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>Salmonella typhimurium</td>
<td>8</td>
<td>2.75</td>
</tr>
<tr>
<td>Salmonella typhi</td>
<td>13</td>
<td>2.00</td>
</tr>
<tr>
<td>Serratia ficaria</td>
<td>15</td>
<td>2.87</td>
</tr>
<tr>
<td>Serratia fonticola</td>
<td>33</td>
<td>3.18</td>
</tr>
<tr>
<td>Serratia grimes</td>
<td>8</td>
<td>4.75</td>
</tr>
<tr>
<td>Serratia liquefaciens</td>
<td>6</td>
<td>2.17</td>
</tr>
<tr>
<td>Serratia marcescens</td>
<td>123</td>
<td>4.55</td>
</tr>
<tr>
<td>Serratia marcescens 1</td>
<td>22</td>
<td>5.60</td>
</tr>
<tr>
<td>Serratia odorifera 1</td>
<td>20</td>
<td>2.75</td>
</tr>
<tr>
<td>Serratia odorifera 2</td>
<td>28</td>
<td>2.86</td>
</tr>
<tr>
<td>Serratia plymuthica</td>
<td>16</td>
<td>4.44</td>
</tr>
<tr>
<td>Serratia proteamaculans 1</td>
<td>5</td>
<td>9.80</td>
</tr>
<tr>
<td>Serratia proteamaculans 2</td>
<td>5</td>
<td>4.20</td>
</tr>
<tr>
<td>Serratia rubidae</td>
<td>12</td>
<td>3.17</td>
</tr>
<tr>
<td>Shigella boydii</td>
<td>120</td>
<td>3.63</td>
</tr>
<tr>
<td>Shigella dysenteriae</td>
<td>57</td>
<td>3.67</td>
</tr>
<tr>
<td>Shigella flexneri</td>
<td>181</td>
<td>4.56</td>
</tr>
<tr>
<td>Shigella sonnei</td>
<td>72</td>
<td>3.88</td>
</tr>
<tr>
<td>Shigella prov</td>
<td>67</td>
<td>3.22</td>
</tr>
<tr>
<td>Tatumella pytseos</td>
<td>53</td>
<td>1.98</td>
</tr>
<tr>
<td>Xenorhabdus luminescens 5</td>
<td>5</td>
<td>2.20</td>
</tr>
<tr>
<td>Yersinia aldovae</td>
<td>5</td>
<td>3.80</td>
</tr>
<tr>
<td>Yersinia enterocolitica</td>
<td>207</td>
<td>4.59</td>
</tr>
<tr>
<td>Yersinia frederiksenii</td>
<td>30</td>
<td>3.43</td>
</tr>
<tr>
<td>Yersinia intermedia</td>
<td>35</td>
<td>4.80</td>
</tr>
<tr>
<td>Yersinia kristensenii</td>
<td>25</td>
<td>4.32</td>
</tr>
<tr>
<td>Yersinia pestis</td>
<td>2</td>
<td>1.50</td>
</tr>
<tr>
<td>Yersinia pseudotuberculosis</td>
<td>24</td>
<td>1.83</td>
</tr>
<tr>
<td>Yersinia rohdei</td>
<td>2</td>
<td>1.50</td>
</tr>
<tr>
<td>Yersinia ruckeri</td>
<td>4</td>
<td>2.75</td>
</tr>
</tbody>
</table>
Classification of Enterobacteriaceae

with the inclusion of a few odd items). (2) *Escherichia coli* is certainly also confirmed but it splits into eight pure (i.e. containing only *E. coli* strains) and four almost pure classes (containing in addition a few odd strains). Specimens labelled *E. coli* show affinity particularly to shigellas (but also salmonellas and yersinias). (3) Several of the genera, e.g. *Enterobacter*, *Klebsiella*, and especially *Serratia*, *Salmonella*, *Shigella* and *Yersinia*, are split into several pure and almost pure genus classes, which do not, however, always correspond to the species labels of the included specimens.

Adding or subtracting items to a classified collection may in principle alter the already established classification with respect to both its structure and the number of classes. A classification method of any use in practical taxonomy should of course be rather rigid with respect to adding new items. To test whether the SC-based classification method fulfils this requirement we performed two further classifications. One classification comprised all escherichias (including *Leclercia*), and all salmonellas, shigellas and yersinias in ENTE, 2910 strains altogether. This material is referred to as ESSY. The other classification comprised the 1708 *E. coli* labelled strains (material COLI).

### SC-minimizing classification of ESSY (SCESSY)

At the SC minimum (20.273) ESSY is divided into 26 classes. The results of SCCESSY are condensed into Table 3. The genera *Escherichia*, *Salmonella*, *Shigella* and *Yersinia* are briefly commented on below.

*Escherichia.* Four species are (almost) strictly confirmed: *E. fergusonii*, *E. hermanii*, *E. vulneris* and *Leclercia adecarboxylata*. Farmer et al. (1985) concluded that the last-mentioned species, although originally known as *Escherichia adecarboxylata*, does not belong to *Escherichia*. In all our runs *L. adecarboxylata* was strictly separated and attachment to other classes could not be found. *E. coli* produces six pure and five almost pure classes. These are discussed in more detail in connection with material COLI.

*Salmonella.* In SCENTE the salmonellas are divided among four classes (34, 40, 55 and 63). Of these, class 34 corresponds exactly to SCESSY class 18, as does class 40 to SCESSY class 17. The SCENTE classes 55 and 63 are combined in SCESSY classes 18 and 19. (By correspondence between a SCESSY class A and a SCENTE class B we mean that every element of A is also an element of B and that B does not contain other ESSY elements than those belonging to A.)

*Shigella.* In SCENTE shigellas occur in classes 2, 12, 14 and 23. These four classes have their identical counterparts in SCESSY classes 21, 10, 8 and 22, respectively.

*Yersinia.* For the yersinias as well there is a complete correspondence between SCENTE (classes 7, 77, 50 and 51) and SCESSY (classes 23, 24, 26 and 25, respectively). Summarizing the above observations, we conclude that as far as the genera *Salmonella*, *Shigella* and *Yersinia* are concerned, the SC-minimizing classification of the material ESSY gives rise to a classification identical with that resulting from minimizing SC of the considerably larger material ENTE. This shows that the SC-based classification method has a certain robustness with respect to altering the material.

### SC-minimizing classification of Escherichia coli (SCCOLI)

The SC minimum of the COLI material was 20.219 at 10 classes. The analysis of the 10 SCCOLI classes was performed by comparison of the distribution of the individual *E. coli* strains within each of the three optimal classifications (materials COLI, ESSY and ENTE, respectively). The results are presented in Table 4.

As can be seen from Table 4, there is an obvious correspondence between certain SCCOLI classes and classes of the SCESSY and SCENTE classifications. SCCOLI classes 1, 2, 3 and 4 are well concentrated (cf. also Table 5) and represent the bulk of *E. coli*(932/1708), corresponding well to the centroid and HMO of *E. coli* as defined by Farmer et al. (1985). These SCCOLI classes are also quite homogeneous (distortion figures <4.2). It is, therefore, easy to conclude that the ‘real’ *E. coli* is included within these SCCOLI classes.

Farmer et al. (1985) also considered an ‘inactive’ *E. coli.* The description given corresponds quite well to the SCCOLI class 6 (Table 5). This class (which contains 156/1708 *E. coli*) is also homogeneous (distortion figure: 3.763). The remaining *E. coli*-labelled specimens (620/1708) are distributed over five SCCOLI classes (5, 7, 8, 9 and 10). These classes are all heterogeneous (distortion figures >4.8). As shown in Table 5, the *E. coli*-labelled specimens related to shigellas are found particularly in the SCCOLI classes 5 and 8. *E. coli* related to *Yersinia pseudotuberculosis* are found in the SCCOLI classes 7 and 9. It thus seems that the heterogeneous SCCOLI classes 5, 7, 8, 9 and 10 contain *E. coli* which are related to other organisms (*Shigella*, *Y. pseudotuberculosis*, etc.) or are otherwise atypical specimens, perhaps also misidentified isolates.

### DISCUSSION

The purpose of the present study is to point out the applicability of a particular mathematical method (minimization of stochastic complexity) to classification of bacteria. As a specific example we have chosen phenotypic classification of *Enterobacteriaceae* based on biochemical tests. The SC classification considers, as does every numerical taxonomy approach, all characters (tests) as equally valid contributors of information and it thus takes into account the whole binary vector.
Table 2. The SC-minimizing classification of ENTE

Taxonomy refers to the classification given in Table 1.

<table>
<thead>
<tr>
<th>Class</th>
<th>Size</th>
<th>Distortion</th>
<th>Taxonomy</th>
<th>No. of strains in material</th>
<th>No. of strains in class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>264</td>
<td>3.54</td>
<td>E. coli</td>
<td>1708</td>
<td>264</td>
</tr>
<tr>
<td>2</td>
<td>247</td>
<td>3.80</td>
<td>E. coli</td>
<td>1708</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>120</td>
<td></td>
<td>S. boydii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>57</td>
<td></td>
<td>S. dysenteriae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>181</td>
<td></td>
<td>S. flexneri</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>67</td>
<td></td>
<td>S. prov</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>72</td>
<td></td>
<td>S. sonnei</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td></td>
<td>Y. kristensenii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>26</td>
<td></td>
<td>E. aerogenes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>90</td>
<td></td>
<td>K. oxytoca</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>17</td>
<td></td>
<td>K. planticola</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>167</td>
<td></td>
<td>K. pneumoniae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td></td>
<td>K. terrigena</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>31</td>
<td></td>
<td>E. amalonaticus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td></td>
<td>S. arizonae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>35</td>
<td></td>
<td>E. coli</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td></td>
<td>C. freundii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>112</td>
<td></td>
<td>Y. enterocolitica</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>30</td>
<td></td>
<td>Y. frederiksenii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td></td>
<td>Y. intermedia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>119</td>
<td></td>
<td>Y. kristensenii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td></td>
<td>Y. rohdei</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>57</td>
<td></td>
<td>E. coli</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>181</td>
<td></td>
<td>S. boydii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>159</td>
<td></td>
<td>S. flexneri</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td></td>
<td>S. prov</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>58</td>
<td></td>
<td>Y. kristensenii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>25</td>
<td></td>
<td>E. coli</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>112</td>
<td></td>
<td>Y. enterocolitica</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>30</td>
<td></td>
<td>Y. frederiksenii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>35</td>
<td></td>
<td>Y. intermedia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>25</td>
<td></td>
<td>Y. kristensenii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>5</td>
<td></td>
<td>E. fergusonii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>119</td>
<td></td>
<td>K. ascorbata</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>35</td>
<td></td>
<td>K. cryocrescens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>1</td>
<td></td>
<td>K. species 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>10</td>
<td></td>
<td>C. lapagei</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>14</td>
<td></td>
<td>E. amnigenus 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>35</td>
<td></td>
<td>E. cloacae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>1</td>
<td></td>
<td>E. dissolvens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>10</td>
<td></td>
<td>E. intermedium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>1</td>
<td></td>
<td>E. coli</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>139</td>
<td></td>
<td>E. coli</td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>14</td>
<td></td>
<td>S. flexneri</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>9</td>
<td></td>
<td>S. prov</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>1</td>
<td></td>
<td>E. rubrifaciens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>1</td>
<td></td>
<td>E. stewartii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>1</td>
<td></td>
<td>E. coli</td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>2</td>
<td></td>
<td>S. boydii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>124</td>
<td></td>
<td>S. flexneri</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>8</td>
<td></td>
<td>S. prov</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>1</td>
<td></td>
<td>Y. rohdei</td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>133</td>
<td></td>
<td>E. coli</td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>4</td>
<td></td>
<td>S. boydii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>124</td>
<td></td>
<td>S. flexneri</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>8</td>
<td></td>
<td>S. prov</td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>1</td>
<td></td>
<td>Y. rohdei</td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>66</td>
<td></td>
<td>E. coli</td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>9</td>
<td></td>
<td>S. boydii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>21</td>
<td></td>
<td>S. dysenteriae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>17</td>
<td></td>
<td>S. flexneri</td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>109</td>
<td></td>
<td>E. coli</td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>2</td>
<td></td>
<td>S. sonnei</td>
<td></td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>109</td>
<td></td>
<td>E. coli</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 2. (cont.)

<table>
<thead>
<tr>
<th>Class</th>
<th>Size</th>
<th>Distortion</th>
<th>Taxonomy</th>
<th>No. of strains in material</th>
<th>No. of strains in class</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>109</td>
<td>3.69</td>
<td>Y. aldoae</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Y. entercolitica</td>
<td>207</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Y. kristenseni</td>
<td>25</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Y. rohdei</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>108</td>
<td>1.58</td>
<td>E. amnigenus 2</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. cloaca</td>
<td>156</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. niimpressuralis</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. taylorae</td>
<td>98</td>
<td>98</td>
</tr>
<tr>
<td>19</td>
<td>98</td>
<td>3.68</td>
<td>S. marcescens 1</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. marcescens</td>
<td>123</td>
<td>97</td>
</tr>
<tr>
<td>20</td>
<td>98</td>
<td>4.85</td>
<td>M. morganii</td>
<td>66</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>P. alcalifaciens</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>P. rustigianii</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>P. stuartii</td>
<td>67</td>
<td>61</td>
</tr>
<tr>
<td>21</td>
<td>95</td>
<td>5.26</td>
<td>E. coli</td>
<td>1708</td>
<td>95</td>
</tr>
<tr>
<td>22</td>
<td>94</td>
<td>2.28</td>
<td>E. amnigenus 1</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. sakazakii</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>23</td>
<td>90</td>
<td>4.08</td>
<td>E. coli</td>
<td>1708</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. enteritidis</td>
<td>99</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. boydii</td>
<td>120</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. flexneri</td>
<td>181</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. sonnei</td>
<td>72</td>
<td>68</td>
</tr>
<tr>
<td>24</td>
<td>89</td>
<td>4.91</td>
<td>H. alvei</td>
<td>99</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>K. pneumoniae</td>
<td>167</td>
<td>3</td>
</tr>
<tr>
<td>25</td>
<td>83</td>
<td>3.84</td>
<td>M. morganii 1</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>M. morganii</td>
<td>66</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>P. rettgeri</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>X. luminescens 5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>26</td>
<td>81</td>
<td>3.68</td>
<td>C. amalonaticus 1</td>
<td>40</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C. amalonaticus</td>
<td>31</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C. freundii</td>
<td>187</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. coli</td>
<td>1708</td>
<td>2</td>
</tr>
<tr>
<td>27</td>
<td>78</td>
<td>3.60</td>
<td>P. penneri</td>
<td>51</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>P. vulgaris</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>28</td>
<td>77</td>
<td>3.51</td>
<td>E. vulneris</td>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td>29</td>
<td>72</td>
<td>0.86</td>
<td>E. gergoviae</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>30</td>
<td>70</td>
<td>3.36</td>
<td>E. asburiae</td>
<td>45</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. cloaca</td>
<td>156</td>
<td>26</td>
</tr>
<tr>
<td>31</td>
<td>70</td>
<td>6.00</td>
<td>E. intermedium</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. rhapontici</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>K. oxytoca</td>
<td>90</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>K. ozaenae</td>
<td>53</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>K. pneumoniae 1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>K. pneumoniae</td>
<td>167</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>K. rhinoscleromatis</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>K. terrigena</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>K. cryocrescens</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>68</td>
<td>3.61</td>
<td>E. hermanii</td>
<td>68</td>
<td>68</td>
</tr>
<tr>
<td>33</td>
<td>65</td>
<td>5.95</td>
<td>E. coli</td>
<td>1708</td>
<td>65</td>
</tr>
<tr>
<td>34</td>
<td>64</td>
<td>3.25</td>
<td>C. freundii</td>
<td>187</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. enteritidis</td>
<td>99</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. typhimurium</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>35</td>
<td>57</td>
<td>3.39</td>
<td>E. americana</td>
<td>87</td>
<td>57</td>
</tr>
<tr>
<td>36</td>
<td>55</td>
<td>3.18</td>
<td>C. amalonaticus 1</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. ictaluri</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. tarda</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. coli</td>
<td>1708</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. americana</td>
<td>87</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H. alvei 1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O. proteus 2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Classification of Enterobacteriaceae
### Table 2. (cont.)

<table>
<thead>
<tr>
<th>Class</th>
<th>Size</th>
<th>Distortion</th>
<th>Taxonomy</th>
<th>No. of strains in material</th>
<th>No. of strains in class</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>55</td>
<td>3.07</td>
<td>C. diversus</td>
<td>54</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. fergusonii</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>38</td>
<td>54</td>
<td>5.59</td>
<td>K. ozzaeae</td>
<td>53</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>K. pneumoniae</td>
<td>167</td>
<td>7</td>
</tr>
<tr>
<td>39</td>
<td>53</td>
<td>1.57</td>
<td>P. stuartii</td>
<td>67</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T. ptyseos</td>
<td>53</td>
<td>52</td>
</tr>
<tr>
<td>40</td>
<td>52</td>
<td>5.65</td>
<td>E. coli</td>
<td>1708</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. choleraesuis</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. enteritidis</td>
<td>99</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. typhi</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Y. ruckeri</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>41</td>
<td>50</td>
<td>4.58</td>
<td>E. coli</td>
<td>1708</td>
<td>50</td>
</tr>
<tr>
<td>42</td>
<td>49</td>
<td>5.40</td>
<td>E. coli</td>
<td>1708</td>
<td>49</td>
</tr>
<tr>
<td>43</td>
<td>47</td>
<td>5.96</td>
<td>S. marcescens 1</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. marcescens</td>
<td>123</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. proteamaculans 1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>44</td>
<td>46</td>
<td>3.52</td>
<td>C. daviseae</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C. lapagei</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C. neteri</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C. species 3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C. species 5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>45</td>
<td>44</td>
<td>4.27</td>
<td>P. mirabilis</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>P. myxofaciens</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>P. penneri</td>
<td>51</td>
<td>2</td>
</tr>
<tr>
<td>46</td>
<td>40</td>
<td>3.30</td>
<td>E. aerogenes</td>
<td>26</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. fonticola</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>47</td>
<td>39</td>
<td>5.41</td>
<td>S. arizonae</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. grimesii</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. liquefaciens</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. odorifera 1</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. proteamaculans 1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. proteamaculans 2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>48</td>
<td>39</td>
<td>3.51</td>
<td>E. coli</td>
<td>1708</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. fergusonii</td>
<td>40</td>
<td>38</td>
</tr>
<tr>
<td>49</td>
<td>38</td>
<td>4.71</td>
<td>S. odorifera 1</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. odorifera 2</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. rubidae</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>50</td>
<td>35</td>
<td>3.20</td>
<td>E. herbicola</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. coli</td>
<td>1708</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Y. pestis</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Y. pseudotuberculosis</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>51</td>
<td>35</td>
<td>4.69</td>
<td>C. amalonaticus</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>K. cryocrescens</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Y. intermedia</td>
<td>35</td>
<td>33</td>
</tr>
<tr>
<td>52</td>
<td>35</td>
<td>5.51</td>
<td>S. ficaria</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. plymuthica</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. proteamaculans 1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. rubidae</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>53</td>
<td>30</td>
<td>4.80</td>
<td>P. rettgeri</td>
<td>28</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>P. stuartii</td>
<td>67</td>
<td>5</td>
</tr>
<tr>
<td>54</td>
<td>30</td>
<td>5.17</td>
<td>E. americana</td>
<td>87</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. marcescens 1</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. proteamaculans 1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Y. aldovae</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>55</td>
<td>28</td>
<td>3.04</td>
<td>E. blattae</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E. coli</td>
<td>1708</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. enteritidis</td>
<td>99</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. pullorum</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. sonnei</td>
<td>72</td>
<td>1</td>
</tr>
</tbody>
</table>
### Table 2. (cont.)

<table>
<thead>
<tr>
<th>Class</th>
<th>No. of strains in material</th>
<th>Class</th>
<th>No. of strains in material</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>26</td>
<td>57</td>
<td>26</td>
</tr>
<tr>
<td>58</td>
<td>21</td>
<td>59</td>
<td>19</td>
</tr>
<tr>
<td>60</td>
<td>19</td>
<td>61</td>
<td>15</td>
</tr>
<tr>
<td>62</td>
<td>13</td>
<td>63</td>
<td>12</td>
</tr>
<tr>
<td>64</td>
<td>10</td>
<td>65</td>
<td>10</td>
</tr>
<tr>
<td>66</td>
<td>3</td>
<td>67</td>
<td>2</td>
</tr>
<tr>
<td>68</td>
<td>1</td>
<td>69</td>
<td>1</td>
</tr>
</tbody>
</table>

### Table 3. Distribution of strains over the 26 classes of the SC-minimizing classification of ESSY and the species in the classification given in Table 1

The entries are absolute numbers. Dist., distortion.
Table 4. The distribution of E. coli-labelled strains over the classes of SCCOLI and the pure and almost pure E. coli classes of SCESSY and SCENTE

The entries are absolute numbers. Dist., distortion.

<table>
<thead>
<tr>
<th>SCCOLI</th>
<th>No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size...</td>
<td>259</td>
<td>240</td>
<td>233</td>
<td>200</td>
<td>188</td>
<td>156</td>
<td>138</td>
<td>131</td>
<td>108</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>Dist...</td>
<td>3.52</td>
<td>4.19</td>
<td>3.66</td>
<td>3.50</td>
<td>6.01</td>
<td>3.76</td>
<td>5.82</td>
<td>6.08</td>
<td>4.82</td>
<td>5.95</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SCESSY</th>
<th>No.</th>
<th>Size</th>
<th>Dist.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>265</td>
<td>3.57</td>
<td>254</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>222</td>
<td>4.00</td>
<td>220</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>216</td>
<td>3.39</td>
<td>215</td>
<td>2</td>
<td>181</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>186</td>
<td>3.45</td>
<td>2</td>
<td>117</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>149</td>
<td>5.02</td>
<td>2</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>141</td>
<td>3.69</td>
<td>2</td>
<td>133</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>131</td>
<td>5.71</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>53</td>
<td>4.75</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>101</td>
<td>4.63</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>45</td>
<td>5.53</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>56</td>
<td>6.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SCENTE</th>
<th>No.</th>
<th>Size</th>
<th>Dist.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>264</td>
<td>3.54</td>
<td>248</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>228</td>
<td>4.04</td>
<td>225</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>212</td>
<td>3.37</td>
<td>210</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>146</td>
<td>2.84</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>138</td>
<td>5.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>141</td>
<td>3.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>111</td>
<td>5.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>50</td>
<td>4.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>109</td>
<td>4.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>49</td>
<td>5.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>65</td>
<td>5.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>95</td>
<td>5.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

representing the various strains; it does not consider divisions produced by individual attributes, however interesting they may be.

We emphasize that the SC-minimizing classification maximizes the information content of the classification and thus provides the ‘best’ classification in the sense of information theory, which does not imply that the classification would be the most relevant from the biological point of view.

The results of the present study have to be evaluated within the limits given by the bacterial material involved and the data by which this material is defined. As to the bacterial material, one may argue that it was skewed since almost a third of the strains in the study on Enterobacteriaceae were labelled Escherichia coli. On the other hand, as Farmer et al. (1985) emphasized, the proportion of E. coli is even higher in clinical practice. Taxonomy is not dependent on clinical practice, but clinical practice is dependent on a taxonomy, which is relevant with respect to the isolates it deals with, or as Staley & Krieg (1984) put it: ‘bacterial classifications are devised for microbiologists, not for the entities being classified’.

Another point worth discussion is the means to characterize Enterobacteriaceae. Over time new traits have been introduced and this trend is reflected in the steady increase in the number of genera and species of Enterobacteriaceae. This also concerns the test battery for phenetic description of Enterobacteriaceae which provided the data for the present study. However, no such change can be discerned during the period over which the material considered in this paper was collected. The material contained strains isolated during 20 years (1969–1988) and the strains of different years were distributed uniformly over the respective classes in all the classifications obtained by SC-minimization.

The structure of the SC-based classification of Enterobacteriaceae was very rigid with respect to altering the material and corresponded well to the general taxo-
earlier thoughts. The SC classification applied to a set of
However, new approaches have to be checked against
Providencia, name-labelled specimens compares taxospecies (or
classes. SC classification is thus neither a splitter nor a
agreement with the outcomes of other approaches.
The SC-based classification has been compared with
methods. However, some differences were obvious. E.
coli, the key species of Enterobacteriaceae to which all
other members of the family have always been com-
pared, was split into several entities by the SC-based
classification. In contrast, other species of Escherichia
(E. fergusonii, E. hermannii, E. vulneris) were distinctly
separable and definable. In other entities of Enterobac-
teriaeae, within genera such as Proteus and Providencia, known species were combined into joint
classes. SC classification is thus neither a splitter nor a
lumper: it both splits and lumps.

The SC-based classification has been compared with
present, generally accepted classifications of Enterobac-
teriaeae. This does not mean that the authors seek
some kind of authorization for their method from agreement with the outcomes of other approaches.
There are no ‘official classifications’ in microbiology.
However, new approaches have to be checked against
earlier thoughts. The SC classification applied to a set
of name-labelled specimens compares taxospecies (or
taxocharacters) with nomenclature species [cf. Sneath, 1984]. This
is the usual approach of numerical taxonomy. Sneath
(1995b) concluded that results of numerical taxonomy have not necessarily confirmed earlier views. The
present paper comes close to confirmation of earlier
views, but may indicate some further problems requiring
attention.

Vandamme et al. (1996) reviewed the various
to classification of bacteria, and illustratively concluded that a combination of the results
from all of them (polyphasic taxonomy) is to be
recommended. SC-based classification may provide a
further tool for polyphasic taxonomy. It also provides a
useful tool for comparing the information content of
alternative classifications of common materials. The SC
method is well suited for the treatment of large (and thus statistically representative) materials. Finally, as already
emphasized in the Introduction, the SC-based classi-
fication sets no preconditions. There is no predefined
level of similarity that an item has to fulfil in order to be
included as a member of a class.

The amount of comparative molecular sequence data is
increasing at an exponential rate and as a consequence
the importance of phenotypic numerical taxonomy
based on biochemical tests is declining. If the SC method
were applicable only to classifying items characterized by
their phenotype its value for bacterial taxonomy
would be rather meagre. From a mathematical point of
view minimizing SC is a method for classifying (binary)
Vectors (Gyllenberg et al., 1994b). Since the molecular
sequence data of a bacterium can be given a binary code
equivalent to the standard code using the four-letter alphabet ACCT, the SC approach can, at least in
principle, be applied to molecular sequence data. It
should, however, be noted that the statistical model (2)
of the classification based on phenotypic data would
most probably be a poor model for a classification based
on genotypic data. It is an important and challenging
task to construct a relevant model for classifications
based on molecular sequence data. Once this has been
done the SC method presented in this paper can
immediately be applied.

Acknowledgements
The authors express their sincerest thanks to Dr J. J. Farmer
III for the permission to use the data material on which this
study was based.

References
Balows, A., Trüper, H. G., Dworkin, M., Harder, W. & Schleifer,
Verlag.
Brenner, D. J. (1992). Introduction to the family Enterobac-
by A. Balows, H. G. Trüper, M. Dworkin, W. Harder & K.-H.
authors (1985). Biochemical identification of new species and
biogroups of Enterobacteriaceae isolated from clinical specimens.
J Clin Microbiol 21, 46–76.
30, 643–654.
Non-uniqueness of numerical taxonomic structures. Binary 5,
138–144.


Sneath, P. H. A. (1979). BASIC program for determining the best identification scores possible from the most typical examples when compared with an identification matrix of percent positive characteristics. Comput Geosci 6, 27–34.


Received 15 March 1996; revised 25 September 1996; accepted 21 October 1996.