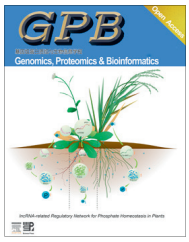
Genomics Proteomics Bioinformatics 13 (2015) 192–199

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



APPLICATION NOTE

BioCluster: Tool for Identification and Clustering of Enterobacteriaceae Based on Biochemical Data

Ahmed Abdullah [a], S.M. Sabbir Alam [b], Munawar Sultana [c], M. Anwar Hossain [*,d]

Department of Microbiology, University of Dhaka, Dhaka 1000, Bangladesh

Received 24 November 2014; revised 11 February 2015; accepted 10 March 2015
Available online 26 July 2015

Handled by Wuju Li

KEYWORDS

Bacterial identification;
Enterobacteriaceae;
Biochemical properties;
Clustering tool;
Identification tool;
Hierarchy algorithm

Abstract Presumptive identification of different **Enterobacteriaceae** species is routinely achieved based on **biochemical properties**. Traditional practice includes manual comparison of each biochemical property of the unknown sample with known reference samples and inference of its identity based on the maximum similarity pattern with the known samples. This process is labor-intensive, time-consuming, error-prone, and subjective. Therefore, automation of sorting and similarity in calculation would be advantageous. Here we present a MATLAB-based graphical user interface (GUI) tool named BioCluster. This tool was designed for automated clustering and identification of **Enterobacteriaceae** based on biochemical test results. In this tool, we used two types of algorithms, *i.e.*, traditional hierarchical clustering (HC) and the Improved Hierarchical Clustering (IHC), a modified algorithm that was developed specifically for the clustering and identification of **Enterobacteriaceae** species. IHC takes into account the variability in result of 1–47 biochemical tests within this **Enterobacteriaceae** family. This tool also provides different options to optimize the clustering in a user-friendly way. Using computer-generated synthetic data and some real data, we have demonstrated that BioCluster has high accuracy in clustering and identifying enterobacterial species based on biochemical test data. This tool can be freely downloaded at http://microbialgen.du.ac.bd/biocluster/.

Introduction

Enterobacteriaceae are a family of gram-negative, rod-shaped, facultative anaerobic bacteria, which are mainly recognized for their ability to cause intestinal diseases [1]. Enterobacteriaceae are responsible for a variety of human and animal illnesses, including urinary tract infections, gastroenteritis, meningitis, pneumonia, and septicemia [2,3]. Microbiological diagnosis for detecting the presence and type of Enterobacteriaceae from clinical samples is potentially important. Various biochemical

* Corresponding author.
E-mail: hossaina@du.ac.bd (Hossain MA).
[a] ORCID: 0000-0001-5550-9574.
[b] ORCID: 0000-0002-9148-3727.
[c] ORCID: 0000-0002-8563-3661.
[d] ORCID: 0000-0001-9777-0332.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2015.03.007

tests are traditionally used for presumptive identification and clustering of different enterobacterial species [4,5]. Biochemical tests such as indole production test, methyl red test, Voges-Proskauer test, citrate utilization etc. are usually performed [1,4–6]. Results of different tests are manually evaluated either as positive or negative for identification of a particular group of bacteria [2].

Manual check and comparison of biochemical test results are cumbersome and the results are sometimes hard to interpret, especially if the number of isolates is large. When there are a large volume of isolates, it becomes error-prone and difficult to reproduce, which is further confounded by the fact that the test result for a given species is not completely fixed: a given species may provide several combinations of biochemical test results [1].

By automating the analysis process of biochemical results, sorting (clustering), and identification of particular genera, the difficulties associated with manual sorting could be resolved. Here we propose a MATLAB-based tool, which was specifically designed for the clustering and identification of 128 species of Enterobacteriaceae from 30 genera based on the results of different biochemical tests (1–47 selected tests in Bergey's Manual of Systematic Bacteriology, [1]). We used two types of algorithms. One is the agglomerative hierarchical clustering algorithm (HC) and the other is a modified hierarchical clustering algorithm, which we termed as Improved Hierarchical Clustering algorithm (IHC). Agglomerative HC is a "bottom up" approach. Each observation starts in its own cluster, and pairs of clusters are merged as one to move up along the hierarchy [7,8]. Using BioCluster, HC can be applied directly to cluster Enterobacteriaceae isolates based on the biochemical properties. However, HC-based clustering may provide a misleading result due to the variability of the test results present within the same species. For example, closely-related *Escherichia* spp. can be sorted into different clusters, whereas different *Salmonella* spp. can be clustered with *Escherichia* spp. Therefore the algorithm was improved to take into account the variability of test results in Enterobacteriaceae by maximal utilization of relevant biochemical information for isolate clustering. We tested the accuracy of the new algorithm using computer-generated synthetic data and some real data, and it showed improved performance as compared to the naive HC algorithm.

BioCluster provides a user-friendly, easy method for the rapid clustering and identification of Enterobacteriaceae species based on biochemical properties. This tool is freely available for non-profit use at: http://microbialgen.du.ac.bd/biocluster/.

Methods

Algorithm

BioCluster uses HC as the clustering algorithm in two different ways. In one case, HC is directly applied to cluster the biochemical test results. However, biochemical test results are not numerical but are categorical (with binomial output as + or − for a given test). Hamming distance was thus chosen to measure the distance among different isolates, since it only considers the identity or non-identity of a test at a given position but not the actual numerical distance [9,10].

The clustering is further improved in IHC to provide a more refined output of biochemical test data. Species/isolates of a given genus show different levels of variability in their biochemical test results. For a given species, every test result may not be equally informative in clustering. For instance, the frequency for the *Edwardsiella hoshinae* isolates to produce a positive test result for indole production is about 50% [11]. So, the indole production test does not provide useful information for *E. hoshinae* identification/exclusion. As a result, the biochemical test results weight differentially when classifying a certain species.

Bergey's Manual of Systematic Bacteriology is a systematic catalog that contains information on the variability of the biochemical test results of a particular species [1]. Data tables for the frequency of positive biochemical test results for possible species of Enterobacteriaceae were taken from Bergey's Manual (Table S1) [2]. For IHC, the frequency table for positive results of biochemical tests (1–47 tests) is converted to conditional probability score matrixes for 128 Enterobacteriaceae species in 30 genera.

Naïve Bayesian model was used to find the probability of different instances of test results belonging to the set of 128 Enterobacteriaceae species [10,12,13]. It is assumed that the isolate belongs to one of the 128 members. If an isolate (*e.g.*, isolate 1) has the biochemical result T (T is a string of result, *e.g.*, $T = + - + + + \cdots -$), then probability score for species S_i is given by Bayesian probability as

$$P(S_i|T) = \frac{P(S_i)P(T|S_i)}{\sum_{j=1}^{n} P(S_j)P(T|S_j)} \quad (1)$$

Prior probability of being in one or other species is equal, which means:

$$P(S_1) = P(S_2) = P(S_3) = \ldots P(S_k) = \frac{1}{n} = 1/128$$

$$P(S_i|T) = \frac{\frac{1}{n}P(T|S_i)}{\frac{1}{n}\sum_{j=1}^{n} P(T|S_j)} = \frac{P(T|S_i)}{\sum_{j=1}^{n} P(T|S_j)} \quad (2)$$

There are t tests and test results are independent from each other according to the naivety assumption:

$$P(T|S_i) = \Pi_{j=1}^{t} Q_j \quad (3)$$

where Q_j stands for the probability of the Species S_i to show the same result for jth test as in T and n is the total number of species (128 in this case).

The choice of an appropriate distance metric is crucial for multidimensional clustering analysis. It is not always obvious what the distance metric means for a particular situation. In this study, we have chosen a distance metric, which we considered as one of the best possible solutions in the context, for the distance between the two isolates is defined as the probability that they belong to different species. It can be easily calculated from the conditional probability matrix. If C_i and C_j stands for conditional probability vector of ith and jth isolates, the distance is:

$$D(i,j) = 1 - C_i^T.C_j \quad (4)$$

For future reference, we will call this distance metric the Bayesian probability distance (BPD), which is a special case for canonical distance measure (CDM) developed by Baxter [14]. Although BPD is symmetrical and non-negative, it does not follow the identity indiscernible axiom. $D(x,y)\neq 0$ when $x = y$; and also, there exist x, y for which $D(x,y) < D(x,y)$. In other words, it is not self-minimal. Since this distance can be mathematically shown to be optimum, metric axioms have no special status for 1-nearest-neighbor classification [7,15]. In BPD, distance is expressed as probabilities.

Finally, using the distance measured from the preceding steps, a dendrogram is constructed using the HC algorithm. A cutoff value is chosen to divide the dendrogram into distinct clusters. A dendrogram potentially contains more information than simply clusters, such as how clusters (which are normally species) themselves form higher-level clusters (*i.e.*, genus or higher hierarchy). Unfortunately BPD metric is not useful for this purpose. For BPD, in addition to probability distribution of tests we need to know the actual phylogeny of the species. We have taken the phylogenetic tree generated from 16S rRNA gene as the model tree for this purpose. Although BPD measure is optimum for forming the lowest level cluster (*i.e.*, species level), it could not establish the phylogenetic relationship of the isolates examined. So we have employed another distance metric that captures this aspect, by using cost function.

$$d(x,y) = \sum_{(S_i,S_j)} R(S_i|S_j)P(S_i|x)P(S_j|y) \quad (5)$$

where $S_i \neq S_j$

$(R(S_i|S_j))$ is cost for miss-classifying species S_i for species S_j and $P(S_i|x)$ is the conditional probability. BPD can be considered as a special case of cost function (formulation can be found in [7]), where $(R(S_i|S_j))$ is the same for all pairs. In this work $R(S_i|S_j)$ represents the phylogenetic distance between species S_i and S_j, which is the percent dissimilarity between the pair after multiple sequence alignment (MSA) of 16S rRNA gene sequence of the corresponding species. BPD gives better clustering accuracy, whereas cost function captures the phylogenetic relatedness of the clusters better.

Implementation and availability

All algorithmic tools in BioCluster are implemented in MATLAB. This software can be accessed easily from a user-friendly graphical user interface (GUI). BioCluster is available as an executable file that runs on Windows XP/Vista/7/8 (32/64 bit), provided that the MATLAB Compiler 7.17 is installed. This tool is licensed under GNU GPL version 3.0 and does not have any restrictions to use by non-academics.

Software architecture and data input/output

In BioCluster, three different tasks are available: traditional HC, IHC, and identification. They are located on the upper panel of the GUI. Within the BioCluster, data are arranged in such a manner that columns correspond to tests and rows correspond to isolates. So each row corresponds to test results of a particular bacterial isolate. Positive test results are denoted as 1 and negative test results are denoted as 0. An example of input can be:

1100000101100000 1001
1101000101101111 0001
1100011001101001 0001
1101011000101001 1001
0011000000010111111nan

Here, 1 means positive result, 0 means negative result, and nan means absence of a test result. The input data have to be pasted into the text-box labeled input "Biochemical Data". For HC, test number is not essential as all tests are considered with equal weight. However, for IHC it is mandatory to assign the sequential number for each test. This can be done by selecting biochemical tests serially from the upper left panel of the IHC window, or by entering test numbers from 1 to 47 manually in the "sequence of selected test" window.

The name of the isolates can be provided as "enter" delimitated (one under other). This is optional, if no names are provided, then the numbered labeling would be used in the dendrogram. The number corresponds to the sequential position (row number) of the isolates in the data. For colored partitioning of the dendrogram, a cutoff value of the distance measure can be provided with the default cutoff value as 0.3. A dataset that is expected to form a single cluster can also be pasted in the same box, from which the cutoff value will be computed.

Unlike HC, in IHC the tests used should be specified, so as the order that they appear in the data. There is a list box under "Select test", from which the appropriate tests can be selected. The order in which these tests appear in the original data set has to be specified within the "Sequence of the selected test" box. For example, "3 1 2 5 4" (without quotation) indicate that the 3rd test from the list is the 1st test in the data, the 1st test appeared 2nd in the data and so forth. The list of all 47 tests can be found in Table S1. Then the name of the isolates can be typed (optional) and the cutoff value for cluster-formation can be designated or by default (0.3). Finally pressing the "Make Dendrogram" button will create a dendrogram. One of two distance metrics, either the "BPD" or "cost function" can be selected from a drop-down menu with the former set as default. BPD gives better accuracy in clustering but doesn't reflect phylogenetic relationship. For a better interpretation of clustering results, tests with both BPD and cost-function should be performed to compare. Clicking on the "Show in Enterobacteriaceae tree" button shows where the isolates stand in the population of the randomly selected 128 Enterobacteriaceae species. In both HC and IHC, there is a button called "show in new window". By clicking on this

button, the dendrogram will be shown in a fresh new window and can be saved in the new window as well.

The function of the identification tab is to view the Bayesian probability score for different species. The tool has to be provided with Biochemical data, Isolate names, Tests, Test sequence as input for bacterial identification. Users can choose the desired number of the highest scoring target species to list and three best results are shown by default in a descending order.

Results

Performance on simulated data

To synthetically generate data, a test number t (number of the test to be used, maximum is 47) was chosen. Then t number of tests was randomly chosen from the whole test set without replacement. Then five species were chosen from 128 species set without replacement. Bergey's Manual was taken as gold standard for synthetic data generation. For each species, five samples were computationally generated with random number generator (RNG) with the probability given by Bergey's manual [11]. These five sets of isolates from five species were defined as "Original cluster set". Two dendrograms were constructed, one using IHC and other using HC. Clustering accuracy (CA) was computed for both dendrograms.

The clustering accuracy is defined as

$$\mathrm{CA} = \frac{\text{Number of clusters correctly identified}}{\text{Total number of original clusters}} \tag{6}$$

We say a cluster is correctly identified when there is at least one node that contains all the elements from an original cluster set but none out of that set.

CA will obviously be different depending on the test amount, and higher CA would be achieved when more tests are used. It will also be related to choice of the tests and species. Here we compared CA of IHC and HC in terms of number of tests used. For each test number, CA was calculated 4000 times for different randomly-chosen combinations of tests and species, and then the average was taken.

The mean clustering accuracy was calculated as:

$$\overline{\mathrm{CA}}_n = \frac{1}{40 \times 100} \sum_{j=1}^{40} \sum_{k=1}^{100} \mathrm{CA}_n(\mathrm{TC}_j, \mathrm{SC}_k) \tag{7}$$

$\overline{\mathrm{CA}}$ is the mean CA and n is the number of tests used. TC and SC stand for particular combinations of tests and species used, respectively.

Different linkage criteria of clustering, *e.g.*, single, complete, average, weighted, ward, and centroid, were tested for both algorithms. The complete linkage criterion was found to be the optimum for both algorithms, which was presented in **Figure 1**. Moreover, IHC offers better CA than HC for all tests and combinations generally, which is not due to species bias as we have randomly selected species and the results are presented as average of random iterations.

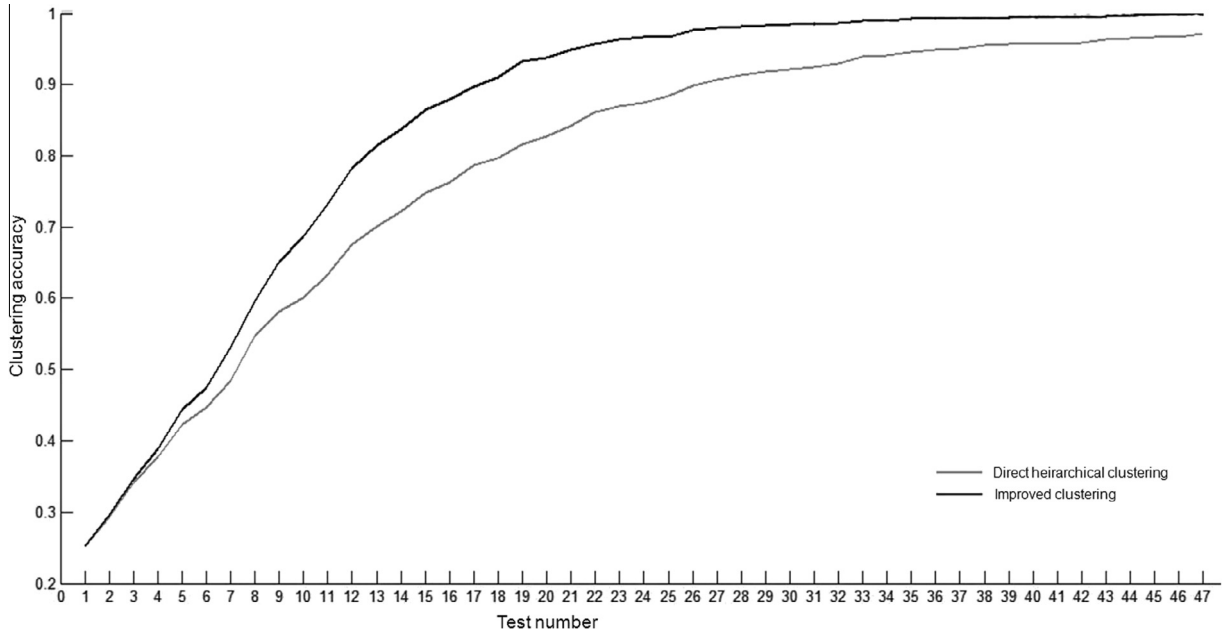


Figure 1 Comparison of clustering accuracy of DHC and IHC

X-axis denotes the test number and *Y*-axis denotes clustering accuracy. Dark line indicates clustering accuracy of Improved Hierarchical Clustering (IHC) and gray line indicates clustering accuracy of direct hierarchical clustering (DHC). Clustering accuracy was calculated 4000 times for different randomly-chosen combinations of tests and species, and then the average was taken (Equation (7)). Random combinations of tests were iterated 40 times and random combinations of species were iterated 100 times.

Accuracy of BPD compared to other distance metric

There are other distance metrics that are quite close to BPD, *e.g.*, Peterson' distance and cosine distance [10,13]. Cosine distance between two observations O_i and O_j is given by

$$D(O_i, O_j) = 1 - \frac{O_i.O_j}{|O_i||O_j|} \tag{8}$$

Peterson distance between two observations O_i and O_j is given by

$$D(O_i, O_j) = 1 - \frac{cov(O_i, O_j)}{S_{O_i} S_{O_j}} \tag{9}$$

In the context of our work the observations are conditional probability vectors as in Equation (4). From the data generated with the algorithm described in the preceding subsection, we compared the performance of Cosine and Peterson distances with BPD. Euclidean distance was also included for comparison, as it is the most natural choice of a distance metric. It should be noted that, these distance metrics are applied on the conditional probability matrix. **Table 1** shows the relative percentage difference (RPD) of CA, which was calculated by BPD over the three distance metrics aforementioned. Positive values shown in Table 1 represent a higher accuracy for BPD than the corresponding distance metrics. The BPD metric scored the highest accuracy for all test numbers except test numbers 1 and 2. The most commonly used distance metric, Euclidean distance scored poorly compared to other distance metrics. The accuracy of BPD, Peterson, and Cosine distance were fairly close, though BPD did better in most of the cases. These distances are mathematically very similar, which may explain comparable CA using these algorithms. When the magnitude of the observation vectors is equal to 1, Cosine distance and BPD become the same. When mean of the observation vector is equal to 0, the Peterson distance becomes equal to Cosine distance. Mathematically, Cosine distance is closer to BPD than Peterson. And as expected, Cosine distance does better than Peterson distance in a majority of the cases.

Comparison between BPD and cost function distance metrics

Dendrograms were constructed using the two distance metrics, BPD and cost function. We only took a dendrogram that has correctly identified all original cluster sets. The topology of the dendrogram was compared with its corresponding dendrogram constructed from 16S DNA sequences. Topological accuracy (TA) is defined as:

$$\text{TA} = \frac{\text{Number of times topology is perfectly matched}}{\text{Total number of iterations}} \tag{10}$$

TA can be viewed as the proportion of 4000 repetitions, in which topology is correctly identified. Only a perfect match of topology is counted while partial match is ignored. Like CA for each test number, TA was calculated 4000 times for different randomly-chosen combination of tests and species, and then the average was taken.

Table 1 Relative percentage difference of CA with different distance metrics compared to BPD in random test combinations

Test number	Peterson (%)	Cosine (%)	Euclidean (%)	Test number	Peterson (%)	Cosine (%)	Euclidean (%)	Test number	Peterson (%)	Cosine (%)	Euclidean (%)
1	0.00	0.00	0.00	17	1.62	0.26	29.85	33	14.14	6.06	149.03
2	−0.07	0.15	2.59	18	1.52	0.51	27.25	34	17.17	10.10	144.43
3	0.14	0.42	5.90	19	1.22	0.31	22.08	35	8.57	1.51	86.66
4	0.31	0.32	11.42	20	1.17	0.36	19.31	36	12.09	5.54	65.46
5	0.93	1.29	19.35	21	0.98	0.26	15.99	37	16.10	9.06	74.47
6	1.50	1.54	24.77	22	0.84	0.25	12.99	38	11.58	6.04	49.83
7	2.17	1.17	32.81	23	0.59	0.13	10.02	39	6.54	3.02	42.76
8	1.94	0.87	36.57	24	0.61	0.21	9.09	40	10.05	5.02	30.15
9	2.33	0.59	39.48	25	0.64	0.24	9.04	41	4.02	2.51	27.14
10	2.58	1.00	42.33	26	0.42	0.11	6.14	42	4.52	2.51	22.11
11	2.61	0.72	43.08	27	0.38	0.11	4.81	43	7.03	6.53	20.08
12	2.71	0.74	40.97	28	0.36	0.14	4.23	44	7.03	4.52	17.06
13	2.18	0.58	38.37	29	0.23	0.09	3.14	45	4.51	3.51	13.54
14	2.33	0.64	37.56	30	0.27	0.07	2.91	46	4.01	3.01	12.54
15	1.88	0.34	34.15	31	0.22	0.11	2.62	47	4.01	3.01	11.54
16	1.87	0.35	31.73	32	0.21	0.08	1.96				

Note: Relative percentage difference (RPD) of x with respect to $y = \frac{(y-x)}{(x+y)/2} \times 100\%$. Here x represents the CA of Cosine, Peterson, or Euclidean distance metrics and y represents the CA of BPD. RPD is calculated for different combinations of 47 tests. For a given number of tests, *e.g.*, 10, 10 tests are randomly selected from the 47 available tests and CA was calculated. This process is carried out 4000 times and the average RPD is presented here. CA, clustering accuracy; BPD, Bayesian probability distance.

The mean topological accuracy was calculated as:

$$\overline{\mathrm{TA}_n} = \frac{1}{40 \times 100} \sum_{j=1}^{40} \sum_{k=1}^{100} \mathrm{TA}_n(\mathrm{TC}_j, \mathrm{SC}_k) \quad (11)$$

$\overline{TA}$ is the mean TA and n is the number of tests used. TC and SC stand for particular combinations of tests and species used, respectively.

Comparison of TA between BPD and cost function distance metrics is shown in Figure 2. It's obvious that cost function distance gives significantly better results than BPD. At the maximum performance with 44 tests, correctly-identified topology was found in 83.27% of the 4000 trials. Performance of cost function is correlated with the number of tests. The overall Spearman rank correlation coefficient is 0.9960 with

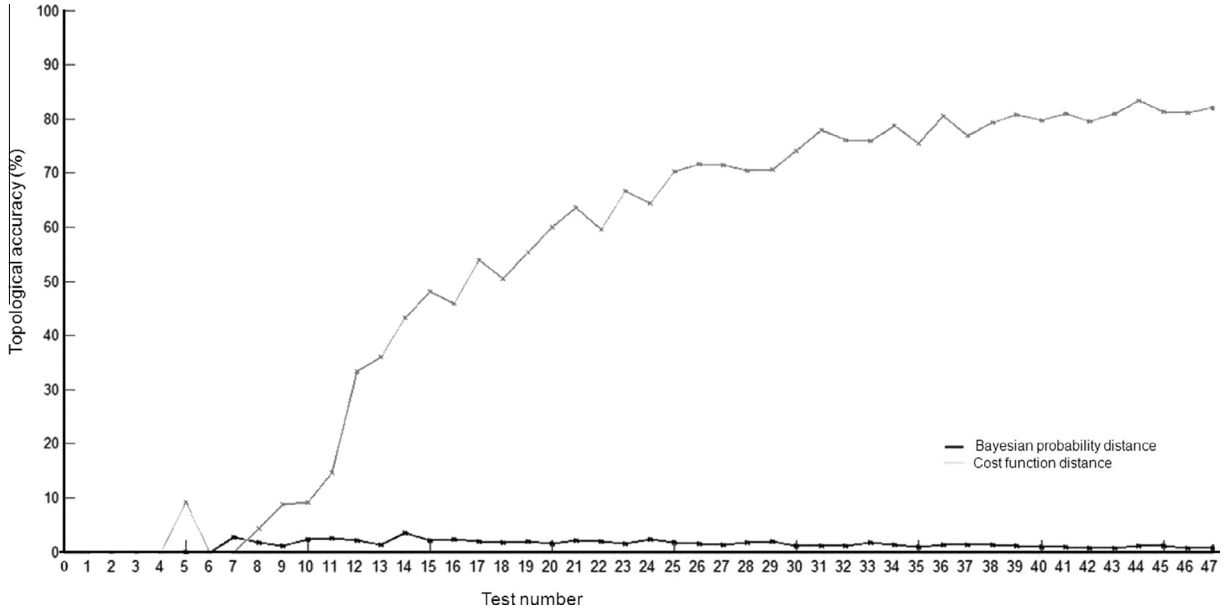


Figure 2 Comparison of topological accuracy of BPD and cost function distance metrics
X-axis denotes the test number and *Y*-axis denotes percentage of times when the topology of dendrogram is correctly identified, *i.e.*, topological accuracy. Dark line indicates topological accuracy of Bayesian probability distance (BPD) and gray line indicates topology accuracy of cost function distance. Clustering accuracy was calculated 4000 times for different randomly-chosen combinations of tests and species, and then the average was taken (Equation (11)). Random combinations of tests were iterated 40 times and random combinations of species were iterated 100 times.

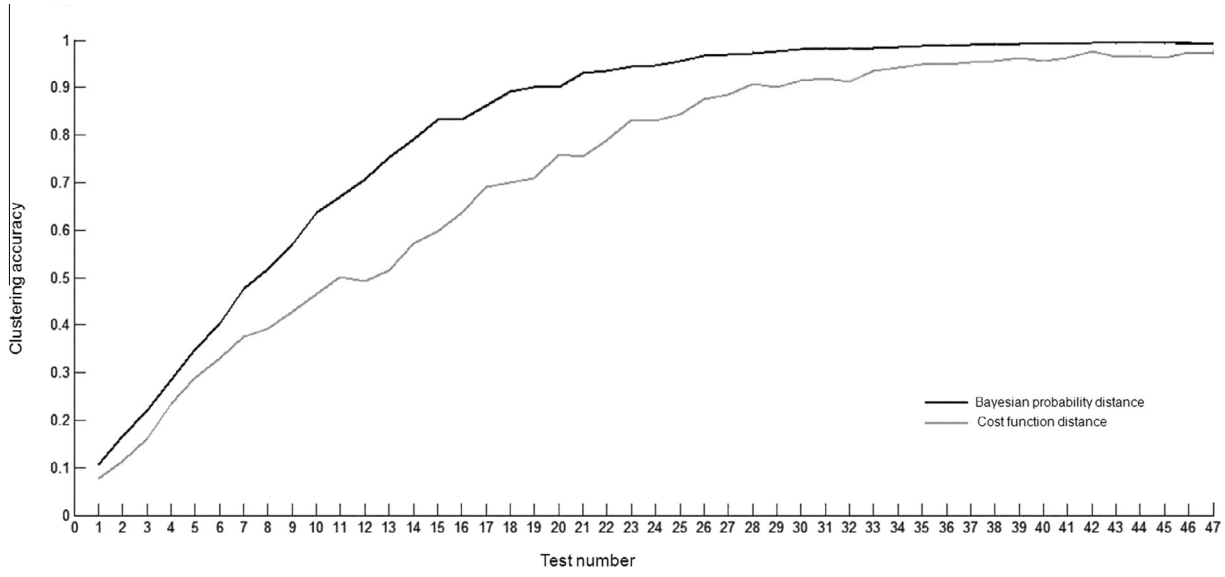


Figure 3 Comparison of clustering accuracy of BPD and cost function distance metrics
X-axis denotes the test number and *Y*-axis denotes clustering accuracy. Dark line indicates clustering accuracy of Bayesian probability distance (BPD) and gray line indicates clustering accuracy of cost function distance metric. Clustering accuracy was calculated 4000 times for different randomly-chosen combination of tests and species, and then the average was taken (Equation (7)). Random combinations of tests were iterated 40 times and random combinations of species were iterated 100 times.

a *P* value of 6.1189E−49. Conversely, BPD scores poorly for TA: topology is correctly identified only 3.45% of the times even at its best performance (with 14 tests) and 0% when less than 8 tests were used. Moreover, there is not much correlation between the performance and test number (Spearman rank correlation coefficient: 0.2911; *P* value: 0.0471).

Topological analysis has shown that cost function is better at identifying phylogenetic relationship between the species (*i.e.*, higher topological accuracy). Using synthetic data, we also compared CA of BPD and cost function. Our data showed that BPD is undoubtedly superior to cost function (**Figure 3**). Therefore, BPD and cost function should be used together to better understand the relationships and clustering of isolates.

Performance on real data

To assess the performance of BioCluster on real data, we collected some biochemical test result data from different published literature and also some data from several lab experiments to validate our tool (Table S2). Consequently, we obtained 36 samples and compared CA using IHC and HC. We found that IHC showed higher CA than HC (**Figure 4**), indicating that results from real data are consistent with results from synthetic data. We also tested BioCluster for its capability to identify different bacterial species correctly using the same dataset. Out of 36 samples, 33 samples were correctly identified (91%). We went further to test BioCluster by selecting different test numbers in random combination to find out the optimum test numbers for bacterial identification (Table S3). Our data indicated that for a test number of 10 or less of randomly-chosen tests, the test accuracy is around 40% (Equation (2)), and the tool can identify enterobacterial species with an accuracy of 90%–98% when the number of tests is 18 or higher. Therefore, higher accuracy can be achieved with the increasing test numbers.

Discussion

Automation of bacterial clustering and identification process has great potential in screening large datasets of biochemical test results of Enterobacteriaceae as opposed to the manual process. In this study, we have shown that improved clustering performs better by taking into account the probabilistic nature of test results, which allows more accurate results than the direct clustering. Here, we have introduced two distance metrics; one is BPD and the other is cost function. BPD is good at lower level clustering but performs poorly at higher level. Therefore, BPD is expected to perform very well at clustering all samples of *Escherichia coli* or samples of *Proteus vulgaris* accurately but not for the clustering at the genus level. On the other hand, cost function shows better results at clustering species of the same genus (or higher) together, but it is relatively error-prone for clustering at species level, compared to BPD. By combining these two metrics (separately in the BioCluster) together, users can infer the clustering both at species level and higher levels.

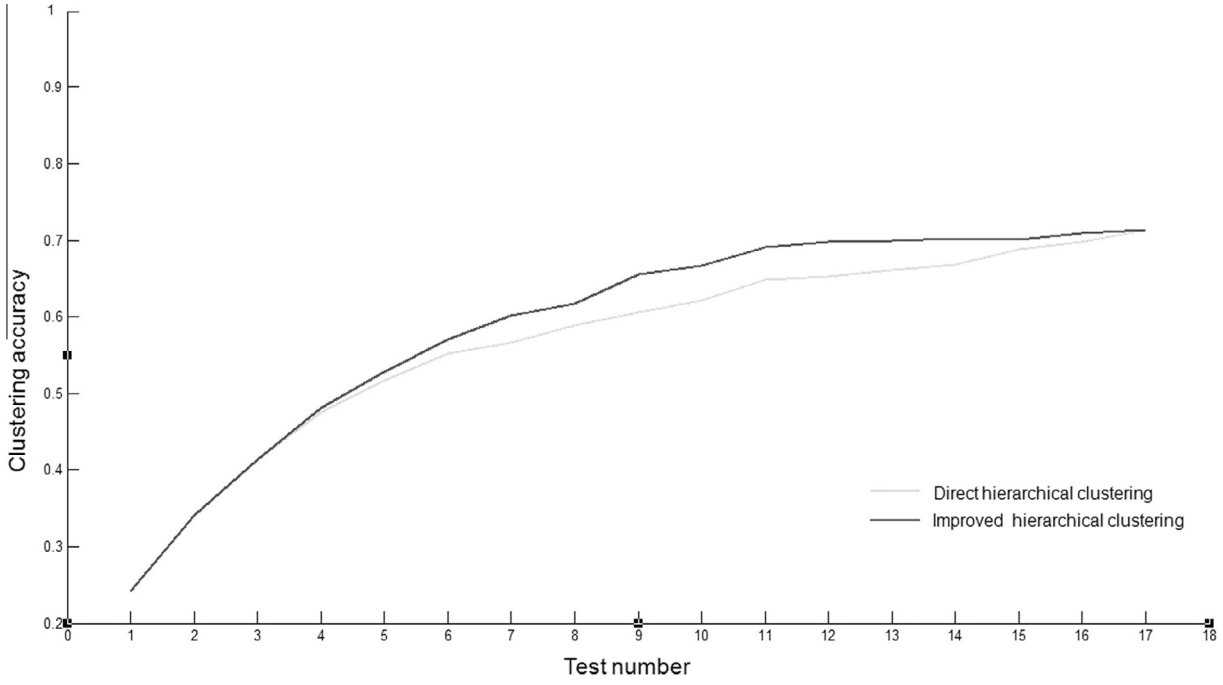


Figure 4 Comparison of clustering accuracy of DHC and IHC using real data

X-axis denotes the test number and *Y*-axis denotes clustering accuracy. Dark line indicates clustering accuracy of Improved Hierarchical Clustering (IHC) and the gray line indicates clustering accuracy of direct hierarchical clustering (DHC). IHC showed better performance than DHC. Clustering accuracy was calculated 4000 times for different randomly-chosen combination of tests and species, and then the average was taken (Equation (7)). Random combinations of tests were iterated 40 times and random combinations of species were iterated 100 times.

To the best of our knowledge, no similar tool for biochemical clustering of Enterobacteriaceae is available in the public domain so far. Hence, our tool can provide a unique role for identifying and clustering Enterobacteriaceae. It allows the presumptive identification process to be more robust and flexible as it can deal with variability. To properly cluster and identify the isolates, higher number of tests is always preferable. With low test numbers, two different organisms can show the same (or very similar) test results by chance alone. But by using a sufficiently-high number of tests, this problem could be resolved. BioCluster is very efficient at discerning organisms when enough information is provided. Therefore, BioCluster could give substantial advantage in terms of more accurate, rapid and reproducible data analysis.

In summary, BioCluster is a user-friendly MATLAB-based tool. All operation of the tool is done through GUI, therefore, no programing skill and acquaintance with programing language is required.

Authors' contributions

AA carried out the design, implementation, modification, data acquisition, and uploading of BioCluster tool, and drafted the manuscript. SMA was involved in tool development, data acquisition, and interpretation, manuscript preparation. MS participated in conception, data interpretation and coordination, and helped in draft preparation and critical revision of the draft. MAH contributed in conception, tool design, data interpretation, and manuscript revision. All authors read and approved the final manuscript.

Competing interests

The authors declare that there are competing interests.

Acknowledgments

The work was supported by the grants from the Ministry of Science and Technology (S&T) of Bangladesh (Grant No. HEQEP CP236) and the University Grants Commission (UGC).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2015.03.007.

References

[1] Garrity G, Brenner DJ, Krieg NR, Staley JRE. Bergey's Manual of Systematic Bacteriology. Vol 2: The Proteobacteria, Part B: The Gammaproteobacteria. Springer, US; 2005.
[2] Linton AH, Hinton MH. Enterobacteriaceae associated with animals in health and disease. Soc Appl Bacteriol Symp Ser 1988;17:71S–85S.
[3] Brenner DJ. Introduction to the family Enterobacteriaceae. In: Balows A, Trüper HG, Dworkin M, Harder W, Schleifer KH, editors. The prokaryotes. New York: Springer-Verlag; 1992. p. 2673–95.
[4] Traub WH, Raymond EA, Linehan J. Identification of Enterobacteriaceae in the clinical microbiology laboratory. Appl Microbiol 1970;20:303–8.
[5] Gibbs BM, Shapton DA. Identification methods for microbiologists. Part B J Appl Bacteriol 1968, xi + 212.
[6] Carpenter KP, Lapage SP, Steel KJ. Biochemical identification of Enterobacteriaceae. In: Gibbs BM, Skinner FA, editors. Identification methods for microbiologists. London: Academic press; 1966. p. 21–3.
[7] Minka T. Distance measures as prior probabilities. Technical report, Carnegie Mellon University, 2000.
[8] Jordan MI, Kearns MJ, Solla SA. Advances in neural information processing systems 10. Proceedings of the 1997 Conference. 1998.
[9] Johnson SC. Hierarchical clustering schemes. Psychometrika 1967;32:241–54.
[10] Tan P-N, Steinbach M, Kumar V. Introduction to data mining. 1st ed. Pearson Education India, 2007.
[11] Bochner BR. Global phenotypic characterization of bacteria. FEMS Microbiol Rev 2009;33:191–205.
[12] Basseville M. Distance measures for signal processing and pattern recognition. Signal Process 1989;18:349–69.
[13] Singhal A. Modern information retrieval: a brief overview. IEEE Data Eng Bull 2001;24:35–43.
[14] Baxter J. The canonical distortion measure for vector quantization and function approximation. In: Thrun S, Lorien P, editors. Learning to learn. US: Springer; 1998. p. 159–79.
[15] Yianilos PN. Metric learning via normal mixtures. Technical report. Princeton, NJ: NEC Research Institute; 1995.