Journal of Industrial Microbiology (1996) 17, 490–497
© 1996 Society for Industrial Microbiology 0169-4146/96/$12.00

Bioinformatics tools for the study of microbial diversity

H Sugawara[1], S Miyazaki[1], J Shimura[2] and Y Ichiyanagi[2]

[1] *Center for Information Biology, National Institute of Genetics, Mishima 411;* [2] *WFCC-MIRCEN World Data Centre for Microorganisms (WDCM), The Institute of Physical and Chemical Research (RIKEN), Wako, Saitama 351-01, Japan*

Although computers are capable of storing a huge amount of data, there is a need for more sophisticated software to assemble and organize raw data into useful information for dissemination. Therefore we developed tools that assist in gathering and categorizing data for the study of microbial diversity and systematics. The first tool is for data retrieval from heterogeneous data sources on the INTERNET. The second tool provides researchers with a polyphasic view of microbes based on phenotypic characteristics and molecular sequence data.

Keywords: biodiversity; systematics; classification; phylogeny; identification; database; INTERNET; multi-media

Introduction

The Convention on Biological Diversity [2] came into effect in December 1993 and has attracted the attention of international, regional and national organizations dealing with biodiversity. The decision makers paid attention to the protection of endangered species of animals and plants and finally recognized the importance of scientific issues related to the utilization of organisms. They are now also recognizing the value of microorganisms [6].

Microorganisms are diverse in species, yet many have not been isolated [5]. Although the definition of species is a little ambiguous, it is estimated that only 1–10% of all species have been described so far. Therefore it is important for research dealing with microbial diversity to include making an inventory and monitoring living microorganisms *in situ*.

Microbial diversity research will yield a number of cultures that should be maintained in culture collections [14] for polyphasic characterization and for studies on products for application. A large amount of data will be produced in a wide variety of categories such as phenotypes, genotypes, environments and pathways. Systematics are needed to categorize new cultures and incorporate raw data into useful information.

If we take into consideration the number and heterogeneity of the cultures and data, it is obvious that no single international, regional or national institute will be able to process all the information from projects on microbial diversity. The solution is the establishment of networks among geographically and scientifically distributed institutions. This networking is technically feasible thanks to bioinformatics and the INTERNET [9].

To meet the needs of such studies, we developed a tool for retrieving data from data sources on the INTERNET and a system for organizing data to aid in taxonomic structuring of microbes.

Materials and methods

We used a SUN4/10 workstation (128 MB memory) that is connected to the global INTERNET via the local area network of the Institute of Physical and Chemical Research, and STAnet and Inter-Ministry Research Information Network (IMnet) sponsored by the Science and Technology Agency of the Japanese government.

Multiple databases, AHMII

During the last 2 years, there has been an explosion of data servers on the INTERNET, especially on World Wide Web (WWW) servers. WWW servers are counted by pages and more than 10 million pages are on the INTERNET, according to Lycos, Inc [10]. Therefore on-line catalogues such as Lycos are in service to help users locate sites relevant to their research. However, it is not easy for them to squeeze data from a site, even when a good site can be located. This is because the data structure and search strategy are different from site to site.

Common gateway interfaces (CGI) make it possible to

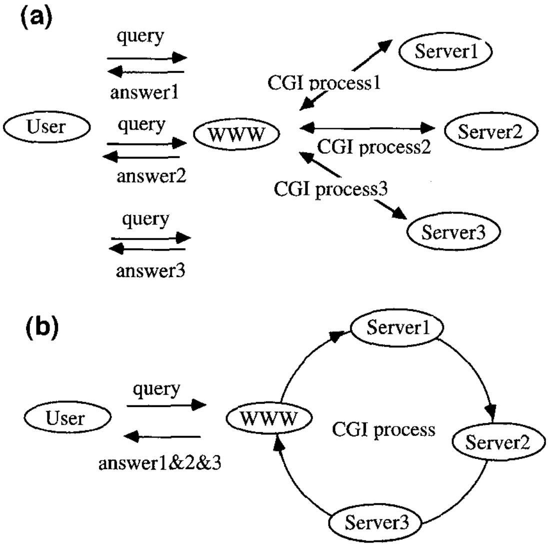


Figure 1 (a) A simple application of CGI to query multiple databases. (b) Query to multiple databases by use of socket.

Correspondence: Dr H Sugawara, Center for Information Biology, National Institute of Genetics, Mishima 411, Japan

Received 15 March 1996; accepted 8 October 1996

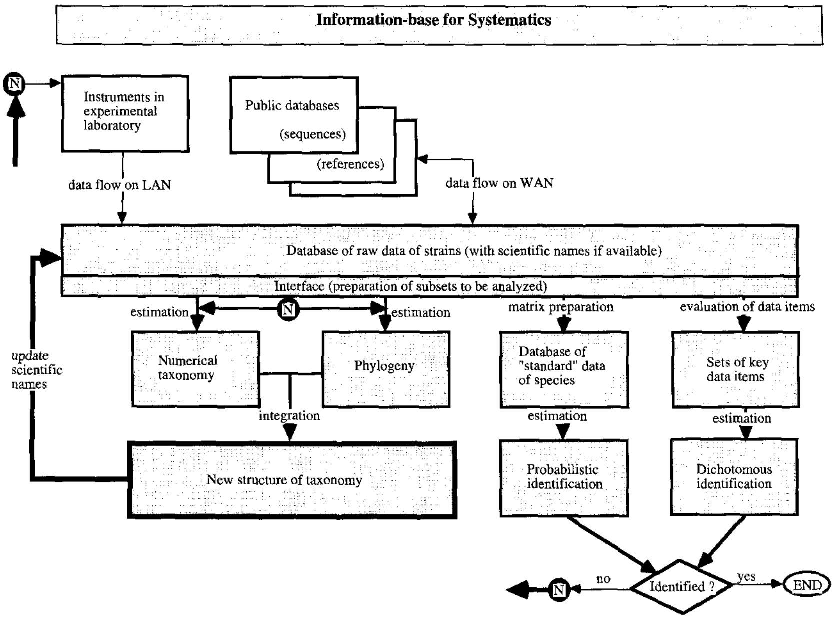

Information-base for Systematics
N
Instruments in experimental laboratory
Public databases
(sequences)
(references)
data flow on LAN
data flow on WAN
Database of raw data of strains (with scientific names if available)
Interface (preparation of subsets to be analyzed)
estimation
N
estimation
matrix preparation
evaluation of data items
update scientific names
Numerical taxonomy
Phylogeny
Database of "standard" data of species
Sets of key data items
integration
estimation
estimation
New structure of taxonomy
Probabilistic identification
Dichotomous identification
no
N
Identified ?
yes
END

Figure 2 The structure of Information-base for systematics.

send a query to multiple remote servers in sequence as shown in Figure 1a. Yet, it is not possible to assemble answers from the servers to display them in an integrated format. In the case of Figure 1a, a remote server takes over the control of the data flow from the user until it returns the answer. The user cannot access another server at the same time.

We developed a tool with which users are able to simultaneously send a query to and get answers from multiple data sources on the INTERNET.

The tool was implemented with an interface-named socket for interprocess communication in UNIX, a CGI and WWW. The CGI proceeds in three steps. In the first step, the CGI establishes links with all potential multiple servers as shown in Figure 1b. In the second step, the CGI creates a query for a server based on keywords that a user types in. In the third step, the CGI converts the result of a search into the HTML format. The CGI repeats the second and third steps for every member of the multiple databases.

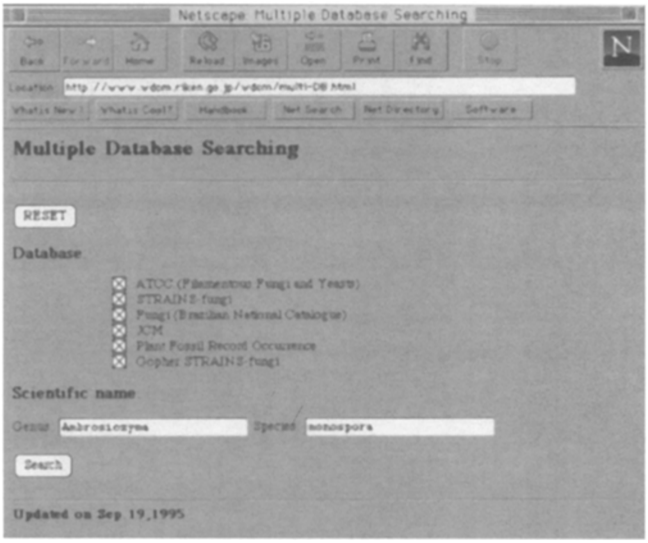
We named this system an Agent for Hunting Microbial Information (AHMII) and it works for bacteria, fungi and cell lines. We implemented the following six servers in AHMII for bacteria:

- the catalogue of the American Type Culture Collection (ATCC);
- the list of scientific names of strains preserved in culture collections registered in the WFCC World Data Centre for Microorganisms (WDCM);
- the catalogue of the Japan Collection of Microorganisms (JCM);
- the catalogue of Centraalbureau voor Schimmelcultures (CBS);
- the approved list of bacterial names digitized by the World Data Centre for Microorganisms (WDCM);
- the taxonomy database of the National Center for Biotechnology Information (NCBI) which is included in the international nucleotide sequence database.

Information-base for systematics

We developed an information-base for systematics to support the classification and identification of numerous microbes that will be isolated in projects involving microbial diversity.

The information-base was designed by analyzing the behavior of taxonomists and its structure is summarized in Figure 2. Taxonomists accommodate experimental data and public data into a database of the relational database management system, SYBASE. They apply, to subsets of the database, such analyses as numerical taxonomy, phylogenetic analysis, probabilistic identification based on a positive



Location: http://www.wdcm.riken.go.jp/wdcm/multi-DB.html

Multiple Database Searching

RESET

Database

☒ ATCC (Filamentous Fungi and Yeasts)
☒ STRAINS-fungi
☒ Fungi (Brazilian National Catalogue)
☒ JCM
☒ Plant Fossil Record Occurrence
☒ Gopher STRAINS-fungi

Scientific name

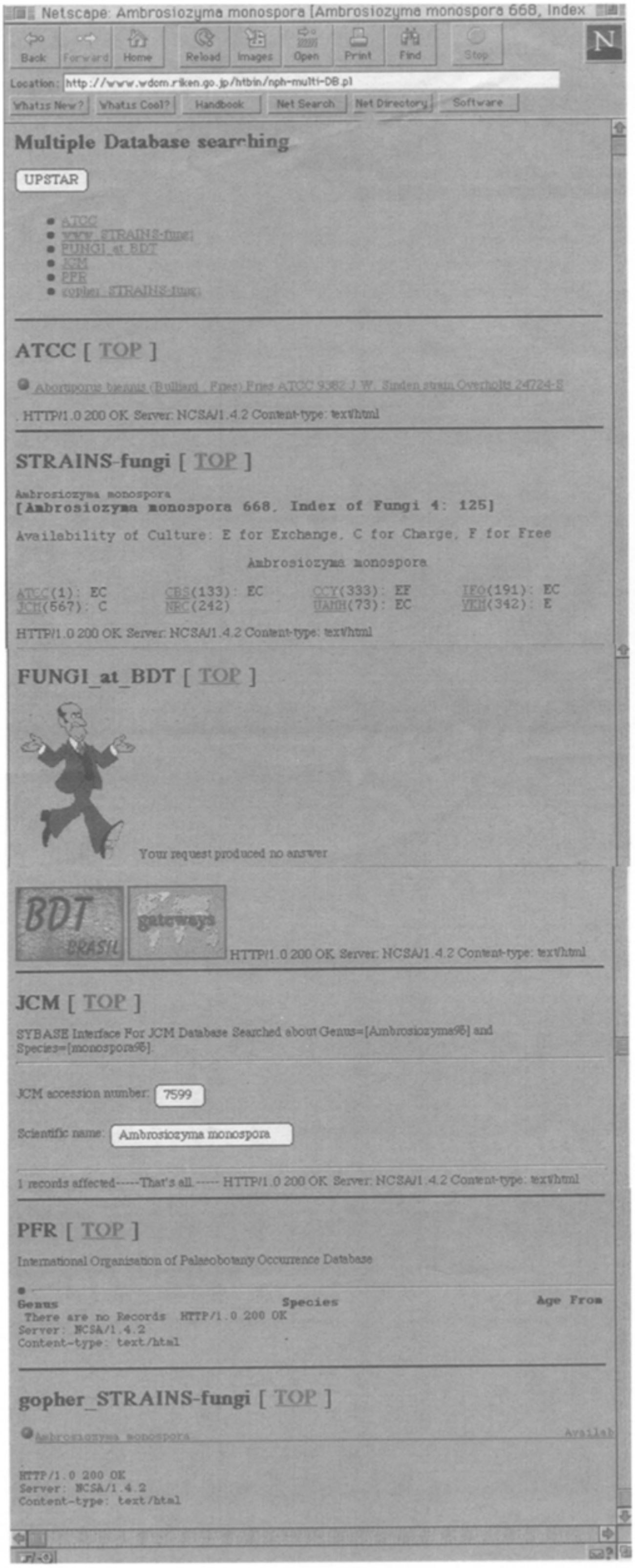
Genus: Ambrosiozyma Species: monospora

Search

Updated on Sep 19,1995

Figure 3 The home page of the multiple databases.

rate matrix, and sequential identification using dichotomous keys [12,13]. They may compare results of different methods for analysis on the same subset to get a reasonable and robust view of the taxonomic structure of microbes.



Location: http://www.wdom.riken.go.jp/htbin/nph-multi-DB.pl

Multiple Database searrhing

- ATCC
- www_STRAINS-fungi
- FUNGI_at_BDT
- JCM
- PFR
- gopher_STRAINS-fungi

ATCC [TOP]

● Abortiporus biennis (Bulliard : Fries) Fries ATCC 9382 J. W. Sinden strain Overholts 24724-S

. HTTP/1.0 200 OK Server: NCSA/1.4.2 Content-type: text/html

STRAINS-fungi [TOP]

```
Ambrosiozyma monospora
[Ambrosiozyma monospora 668, Index of Fungi 4: 125]

Availability of Culture: E for Exchange, C for Charge, F for Free

                    Ambrosiozyma monospora

ATCC(1): EC        CBS(133): EC        CCY(333): EF        IFO(191): EC
JCM(567): C        NRC(242)            UAMH(73): EC        VKM(342): E
```

HTTP/1.0 200 OK Server: NCSA/1.4.2 Content-type: text/html

FUNGI_at_BDT [TOP]

Your request produced no answer

HTTP/1.0 200 OK Server: NCSA/1.4.2 Content-type: text/html

JCM [TOP]

SYBASE Interface For JCM Database Searched about Genus=[Ambrosiozyma%] and Species=[monospora%]

JCM accession number: 7599

Scientific name: Ambrosiozyma monospora

1 records affected-----That's all.----- HTTP/1.0 200 OK Server: NCSA/1.4.2 Content-type: text/html

PFR [TOP]

International Organisation of Palaeobotany Occurrence Database

```
●
Genus                                 Species                              Age From
 There are no Records  HTTP/1.0 200 OK
Server: NCSA/1.4.2
Content-type: text/html
```

gopher_STRAINS-fungi [TOP]

```
●Ambrosiozyma monospora                                                    Availab

HTTP/1.0 200 OK
Server: NCSA/1.4.2
Content-type: text/html
```

Figure 4 The result of a search of the multiple databases.

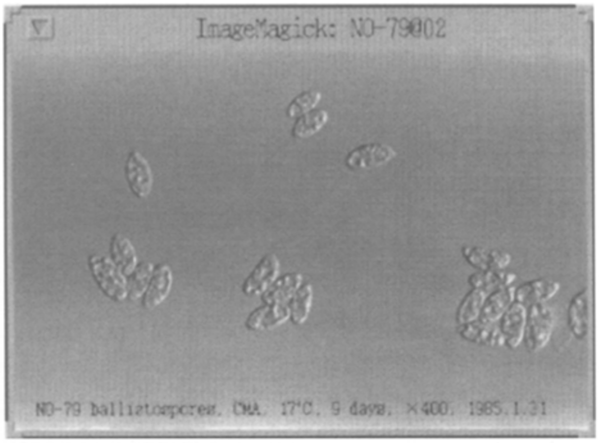
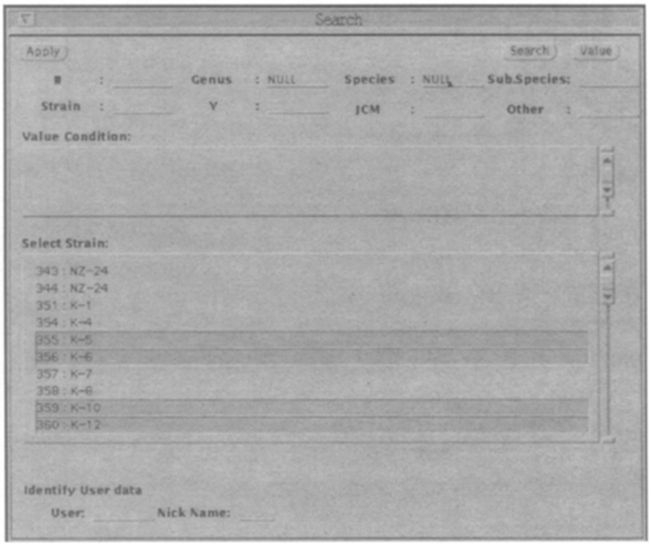


Figure 5 An image of yeasts in the information-base.



```
#       :            Genus   : NULL         Species  : NULL        Sub.Species:
Strain  :            Y       :              JCM      :             Other      :
Value Condition:

Select Strain:
  343 : NZ-24
  344 : NZ-24
  351 : K-1
  354 : K-4
  355 : K-5
  356 : K-6
  357 : K-7
  358 : K-8
  359 : K-10
  360 : K-12

Identify User data
  User:            Nick Name:
```

Figure 6 The preparation of a subset for the identification.

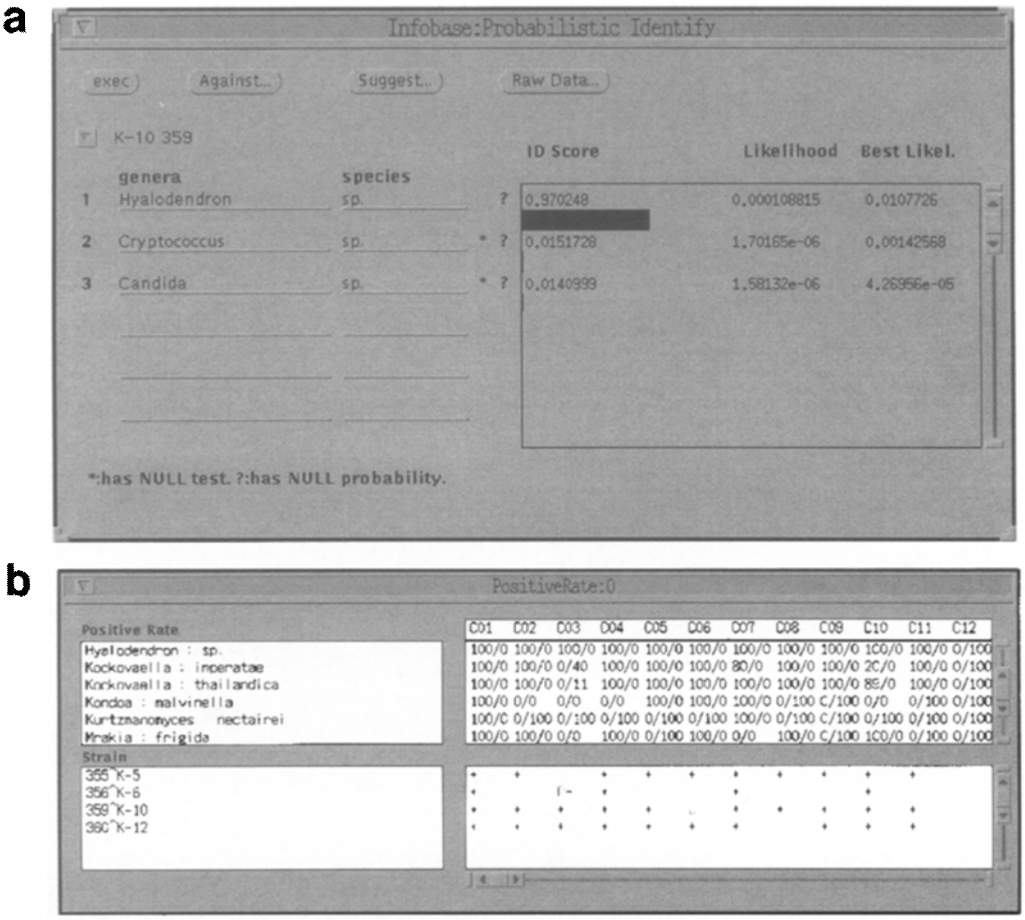
They have tried to integrate the systematics based on phenotypic data and phylogenetic classification. They will try to identify new isolates from the environment. If the identification is not successful, they will either repeat the experiment or carry out classification of the unknown strain and strains already described together. They will propose a new scientific name, if a strain is clearly differentiated from the taxonomic units known.

The information-base uses SYBASE with the APT workbench and library, a C compiler, Sun FORTRAN, SSLII library, DEV Guide, Sun Phigs, Open Windows3 library and modules in PHYLIP [11] and other public domain software [4,7,8,15].

Results

AHMII

The home page of the AHMII for bacteria is introduced in Figure 3, which is a screen dump of a WWW browser, Netscape Navigator. A user selects databases to search by checking the boxes on the home page and typing in a scientific name as a query. It takes about 10 seconds to get



K-10 359

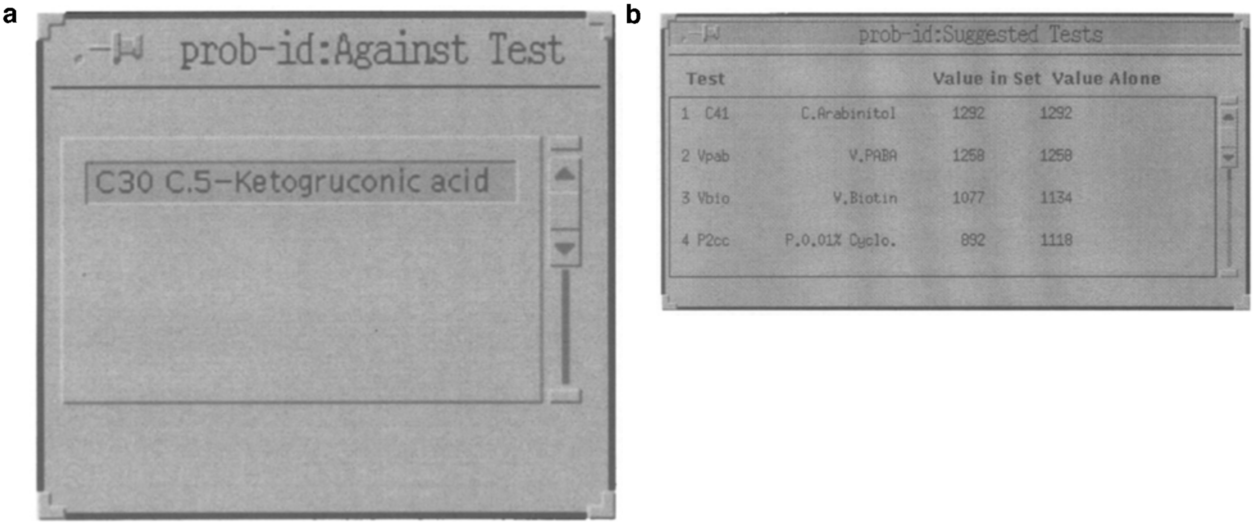
	genera	species		ID Score	Likelihood	Best Likel.
1	Hyalodendron	sp.	?	0.970248	0.000108815	0.0107726
2	Cryptococcus	sp.	* ?	0.0151728	1.70165e-06	0.00142568
3	Candida	sp.	* ?	0.0140999	1.58132e-06	4.26956e-05

*:has NULL test. ?:has NULL probability.

Positive Rate	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12
Hyalodendron : sp.	100/0	100/0	100/0	100/0	100/0	100/0	100/0	100/0	100/0	1C0/0	100/0	0/100
Kockovaella : imperatae	100/0	100/0	0/40	100/0	100/0	100/0	80/0	100/0	100/0	2C/0	100/0	0/100
Korknvaella : thailandica	100/0	100/0	0/11	100/0	100/0	100/0	100/0	100/0	100/0	8S/0	100/0	0/100
Kondoa : malvinella	100/0	0/0	0/0	0/0	100/0	100/0	100/0	0/100	C/100	0/0	0/100	0/100
Kurtzmanomyces nectairei	100/C	0/100	0/100	0/100	0/100	0/100	100/0	0/100	C/100	0/100	0/100	0/100
Mrakia : frigida	100/0	100/0	0/0	100/0	0/100	100/0	0/0	100/0	C/100	1C0/0	0/100	0/100

Strain												
355^K-5	+	+		+	+	+	+	+	+	+	+	
356^K-6	+		|-	+			+			+		
359^K-10	+	+	+	+	+		+	+	+	+	+	
360^K-12	+	+	+	+	+	+	+		+	+	+	

Figure 7 (a) The result of the probabilistic identification. (b) Comparison of the raw data of unknown strains and standard positive rate matrix. The symbols of S, D, W, + and – in the raw data table stand for slow positive, delayed positive, weak positive, positive and negative respectively.



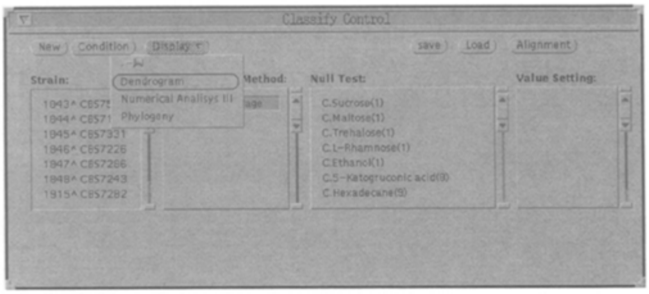
C30 C.5-Ketogruconic acid

Test		Value in Set	Value Alone
1 C41	C.Arabinitol	1292	1292
2 Vpab	V.PABA	1258	1258
3 Vbio	V.Biotin	1077	1134
4 P2cc	P.0.01% Cyclo.	892	1118

Figure 8 (a) Against data items for the result of the identification. (b) Suggested data items for further identification. The first letters like C, V and P of data items represent data categories of carbon assimilation, vitamins required and production. The values in an arbitrary unit represent the power of the data items that classify strains.

answers from the six servers after clicking on the button 'search' in the home page. AHMII does not stick but assembles the results of searches, even if some of the servers are down. The time depends on network congestion.

The results are partly given in Figure 4, when data for *Saccharomyces cerevisiae* were searched for. Users are able to compare results from distributed servers in a common HTML format, even though search engines are different, depending on the servers. For example, the server for WDCM strain bacteria uses WAIS and the server for JCM uses SYBASE. More detailed data will be displayed in the screen immediately after a user clicks buttons or characters underlined. Therefore, it is easy for users to simultaneously check the nucleotide sequence data of a strain of *S. cerevisiae* from the page of NCBI Taxonomy in Figure 4 and the



Strain:
1043^ CBS75
1044^ CBS71
1045^ CBS7331
1046^ CBS7226
1047^ CBS7286
1048^ CBS7243
1915^ CBS7282

Null Test:
C.Sucrose(1)
C.Maltose(1)
C.Trehalose(1)
C.L-Rhamnose(1)
C.Ethanol(1)
C.5-Katogruconic acid(9)
C.Hexadecane(9)

Figure 10 The control panel for classification.

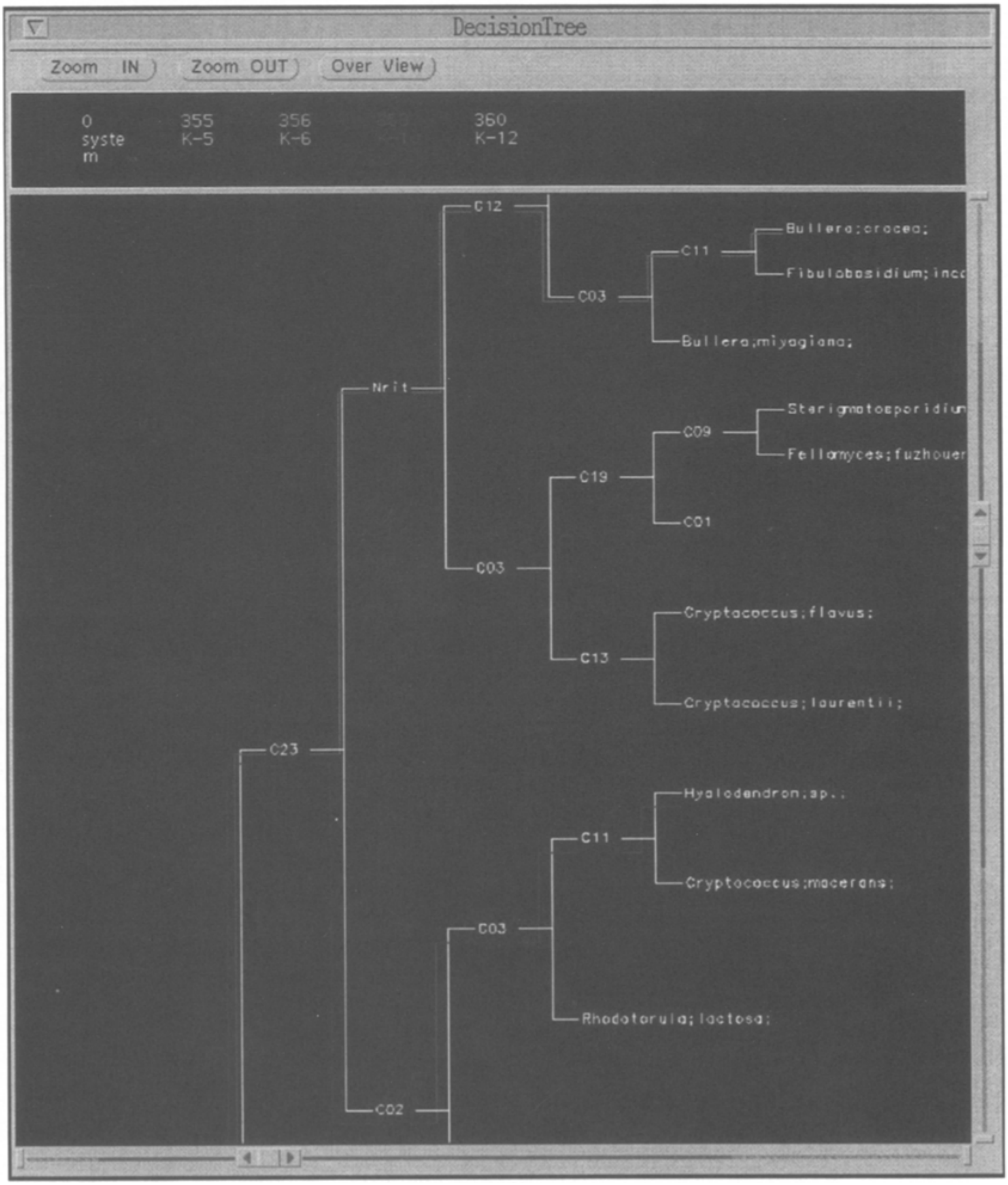
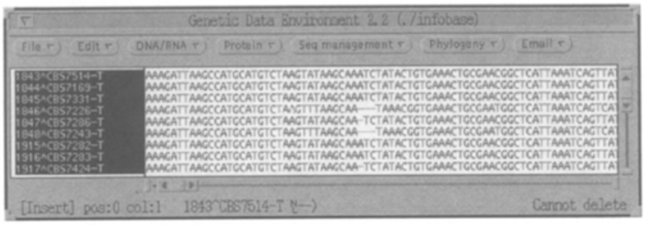


Figure 9 An example of the sequential identification.


Genetic Data Environment 2.2 (./infobase)
File
Edit
DNA/RNA
Protein
Seq management
Phylogeny
Email
[Insert] pos:0 col:1 1843^CBS7514-T
Cannot delete

```
1843^CBS7514-T AAAGATTAAGCCATGCATGTCTAAGTATAAGCAAATCTATACTGTGAAACTGCGAACGGCTCATTAAATCAGTTA
1844^CBS7169-T AAAGATTAAGCCATGCATGTCTAAGTATAAGCAAATCTATACTGTGAAACTGCGAACGGCTCATTAAATCAGTTA
1845^CBS7331-T AAAGATTAAGCCATGCATGTCTAAGTATAAGCAAATCTATACTGTGAAACTGCGAACGGCTCATTAAATCAGTTA
1846^CBS7226-T AAAGATTAAGCCATGCATGTCTAAGTTTAAGCAA---TAAACGGTGAAACTGCGAATGGCTCATTAAATCAGTCA
1847^CBS7286-T AAAGATTAAGCCATGCATGTCTAAGTATAAGCAA-TCTATACTGTGAAACTGCGAACGGCTCATTAAATCAGTTA
1848^CBS7243-T AAAGATTAAGCCATGCATGTCTAAGTTTAAGCAA---TAAACGGTGAAACTGCGAATGGCTCATTAAATCAGTCA
1915^CBS7282-T AAAGATTAAGCCATGCATGTCTAAGTATAAGCAAATCTATACTGTGAAACTGCGAACGGCTCATTAAATCAGTTA
1916^CBS7283-T AAAGATTAAGCCATGCATGTCTAAGTATAAGCAAATCTATACTGTGAAACTGCGAACGGCTCATTAAATCAGTTA
1917^CBS7424-T AAAGATTAAGCCATGCATGTCTAAGTATAAGCAA-TCTATACTGTGAAACTGCGAACGGCTCATTAAATCAGTTA
```

Figure 11 Alignment of sequence data.

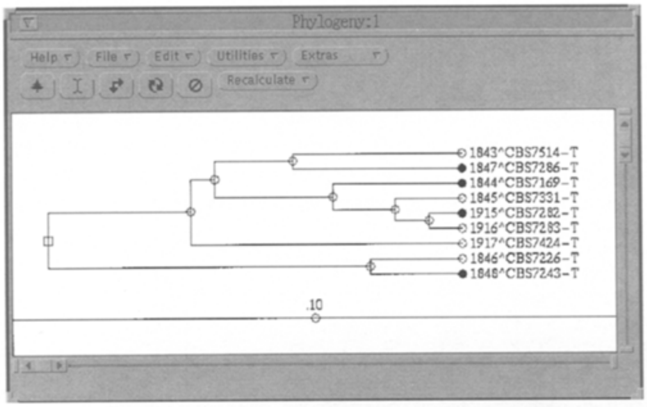


Figure 12 Phylogenetic tree.

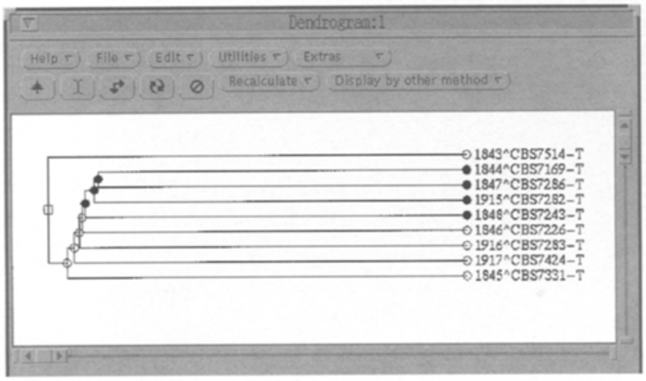


Figure 13 Dendogram from cluster analysis.

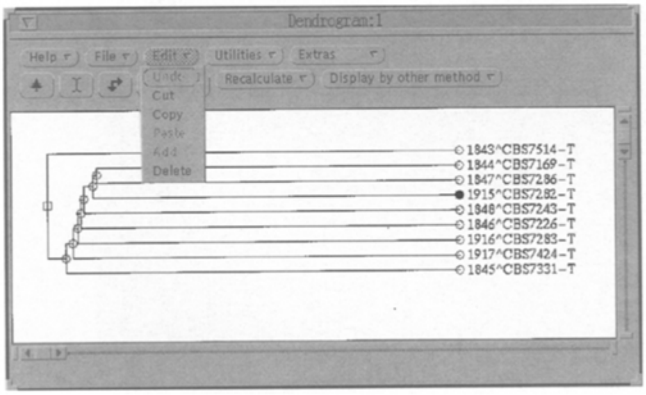


Figure 14 Direct manipulation of the strain in a tree.

availability of strains from ATCC, JCM, and other culture collections. Although the results are not shown in Figure 4, users can also refer to phenotypic characteristics provided by the server of CBS. In addition, users are able to cross-check a name with the list of scientific names authorized by the scientific community.

Information-base for systematics

Database: The information-base was tested for the taxonomy of yeasts. The categories of phenotypic data in the information-base are carbon assimilation, nitrogen assimilation, vitamin requirements, fermentation, GC%, quinone composition, whole cell sugar, cell wall sugar, growth temperature, products, spore and special remarks. It is possible to display data of subsets by specifying simple and/or complex conditions for the search. Image data are also incorporated in the information-base and retrievable onto the display as shown in Figure 5. That is, the user is able to compare the description in the text with the electron micrograph on the screen.

Identification: Data for unknown strains are registered in the information-base with no scientific names. Therefore a subset for identification is created by searching records that have NULL data in the data fields of Genus and Species as shown in Figure 6. The results of probabilistic identification [16] are given in Figure 7a. The solid horizontal bars in the sub-window stand for values of ID scores. In this case, the strain of K-10 359 (a private accession number) seems close to *Hyalodendron* sp. The user is able to analyze the results by referring to the raw data and the positive rate matrix as shown in Figure 7b. Note that expression of slow positive (S), delayed positive (D) and weak positive (W) in addition to positive (+) and negative (−) are used in raw data. The information-base also explicitly lists data items that are controversial in terms of the identification and suggests data items for further analysis as shown in Figure 8a and b.

One feature of the information-base is that a positive rate matrix will be automatically updated if an unknown strain is identified with a known species.

The information-base also produces a set of dichotomous keys from the database of known species. The strain K-10 359 is also tested against sequential identification and the result is graphically presented in Figure 9. The nodes and C** numbers in the figure correspond to key characters for discrimination, and the strain was identified as *Hyalodendron* sp by the set of data items of C23, C02, C03 and C11 in data items numbers.

Classification: In the information-base, a control panel for the classification is implemented (Figure 10). The user can set the subset of strains to be classified and methods and parameters of the analysis. Multiple alignment is also implemented in the classification module of the information-base (Figure 11).

The results of classification of the same subset are compared in Figure 12 and Figure 13. In the phylogenetic tree of Figure 12, the strains of 1847, 1844, 1915 and 1848 are comparatively distributed over the tree. Conversely, they are closely located in the dendrogram in Figure 13. The target strains become noticeable in the figures due to the change from open circles to dark circles by clicking with the mouse. Whenever the user changes a circle of a strain on one of the trees, the circle of the same strain in another tree immediately changes. Thus it is quite easy to evaluate results based on phenotypic data and sequence data to get a polyphasic view of the system of microbes. The user is also able to delete a strain from the tree by picking the circle with the mouse as shown in Figure 14 in order to see the robustness of the tree.

Numerical Analysis-III, which was developed by a group of Japanese statisticians, was applied to the same subset as analyzed in the above. The results are displayed in three-dimensional space (Figure 15) and are similar to the results of phylogenetic analysis. The circles and dots of corresponding strains in Figures 12, 13 and 15 are linked with each other. Thus, it is possible to simultaneously highlight a strain or strains in the three types of representations of relationships of the strains.

The information-base is technically accessible from UNIX workstations connected to the INTERNET, though Open Windows 3.0 and Sun Phigs are required.

Discussion

For the user of the multiple databases, AHMII can carry out a onetime query to remote sites that have heterogeneous data structures and assemble the results into the HTML format. By using scientific names as locators of data sources, the user can easily compare the results of six data sources, in the case of bacteria, without repeating the search. Accordingly AHMII provides a key clearing-house mechanism for the study of biodiversity [1].

In AHMII, expanding the data sources by the development of a CGI for a new search engine is straightforward. We may add useful data sources as much as we need. We will be able to further improve AHMII, if we are able to

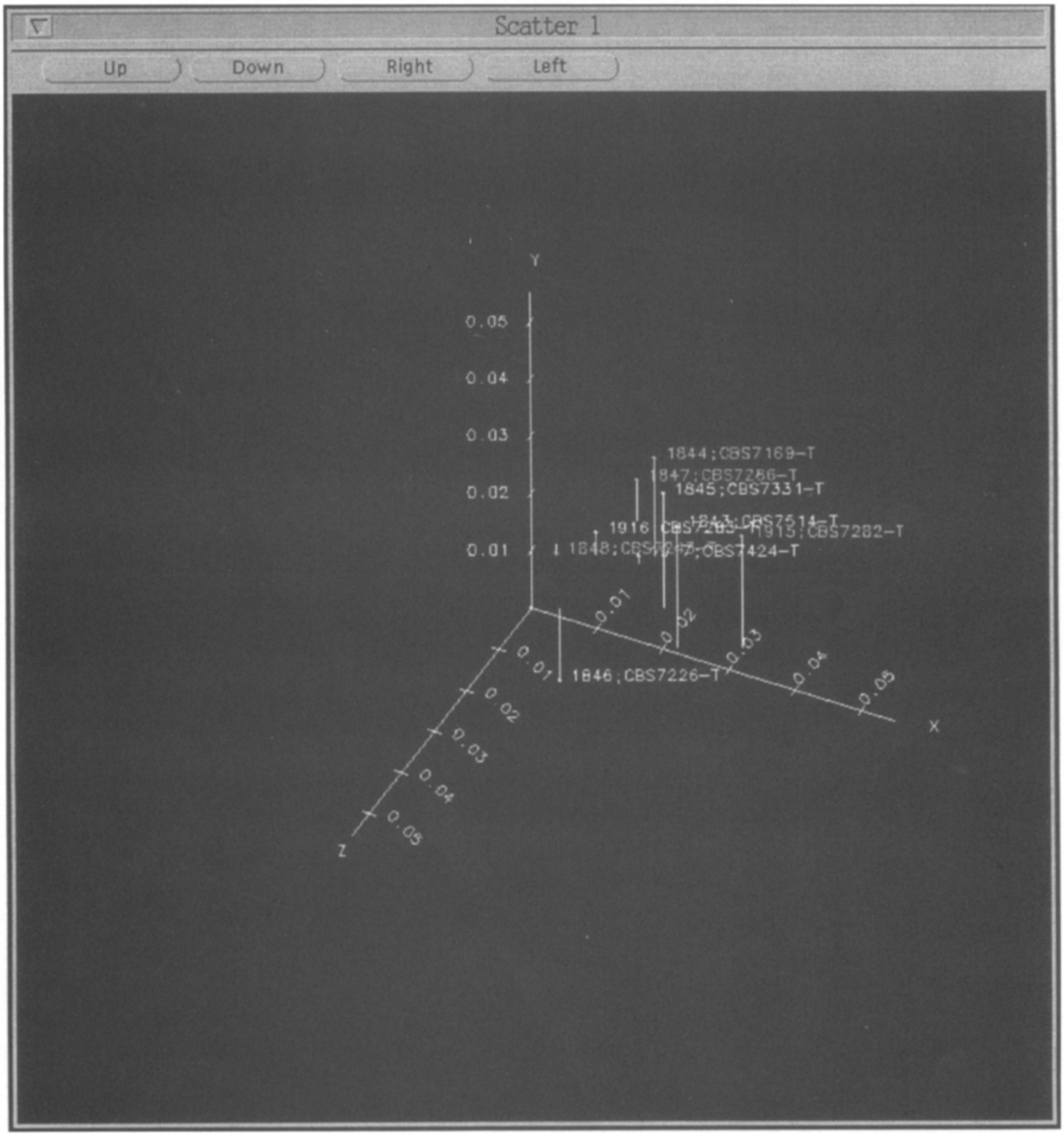


Figure 15 Three-dimensional plot of the strains based on numerical Analyses-III.

develop a CGI that creates interfaces to different search engines on-the-fly.

The information-base aims to provide an integrated information environment for taxonomists. The user is able to apply many kinds of analyses to the same subset from the database and enjoy friendly graphical user interface (GUI). Then he/she is able to compare the results to get a consistent view of the relationship among species and strains. It is proved by the explosive increase of users of Web browsers that a good GUI is powerful, valuable and important.

In the case of identification, sequential identification provides a clear-cut result. However, identification will not always be successful, especially if the data of unknown strains are missing some data items or if data items available for the dichotomous keys are few. In the information-base, the user is able to compare, on the same screen, the results of probabilistic identification and sequential identification. Therefore he/she is able to get insight into the relationship of unknown strains to existing species even if there are many blank data items, and at the same time, evaluate the discrimination power of the data items.

In the case of classification, it is interesting to note that the result of three-dimensional presentation of strains produced a taxonomic structure more similar to the results of phylogenetic analysis than that of numerical taxonomy (dendrogram). We have to carry out multi-dimensional analyses such as Numerical-Analysis III, principal components analysis, etc, if the dendrogram gives an inconsistent tree topology with the phylogenetic tree. It is highly probable that the inconsistency is caused by data compression from multi-dimensional space into one or two dimensions.

The information-base here is a prototype for a system that employs a polyphasic approach to systematics which is an indispensable tool for projects on microbial diversity. It is expected that the information-base will be expanded to include information on experimental methods and also many other strategies for data analysis [3].

Acknowledgements

A part of this study was performed using Special Coordination Funds of the Science and Technology Agency of the Japanese Government. The authors also express their thanks to Dr Takashi Nakase of the Japan Collection of Microorganisms (JCM) who provided the strain data for yeasts.

References

1 BIN21 Secretariat. 1995. Clearing-House Mechanism on Biological Diversity. Base de Dados Tropical, Campinus, Brazil.
2 Convention on Biological Diversity. 1994. Convention on Biological Diversity. Geneva Executive Center of UNEP, Geneva, Switzerland.
3 Fortuner R. 1993. Advances in Computer Methods for Systematics Biology. The Johns Hopkins University Press, Baltimore, MD, USA.
4 GDE (The Genetic Data Environment) (S Smith, ed). gopher://rdpgopher.life.uiuc.edu/11/programs/Editor_GDE.
5 Hammond PM. 1995. Described and estimated species numbers: an objective assessment of current knowledge. In: Microbial Diversity and Ecosystem Function (Allsop D, RR Colwell and DL Hawksworth, eds), pp 29–71, CAB International, Egham.
6 Hawksworth DL and RR Colwell (eds). 1992. Biodiversity amongst microorganisms and its relevance. Biodiversity and Convention 1: 221–345.
7 Higgins DG, AJ Bleasby and R Fuchs. 1991. CLUSTAL V: improved software for multiple sequence alignment. CABIOS 8: 189–191.
8 Higgins DG and PM Sharp. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. CABIOS 5: 151–153.
9 Krol E. 1992. The Whole Internet Users' Guide and Catalogue. O'Keilly and Associates, Sebastopol, CA, USA.
10 Lycos Inc. http://www.lycos.com/lycosinc/comparison/html.
11 PHYLIP (Phylogeny Inference Package) by J Felsenstein. http://evolution.genetics.washington.edu/phylip.html.
12 Priest F and B Austin. 1993. Modern Bacterial Taxonomy, 2nd edn. Chapman & Hall, London, UK.
13 Ridley M. 1986. Evolution and Classification. Longman, London, UK.
14 Sugawara H and B Kirsop. 1994. The WFCC World Data Center on microorganisms and global statistics on microbial resources centres. In: The Biodiversity of Microorganisms and the Role of Microbial Resource Centres (Kirsop B and DL Hawksworth, eds), pp 53–64, World Federation for Culture Collections, Braunschweig, Germany.
15 Treetol by M Maciukenas. gopher://rdpgopher.life.uiuc.edu/11/programs/TreeTool.
16 Willcox WR, SP Lapage, S Bascomb and MA Curtis. 1973. Identification of bacteria by computer: theory and programming. J Gen Microbiol 77: 317–330.