Microbiol. Cult. Coll. Dec. 2003. p. 59－67 Vol. 19, No. 2

An e－Workbench for the Study of Microbial Diversity: the System Design and Basic Functions

Hideaki Sugawara[1, 2]*, Naoto Tanaka[1, 3], and Satoru Miyazaki [1]

[1] Center for Information Biology and DDBJ, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan
[2] SOKENDAI, Hayama, Kanagawa 240-0193, Japan
[3] Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Corporation (JST), 5-3 Yonbancho, Chiyoda-ku, Tokyo 102-8666, Japan

We developed a suite for the study of microbial diversity. The user can seamlessly repeat a work flow from data management and data analysis to evaluation of analytical results. The suite includes functions for such data processing as: design of databases; storage and retrieval of data; numerical analysis; phylogenetic analysis; discriminative and probabilistic identification.

An XML (eXtensible Markup Language) document was adopted to realize flexible database management. It is also a seamless operation to integrate databases and data analysis tools at a local site with those at remote sites. We call the suite e－Workbench through association with a workbench in biological laboratories. The e－Workbench is written in Java so that it is executable in Windows, Macintosh, Linux and UNIX machines and is available free through the Internet to be run on a laptop computer. The e－Workbench is available online at http://www.wdcm.org/.

Key words: e－Workbench, database, XML, CORBA, interoperability, microbial diversity, classification, phylogeny, identification

INTRODUCTION

Just as a workbench is a fundamental instrument for laboratory experiments, so an e－Workbench will be for computational experiments. It is a suite composed of databases and data analysis tools and enables the user to process data intuitively and seamlessly. We implemented an e－Workbench for the polyphasic analysis (14) that is required to understand microbial diversity. Although the term "workbench" has been used in bioinformatics (9, 13, 15), the term "e－Workbench" in this paper was used as an electronic workbench for computational experiments.

Polyphasic analysis has to evaluate both phylogeny and phenotypic characters, and is based on a wide range of data from the molecular (gene sequence) to phenotypic (metabolites and various reactions) level. Where can we capture the data for the analysis? Sequences and their annotations are registered in and retrievable from the International Nucleotide Sequence Databases (INSD: DDBJ/EMBL bank/GenBank, http://www.ddbj.nig.ac.jp). As of May 2003, 54,060 DNA sequences with *unidentified bacterium* (5) are registered in the bacteria division of INSD in addition to gene sequences of authentic strains. There are many software packages available for phylogenetic analysis, the most common of which is ClustalX (12)

In the case of phenotypic data, microbial culture collections have progressively disclosed them besides databases embedded in commercial identification kits. However, there is no central digital repository of pheno-

* Correspondence author

typic data yet, and the data for phenotypic analysis are distributed in many sites on the Internet. In addition, the data format and representation are not standardized, making it difficult to integrate data at local sites with those at remote sites. Software packages for numerical analysis of phenotypic data are available both in the public domain and off the shelf, e.g. NTSYSpc (Exeter Software, http://www.exetersoftware.com/cat/ntsyspc/ntsyspc.html).

Polyphasic analysis requires various types of data and data analysis as described above. However, it is not easy use a series of data analysis tools that are appropriate for a certain workflow. For example, what steps are necessary to evaluate the consistency between a dendrogram and a phylogenetic tree? First, we need a comprehensive database for the research subject. Next, we prepare a dataset for numerical taxonomy (NT), run NT and store the result in a file. Then we prepare a dataset for ClustalX, run it and store the result in another file. Finally we compare the two trees either on the screen or in print. The harder the task is, the larger the data set is.

The e−Workbench proposed here offers seamless data processing as follows: designs databases; captures data from distributed databases including INSD; stores and retrieves data in local databases; carries out numerical taxonomy (11); phylogenetic analysis (10); and identification (4); interactively browses the results of the analysis; and simulates the effect of data variability on the analysis. In this paper, we describe mainly applications of such information technologies as eXtensible Markup Language (XML) (7) and Common Object Request Broker Architecture (CORBA) (3) to the e−Workbench. The scheme of the e−Workbench is illustrated in Fig. 1. Biological data are distributed in remote databases, remote files and a local database. Data analysis tools are also distributed in remote sites and the local computer. The structures and specifications of many databases, files and tools are heterogeneous. The e−Workbench makes it possible to integrate these distributed and heterogeneous resources by wrapping heterogeneity by CORBA and using XML files for data transactions.

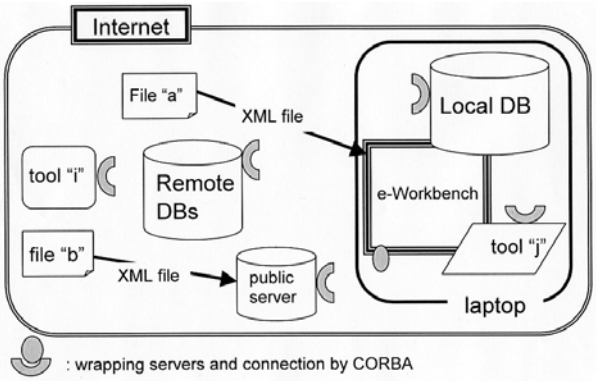

Fig. 1. Scheme of e−Workbench based on XML and CORBA

Wrapping of data resources (databases and data analysis tools) will improve the interoperability and contribute to seamless data processing. Transactions among data resources in XML format will also improve the interoperability and the seamless operation. The user of e-Workbench can capture data and results of analysis from both remote and local resources in a laptop computer and store them locally for subsequent retrieval and analysis .

MATERIALS AND METHODS

XML document for database management

XML has become the norm mainly for data exchange and sharing in bioinformatics, e.g. sequence databases, genome annotation systems, protein databases, pathway databases, system biology and so on (1). This is because XML is more functional than Hyper Text Markup Language (HTML) which is popular in the Web world, and is simpler than Standard Generalized Markup Language (SGML) which is the international standard (ISO 8879) to implement markup into general documents. It is also expected that XML will improve the interoperability of distributed and heterogeneous data resources. It will be very convenient for a community, if the community uses XML documents defined by the common Document Type Definition (DTD) or XML schema (8).

We use an XML document for the master file of the e−Workbench to create a flexible data system with good interoperability. It may be criticism about using XML to develop a robust, common and standard form of document rather than to develop a flexible data system. However, it is an interesting challenge to see whether XML documents are more flexible and expandable, while maintaining interoperability, than a relational database

management system which is the most popular data management system in practice.

We designed the DTD to be suitable for storing the wide range of data items of biological resources, e.g. nomenclature, morphology, physiological data, biochemical data, genetic data, sequence data and even images. There are many ways to set the markup into a document even if we use XML. For example, the markup of scientific names can be done as an *element* of XML:

```
1   <species speciesID ="000123">
    <Scientific name>Escherichia coli
    </Scientific name>
    </species>
```

The set of < −− > and < /−−− > is called a *tag* in XML. It is also possible to formulate it by using an *attribute* as follows:

```
2   <species speciesID ="000123" Scientific
      name ="Escherichia coli"/>
3   <dataitemID ="00003" name ="Scientific
      name" value ="Escherichia coli"/>
```

In the first example, the scientific name is stored as an *element* surrounded by the *tag* of "Scientific name". In the second example, "*Escherichia coli*" is stored as an *attribute* of the XML document. In the third example, both the label of the data item ("Scientific name") and the data value ("*Escherichia coli*") are represented as attributes. There is no golden rule as to what should be represented as elements and what should be represented as *attributes*. We used as few types of *elements* as possible and fully utilized *attributes* for defining data items, *i.e.* we applied the third style to the e−Workbench. This strategy makes the XML document flat and allows the user to easily create and modify data items. In the e−Workbench, all the data items belong to a category and the category contains data item(s) e.g. the category of "Carbon assimilation" contains the data item of "glucose" as a carbon source. In other words, the e−Workbench is composed of data categories, data items and data values. The DTD of the e−Workbench defines a simple container of the data category and item. For example, slots for the identifier, name, and value of the data item are defined as ID, NAME and VALUE respectively as follows:

```
<! ELEMENT TEST ANY>
<! AZTTLIST TEST>
ID          CDATA #REQUIRED
NAME        CDATA #REQUIRED
VALUE       CDATA #IMPLIED
```

The label of the data category and item is not explicitly described in the DTD. This structure makes the database of the e−Workbench flexible.

Wrapping diverse data resources by CORBA

As the Internet and Web have developed, we are now able to search and browse data for the polyphasic analysis from distributed databases. However, it is necessary to write and keep updating programs that interface with diverse and unstable Web sites. Many useful Web sites are unfortunately designed with little attention to guidelines, standards and interoperability of the data system.

We selected CORBA to avoid the endless effort of having to write custom programs for the interface. We also assumed that the name server will be a central registry of CORBA servers that helps us to locate and utilize useful biological CORBA servers anywhere in the world (2). We actually wrapped such databases and computer programs by CORBA as: a key word search system (Sequence Retrieval System (SRS)) in a DDBJ server; a multiple alignment program (ClustalW (6)) in a DDBJ server and in the client; phylogenetic analysis programs (DNAML and DNAPARS in PHYLIP package, http://evolution.genetics.washington.edu/phylip.html) in a DDBJ server; and in the client machine as well. We defined all classes and types of objects to be manipulated in the system by using the standard Interface Definition Language (IDL). Then we compiled the IDL by ORBacus (http://www.orbacus.com/).

We installed CORBA servers for the data retrieval and analysis in the DDBJ server and also in the other server, a PowerEdge 6300 (Pentium II Xeon 450 MHz × 4) with 20 GB hard disk, in the local area network. CORBA servers for the XML document and local data analysis program are stored in the local client machine.

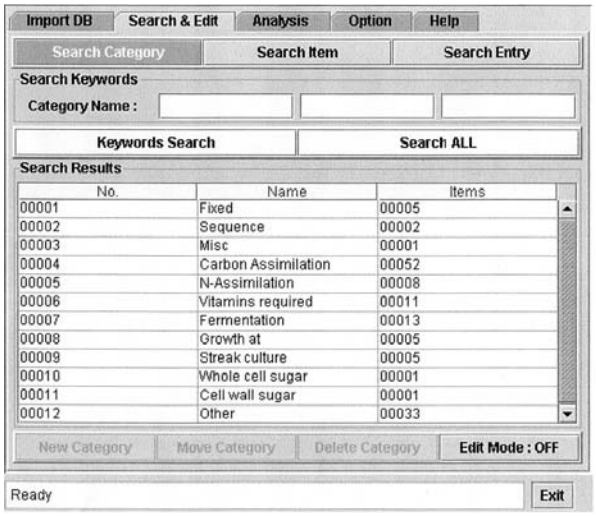

Fig. 2.1. The e－Workbench interface for retrieving, browsing and editing the data category

The user can delete, add and change the data category. Each data category is assigned an identifier as shown in the column of "No.". In the table, the number of data items that belong to the data category is listed in the column of "Items".

Fig. 2.2. The e－Workbench user interface for retrieving, browsing and editing the data item

The data item of "C.Glucose" belongs to the data category of "Carbon assimilation". The user can change the label from "C.Glucose" to "Glucose" in the interface shown here.

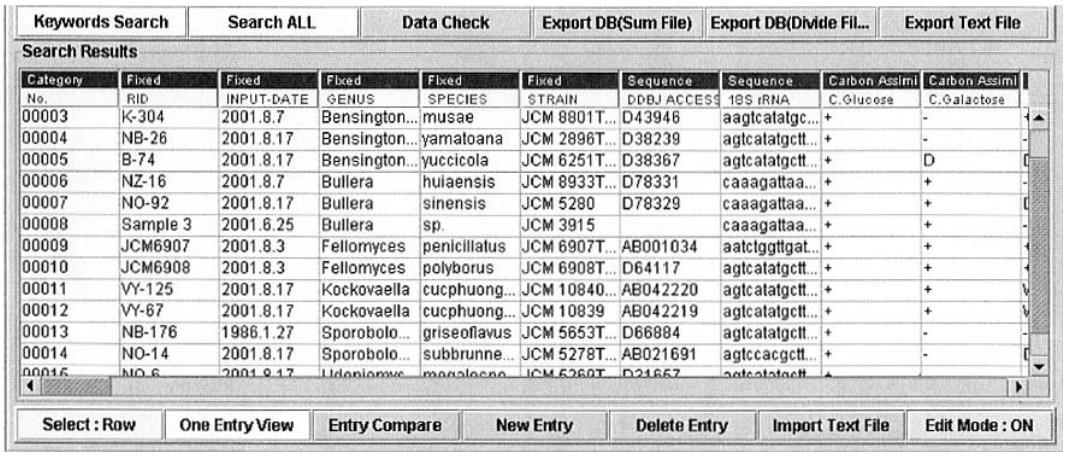

Fig. 2.3. The e－Workbench interface for retrieving, browsing and updating the data value

If a user clicks a row, the row will be highlighted in blue. The user can then update the data in any column in the blue row.

Programming language

The e－Workbench is programmed by Java (JDK 1.1.8) which is one of the most widely used computer languages in bioinformatics (7). Java is platform－independent and the system written in Java is executable on Windows98/Me/NT/2000/XP, Macintosh OS9/X, Linux and UNIX machines. All applications were developed on DIGITAL PC (384 MB memory, 40 GB hard disk) with Windows NT.

RESULTS and DISCUSSION

XML as the container of biological data

The DTD of the XML document was firstly designed and tested for yeasts. Then it was applied to animal cell lines, lactic acid bacteria and *Pseudomonas*. Although the data categories and data items are diverse among the four groups, the common DTD works to construct databases for them. Even binary data can be handled in the e-Workbench. Graphics, figures and spectra are stored not in the XML document but as files either in the remote site or in the client. Instead, the link to the file is stored in the XML document.

With the combination of the DTD and the design of the user interface, the e-Workbench user is able to freely modify data categories, data items and values in a table format as shown in Fig.2.1, Fig.2.2 and Fig.2.3 respectively. Data categories in the table format are introduced in Fig.2.1.

A sample XML document in the back of the user interface is:

```
<CATEGORY ID="00004" NAME="Carbon Assimilation">
    <ITEM ID="00013" NAME="C. Glucose" TYPE=
    "Choice_1" />
```

"Choice_1" in the above parentheses and in Fig.2.2 means that a data value for the data item of Glucose in the data category of "Carbon Assimilation" should be selected from a preset list named "Choice_1". The user of the e-Workbench can also define data types such as codes, Color, Date, Dictionary, Image, Molecule, String, Text and URL. Sample data values in the table format are displayed in Fig.2.3.

Data of the following types can be used for the analysis: Choice, Code and Molecule. Sample results of numerical and phylogenetic analysis using these data are given in Fig.3. The dendrogram in Fig.3 was created in the client (in the local machine). The multiple alignment was done by ClustalW in the DDBJ server and the phylogenetic tree was drawn by a program in the client. In the case of Fig.3, the relationship of strains shown in the dendrogram and the one shown in the phylogenetic tree are consistent with one exception. To see the robustness of the result, the user can repeat the analysis seamlessly after deleting/adding strain(s) and/or data items. Note that the e-Workbench is applicable to any biological objects, if they are described by code data in the same way as microbial data. Concerning the scalability of the system, we could manage data of thousands of strains and hundreds of data items including gene sequences. The scalability of data analysis depends on the hardware, e.g. ClustalW on a large scale PC cluster can process a larger data set than that on a laptop computer.

Wrapping of diverse servers by CORBA and its evaluation

In CORBA, we can use the Interface Definition Language (IDL) and the Internet Inter-ORB Protocol (IIOP) to invoke operations at remote sites from a local program without writing the interface program. The interface with a remote site is automatically generated, if the user compiles the IDL defined and declared by the remote site. Then we can concentrate on programming the application part at the client site without analyzing the structure of the remote site.

On the other hand, CORBA has shortcomings when used in a wide-area network: It is unfriendly to the existing firewall system, as the IIOP requires an additional port number to make the connection between the server and the client program. This requirement is not approved by most system/network administrators because it might degrade the security of their network. The limit of data volume is another issue of CORBA. According to our experiment, the size of transferable data is limited to several kilobytes, so it is difficult to transfer a large transaction such as data of thousands of strains between the client and the server. It is also a problem that CORBA works as a single process. Therefore, we install modules to call CGIs at the remote site in the e-Workbench in addition to wrapping servers by CORBA. Note that we could write custom programs to access the CGI, because we knew the architecture of the remote Web site beforehand. An example of the interoperability between the master file of the e-Workbench and a remote database is introduced in Fig.4.

However, CORBA will be useful even in wide area networks in the future. IIOP is the primary protocol for the ORB compiler but SOAP (Simple Object Access Protocol, http://www.w3.org/TR/SOAP/), which is firewall-friendly, will be selectable for the protocol of

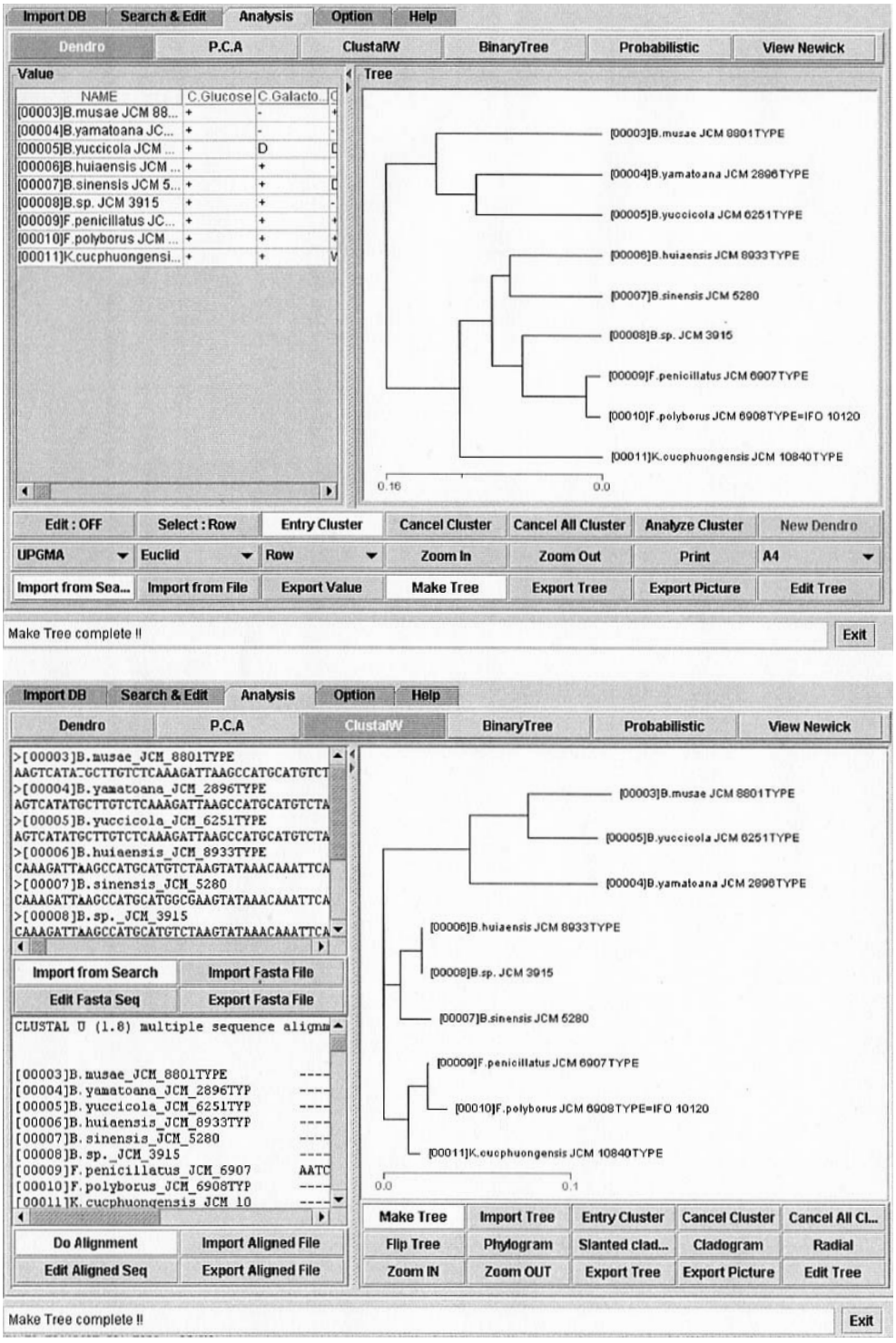

Fig. 3. Sample output of analysis in e-Workbench: dendrogram and phylogenetic tree of the same set of strains

After importing a database and creating a subset for further analysis, the user can apply any analysis by clicking one of the tabs displayed in the second row in the main window, i.e. "Dendro", "P.C.A", "ClustalW", "BinaryTree" and "Probabilistic". In addition, the user can freely go back and forth from "Import DB", "Search & Edit", "Analysis", "Option", to "Help" seamlessly. The seamless operation of multiple data analyses is also realized in e−Workbench. The user can edit the tree by clicking the "Edit Tree" button to prepare a high−quality figure for printing.

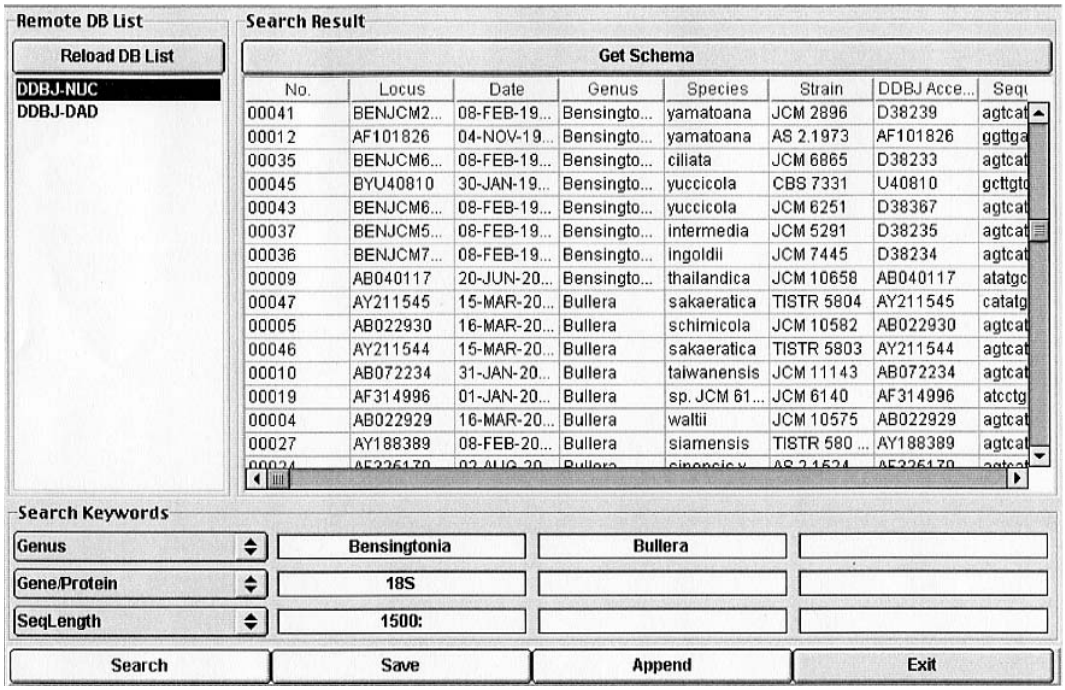

Fig. 4. Import data from the DDBJ server to the local database

In the e−Workbench, the user can connect to a remote database, e.g., the sequence database at the DDBJ server to merge data from the remote site into the local database, even if the data structures are different between the remote and local databases. In the figure, 18s sequences (equal or longer than 1,500 bases) of *Bensingtonia* and *Bullera* strains are retrieved. The data retrieved from DDBJ will be appended to the local database if the "Append" button is clicked.

CORBA. Then CORBA will be indeed a powerful architecture because a computer program autonomously finds and utilizes all the CORBA servers on the Internet.

CONCLUSION

We constructed the e−Workbench for *databasing*, classification and identification of microorganisms by using XML and CORBA. The database is flexible and expandable by fully utilizing *attributes* in XML. The e−Workbench is global to incorporate data from remote sites into a local database, even if their data structures are different. Analytical methods available in the e−Workbench are hierarchical clustering, principal component analysis, phylogenetic analysis, and discriminative and probabilistic identification. The analytical tools are expandable because a computer program of new analytical methods can be implemented as a module to the existing system. All the features of the e−Workbench are applicable to any biological objects from molecules to any organism, provided their data are coded in the same way as microbial data.

The e−Workbench with related public informatics tools and sample databases are available in CD−ROM from us and also downloadable from the menu of "Download e−Workbench" at the server of WFCC−MIRCEN World Data Centre for Microorganisms (http://www.wdcm.org/).

ACKNOWLEDGEMENTS

The authors would like to express their sincere thanks to those who provided test data and encouraged us to develop the e−Workbench: Dr. Takashi Nakase of BIOTEC (Thailand) and Dr. Makiko Hamamoto of the Institute of Physical and Chemical Research (RIKEN) for yeasts; Professor Sanae Okada at Tokyo University of Agriculture (TUA) and Dr. Yoshimi Benno of RIKEN for lactic acid bacteria; Dr. Yoshimasa Kosako of RIKEN and Dr. Masataka Uchino at TUA for *Pseudomonas*; and Dr. Hiroshi Mizusawa at National Institute of Health Sciences for animal cell lines. They also acknowledge the

programming skill of Mr. Koji Koorikawa of Software Engineering Co., Ltd. The development of the e−Workbench was partly supported by the project of fundamental research and development for *databasing* and networking bio−resource information in the framework of the Promotion System for Intellectual Infrastructure of Research and Development, Special Coordination Funds for Promoting Science and Technology.

REFERENCES

1. Achard, F., Vaysseix, G. and Barillot E. XML, bioinformatics and data integration. Bioinformatics **17**: 115−125 (2001).
2. Barillot E., Leser U., Lijnzaad P., Cussat−Blanc C., Jungfer K., Guyon F., Vaysseix G., Helgesen C. and Rodriguez−Tome P. A proposal for a standard CORBA interface for genome maps. Bioinformatics **15**: 157−169 (1999).
3. Bolton, F. Pure CORBA, Sams Publishing, Indianapolis (2002).
4. Dunn, G. and Everitt, B. S. Identification and assignment techniques, *In* An introduction to mathematical taxonomy, p. 106−121, Cambridge University Press, Cambridge (1982).
5. Hawksworth D. L. and Kalin−Arroyo, M. T. Magnitude and distribution of biodiversity, *In* Heywood, V. H.(ed.), Global Biodiversity Assessment, p. 107−192, Cambridge University Press, Cambridge (1995).
6. Higgins, D. G., Thompson, J. D. and Gibson, T. J. Using CLUSTAL for multiple sequence alignments. Methods in Enzymology **266**: 383−402 (1996).
7. Maruyama H., Tamura, K., Uramoto, N., Murata, M., Clark, A., Nakamura, Y., Neyama, R., Kosaka, K. and Hada, S. XML and Java: Developing Web Applications, Addison−Wesley, Boston (2002).
8. Maruyama H., Tamura, K., Uramoto, N., Murata, M., Clark, A., Nakamura, Y., Neyama, R., Kosaka, K. and Hada, S. (2) "Working with Schemas: Data types and Namespaces" in XML and Java: Developing Web Applications, p. 259−293, Addison−Wesley, Boston (2002).
9. Medigue, C., Vermat, T., Bisson, G., Viari, A. and Danchin, A. Cooperative computer system for genome sequence analysis. Proc. Int. Conf. Intell. Syst. Mol. Biol. **3**: 249−258 (1995).
10. Miyamoto, M. M. and Cracraft, J. (eds) Phylogenetic Analysis of DNA Sequences, Oxford University Press (1997).
11. Sneath, P. H. A. and Sokal, R. R. Numerical Taxonomy: *In* W.H. Freeman(ed), The Principles and Practice of Numerical Classification, San Francisco (1973).
12. Thompson, J. D., Gibson, J. T., Plewniak, F., Jeanmougin, F. and Higgins, D. G. The Clustal X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **24**: 4876−4882 (1997).
13. Upton, C., Hogg, D., Perrin, D., Boone, M., and Harris, N. L. Viral genome organizer: a system for analyzing complete viral genomes. Virus Res. **70**: 55−64 (2000).
14. Vandamme, Peter, A. R. Polyphasic Taxonomy in Practice: the *Burkholderia cepacia* Challenge. WFCC Newsletter **34**: 17−24 (2002).
15. Yoshida, M., Fukuda, K., and Takagi, T. PNAD−CSS: a workbench for constructing a protein name abbreviation dictionary. Bioinformatics **16**: 169−175 (2000).

微生物多様性研究を支援する電子的ワークベンチ

菅原秀明[1),2)]，田中尚人[1),3)]，宮崎　智[1)]

[1)] 国立遺伝学研究所生命情報・DDBJ 研究センター
[2)] 総合研究大学院大学遺伝学専攻
[3)] 科学技術振興機構

微生物多様性研究用に総合ソフトウエアパッケージ（suite）を開発した．このパッケージによって，ある流れにそってデータ管理，データ解析そしてその評価までを淀みなく繰り返すことができる．機能として，データ処理機能としては，データベースの設計，データの蓄積と検索，数値解析，系統解析ならびに決定木による同定と確率的同定が含まれている．主ファイルにはXML（拡張可能なマークアップ言語）文書を採用して柔軟なデータベース操作を実現した．また，利用者の手元のデータ資源とインターネット上に分散しているデータベースとデータ解析ツールとも淀みなく統合できる．このパッケージを生物系実験室におけるワークベンチになぞらえて，電子的ワークベンチ（e−Workbench）と呼ぶことにする．本パッケージは，Java で記述されており，Windows，Macintosh，Linux および UNIX のいずれの計算機でも実行可能であり，インターネット経由で自由に入手してラップトップコンピュータで利用することができる．