Open Access

BioloMICS Software: Biological Data Management, Identification, Classification and Statistics

Vincent Robert*[,1,2], Szaniszlo Szoke[1,2], Bernard Jabas[1,2], Duong Vu[1], Oussema Chouchen[2], Erik Blom[1,2] and Gianluigi Cardinali[3]

[1]*CBS-KNAW Fungal Biodiversity Centre, Utrecht, The Netherlands*

[2]*BioAware, Hannut, Belgium*

[3]*DipartimentoBiologiaApplicata- Microbiologia, UniversitàdegliStudi di Perugia, Perugia, Italy*

Abstract: The BioloMICS software is briefly described in this application note. BioloMICS is a tool allowing specialized and scientific biological databases to be created to fit the specific needs of researchers working on any organisms (from viruses, bacteria, fungi, plants, insects to animals) by abroad base of users such as taxonomists, ecologists, human, animal or plant pathologists, molecular biologists, pharmacists, industrial researchers, etc. They can all use the system in completely different ways and with different goals in mind. One can create own custom databases without any prior knowledge of either databasing or programming. The system is completely dynamic and can evolve with the needs of the users. A number of tools for data retrieval and analysis are also included in the system. The system also allows linking with many other databases. BioloMICS is proposed as a suite of tools capable of archiving, analyzing and publishing data on local computers and Internet servers.

Keywords: BioloMICS, software, biology, management, algorithms, identification, classification, statistics, analysis, database, Internet publication.

INTRODUCTION

In 1990, while starting a biodiversity study of the yeast flora in wet and dry forests of central Africa, it became obvious that identification using conventional methods was barely possible. Standard physiological tests were tedious, slow, difficult to interpret unambiguously, and the identification of isolates using classical dichotomic keys was extremely difficult. As a consequence, it was decided to develop software to replace the existing keys (Robert *et al.* [1]). Soon afterwards, a miniaturized system using 96-well microplates to perform physiological tests (Robert *et al.* [2]) was created. This was a success, but the number of characteristics that needed to be introduced into the system was growing continuously with the addition of morphological and molecular features (among others). To endow the system with improved scalability it was necessary to develop new and adapted algorithms, along with tools to first retrieve the data, then to manage and to analyze them.

Software, now called BioloMICS, was first developed as a MS-Windows client/server multi-user application (Robert and Szoke [3]). Over the years, It has been developed further and can now be used for retrieval, management, and analysis (e.g., identification, classification and statistical summary) of any biological material or research experiments. In 2000, another software package called BioloMICSWeb (Robert [4]) was introduced to complement the first one and to allow online Internet publication of the data prepared with BioloMICS. It also permitted real polyphasic identifications to be performed online against species or strain databases. A number of taxonomists working on a wide diversity of groups have used the system to produce advanced polyphasic online identification keys (yeasts, www.cbs.knaw.nl/Yeast.htm, Robert *et al.* [5]; *Penicillium,* www.cbs.knaw.nl/Penicillium.htm, Samson and Frisvad [6]; *Phaeoacremonium,* www.cbs.knaw.nl/Phaeoacremonium.htm, Mostert*et al.* [7]; *Fusarium,* www.cbs.knaw.nl/fusarium, O'Donnell *et al.* [8]; bacteria, *Phoma*, *Phytophthora*, *Colletototrichum*, viruses, insects, nematodes and plants, www.q-bank.eu).

Many other biological and/or molecular databases projects such as Global Biodiversity Information Facility (GBIF, www.gbif.org), Encyclopedia Of Life (EOL, www.eol.org), European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI, www.ebi.ac.uk/embl), National Center for Biotechnology Information (Genbank-NCBI, www.ncbi.nlm.nih.gov), DNA Data Bank of Japan (DDBJ, www.ddbj.nig.ac.jp), or Barcoding Of Life Data (BOLD, www.boldsystems.org) and many others as well have been launched during the last few years.

They all have their strong and weak points of course. EMBL-EBI, Genbank-NCBI and DDBJ are very similar and data available from the first one are essentially available from the others as well. The difference lies mainly in the user interface and some of the tools for searching that are available. The core business of the databases is to provide information related to sequences. The latter being the central point of interest. They are connected to other databases like GBIF, Mycobank, Index of Fungi, PubMed, etc. EMBL-EBI

*Address correspondence to this author at the CBS-KNAW Fungal Biodiversity Centre, Utrecht, The Netherlands; Tel: +31 (0)30 25 12 097; Fax: +32 478 28 57 64; E-mail: v.robert@cbs.knaw.nl and Genbank-NCBI are absolutely not adapted for the storage, retrieval and analyses of data on species or strains. They should be used for data related to DNA or protein sequences.

The BOLD system is aimed at archiving the so-called universal barcode of life (mainly COX1 gene). The main idea is that there is to archive COX1 sequences of all species on Earth. A few restricted additional data relating to the analyzed specimens are also recorded but are quite sparse. Scientists cannot change the structure of the databases to fit with the needs and requirements of their organisms of interest. Search engines and identification schemes are essentially aimed at sequence alignments.

EOL and GBIF are more generalized databasing systems that aim to record all species on Earth and to "rapidly, openly and freely delivering primary data about biodiversity to everyone in the global community, using digital technologies. Another part is ensuring that the primary data being collected today are stored in such a way that they will remain accessible to future generations" (from www.gbif.org). GBIF could be an interesting centre point of information but the data are extremely fragmented and very little scientifically usable data are available apart from species names and geographic locations.

BioloMICS is a flexible tool allowing specialized and scientific databases to be created to fit the specific needs of researchers working on any organism (from viruses, bacteria, fungi, plants, insects to animals) using a wide array of scenarios. Users can be taxonomists, ecologists, human, animal or plant pathologists, molecular biologists, pharmacists, industrial researchers, etc. The system is adaptable to use in very different ways and for various goals. It can also link to many other databases such as those mentioned above. BioloMICS is proposed as a suite of tools capable of archiving, analyzing and publishing data on local computers and Internet servers.

MANAGING ANY BIOLOGICAL DATA

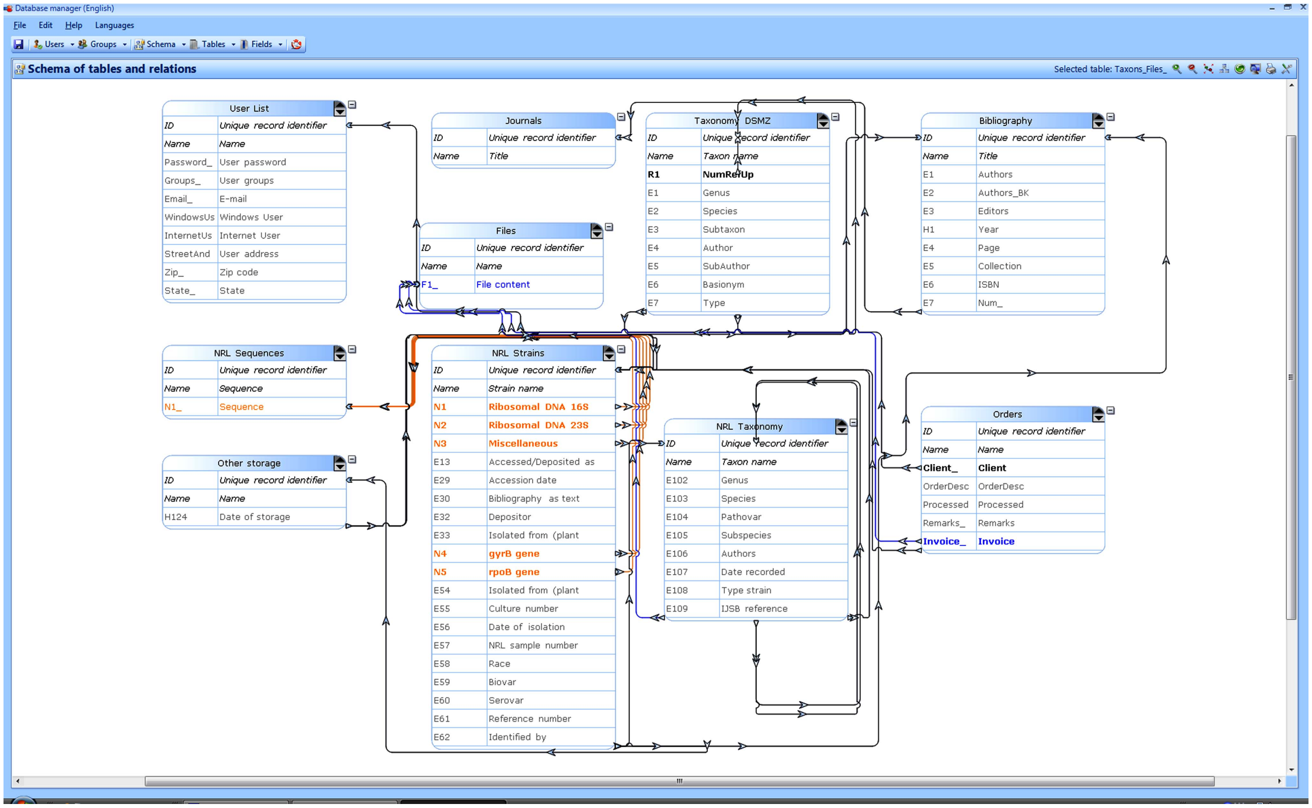
Most researchers have no time to spend on the creation of databases and the necessary database-creation software. Therefore, BioloMICS was developed as a system that allows any users to create their own custom databases without prior knowledge of databasing or programming. MySQL, Oracle or Microsoft SQL Server (coming in the near future) databases can be used to store the information, this allows great flexibility and exportability.

Tables

Database administrators can add as many tables as needed and all tables can contain as many fields as wanted. The structure of the table is completely free, dynamic and relational (Fig. **1**). However, each table contains a number of header fields that are always present and cannot be removed by the administrators. The header fields are used to store important data related to the records of a given table such as, read, write and deletion rights, record creator, dates of creation and last update, locking information and other details.

Field and Characters

Administrators choose the type of fields they need to store their data and define a few parameters related to them,


Database manager (English)
File
Edit
Help
Languages
Users
Groups
Schema
Tables
Fields
Schema of tables and relations
Selected table: Taxons_Files_

User List

ID	Unique record identifier
Name	Name
Password_	User password
Groups_	User groups
Email_	E-mail
WindowsUs	Windows User
InternetUs	Internet User
StreetAnd	User address
Zip_	Zip code
State_	State

Journals

ID	Unique record identifier
Name	Title

Taxonomy DSMZ

ID	Unique record identifier
Name	Taxon name
R1	**NumRefUp**
E1	Genus
E2	Species
E3	Subtaxon
E4	Author
E5	SubAuthor
E6	Basionym
E7	Type

Bibliography

ID	Unique record identifier
Name	Title
E1	Authors
E2	Authors_BK
E3	Editors
H1	Year
E4	Page
E5	Collection
E6	ISBN
E7	Num_

Files

ID	Unique record identifier
Name	Name
F1_	File content

NRL Sequences

ID	Unique record identifier
Name	Sequence
N1_	Sequence

NRL Strains

ID	Unique record identifier
Name	Strain name
N1	**Ribosomal DNA 16S**
N2	**Ribosomal DNA 23S**
N3	**Miscellaneous**
E13	Accessed/Deposited as
E29	Accession date
E30	Bibliography as text
E32	Depositor
E33	Isolated from (plant
N4	**gyrB gene**
N5	**rpoB gene**
E54	Isolated from (plant
E55	Culture number
E56	Date of isolation
E57	NRL sample number
E58	Race
E59	Biovar
E60	Serovar
E61	Reference number
E62	Identified by

Other storage

ID	Unique record identifier
Name	Name
H124	Date of storage

NRL Taxonomy

ID	Unique record identifier
Name	Taxon name
E102	Genus
E103	Species
E104	Pathovar
E105	Subspecies
E106	Authors
E107	Date recorded
E108	Type strain
E109	IJSB reference

Orders

ID	Unique record identifier
Name	Name
Client_	**Client**
OrderDesc	OrderDesc
Processed	Processed
Remarks_	Remarks
Invoice_	**Invoice**

Fig. (1). Database management module allowing creating tables, adding new fields and defining the rights of groups and users.

such as the default values, the states, the labeling, the algorithms that should be used when performing comparisons (i.e. identifications, classifications, etc.), the tolerance (for fuzzy comparisons), the weighting, etc. A number of wizards are used to guide the administrators in setting the basic settings of the system. The generic list of field types that can be used is given in Table **1**.

Using the right type of fields is essential in such a system since it guides the way data are stored and can be used and analyzed. A short and trivial example demonstrates how important the design of a database and the choice of characters and fields can be. Size data such as the length or the width of an organ for example can be stored as pure text such as small, medium or large or by providing the original values. It can also be recorded as the average of the observed measurement. One can also keep the minimum and the maximum and the average, the median, the percentiles, the variance, etc. Storing everything as text is only useful if one wants to read the information. Discretization of measurements data can have some obvious advantages but could lead to loss of useful information for future analyses that would require more detailed data (Grzymala-Busse [9]; Legendre and Legendre [10]; Robert [11]).

To compare, to analyze and to obtain a number of statistics, it is essential to store the information using the most appropriate method and to keep data in a state as close as possible to the original value. Algorithms can also play a huge role in the way data can be retrieved and compared. The following example shows how important the algorithms can be. Two "objects" A (individual) and B (group of organisms) have to be compared on the basis of the length of some cells. The first one has cells ranging from 4 to 5 while the second "object", a species concept, has cells ranging from 3 to 8. In a first identification algorithm, one could consider that 4 and 5 are within the range of the species represented by object B (3 and 8) and the similarity is 100%. Another algorithm could be used when data are considered imprecise and objects could be compared on the basis of their average or medians, 4.5 and 5.5. Similarity could then be 1-[(5.5-4.5)/5.5] equals to 81.8%. A third way that could be used if both objects A and B were species descriptions established on statistically solid observations, would compare the proportion of overlap between both objects. The resulting similarity would then be 20% obtained from the following formula: (5-4)/(8-3). Many more algorithms could of course be used but with this simple example, one can easily understand that the selection of algorithms is crucial and must be done with great care. Each characteristic, field

Table 1. Field Types that can be Used to Store the Information

	Type of Field	Number of Algorithms	Weighting	Tolerance
1.	Continuous data, range (min, low percentile, high percentile, max), e.g., the size of the cells such as (5) 5.5 - 7 (8).	5	Yes	Yes
2.	Continuous data, single value, e.g., a pH value of 9	1	Yes	Yes
3.	Discrete data, single state, e.g., presence or absence of a property	1	Yes	Yes
4.	Discrete data, multiple states, e.g., colour of the colony (not red, not green, white, cream, not brown, not black, etc.)	3	Yes	Yes
5.	96- or 384-well microplate, continuous data (absorbance values), e.g., microplate containing user-defined tests, data are stored as continuous values	5	Yes	Yes
6.	96- or 384-well microplate, discrete data (-, +, weak, etc), e.g., microplate containing user-defined tests, data are stored as discrete values.	1	Yes	Yes
7.	Free-array data, continuous data, e.g., array of any size of continuous data.	5	Yes	Yes
8.	Free-array data, discrete data, e.g., array of any size of discrete data.	1	Yes	Yes
9.	Molecular weights from electrophoretic analysis, e.g., RAPD, RFLP, PCR, AFLP.	8	Yes	Yes
10.	Time series and chromatograms.	3	Yes	Yes
11.	DNA sequences.	6	Yes	Yes
12.	Protein sequences.	6	Yes	Yes
13.	GIS location, latitude, longitude, altitude and precision.	5	Yes	Yes
14.	Links to other BioloMICS records in the same or in other tables. Links to other BioloMICS databases are also possible even on a different server.	2	Yes	Yes
15.	Text, administrative data.	-	No	No
16.	Date and time.	-	No	No
17.	URL	-	No	No
18.	Link to pictures and others files.	-	No	No
19.	Link to bibliographic information, e.g. bibliographic references such as articles, books, etc.	-	No	No
20.	Link to taxonomic data, e.g. species name, genus, family, synonyms, type specimens, etc.	-	No	No

and associated context of data retrieval should have their own settings (algorithm, weight or importance factor and tolerance level).

Fields/characters can be added, modified (e.g., add/delete states for a given character, change the title of a character) or deleted at any time by the administrators of the database. Fields can be re-ordered, re-organized and grouped by the end-users.

BioloMICS Software allows the basic management of databases, such as the addition, modification, or deletion of records and tables. Data can also be imported from, and exported to, a variety of formats (text, tab delimited, html, XML, MS-Excel or MS-Word, Fasta, etc.).

Records can be linked with the taxonomic database at levels ranging from kingdom to species, subspecies, variety or form. All relevant taxonomic information (e.g., synonymy, basionym, anamorph-teleomorph connections for fungi, type strain, bibliography) are available. When the classification or nomenclature is changed, records are automatically updated. Bibliographic databases can also be created and managed from within the software. These can be queried and linked to external databases such as PubMed, for example. Pictures can be displayed and manual or automated measurements (width, length, perimeter and surface) of structures can be performed on any electronic images, which may be stored in any picture formats. Basic descriptive statistics and information (number of observations, mean, minimum, maximum, median, percentiles, standard deviation, and variance) can be obtained and be stored in fields.

Users, Groups, Accession Rights and Data Protection

Each user must be registered in the system and affiliated with one or several of the groups created by the administrators in BioloMICS. Groups receive read, write and deletion rights on tables, fields and records. Depending on the groups they belong to, users will be able to see, modify and delete data. This system is highly flexible and allows a fine level of data protection.

Because end-users can make mistakes or change their mind, we have also introduced an advanced tracking system that records all changes made to the database by any user at a certain date and time. History of changes can be queried and reverted by administrators.

DATA RETRIEVAL TOOLS

A number of tools have been developed allowing users to easily capture data direct from the original equipment (sequencer, microplate reader, electrophoresis) or analyze images and extract useful information.

Sequence Data

Nowadays, DNA and protein sequences are essential for the understanding and the positioning of strains or specimens and species. While a few decades ago, sequencing proteins and later DNA would take years and could lead to a Nobel price, it is now a routine task in any laboratory dealing with biological material. Sanger sequencing is still widely used and modern laboratories can produce large amounts of sequences daily. To handle, store and track such data, we have created a laboratory management information system (LIMS (Fig. **2**)) module that stores and analyzes all data related to Sanger sequencing.

DNA extracts from a strain or specimen are amplified and the resulting information is stored in a PCR reaction table. PCR products are then used to produce a sequencing

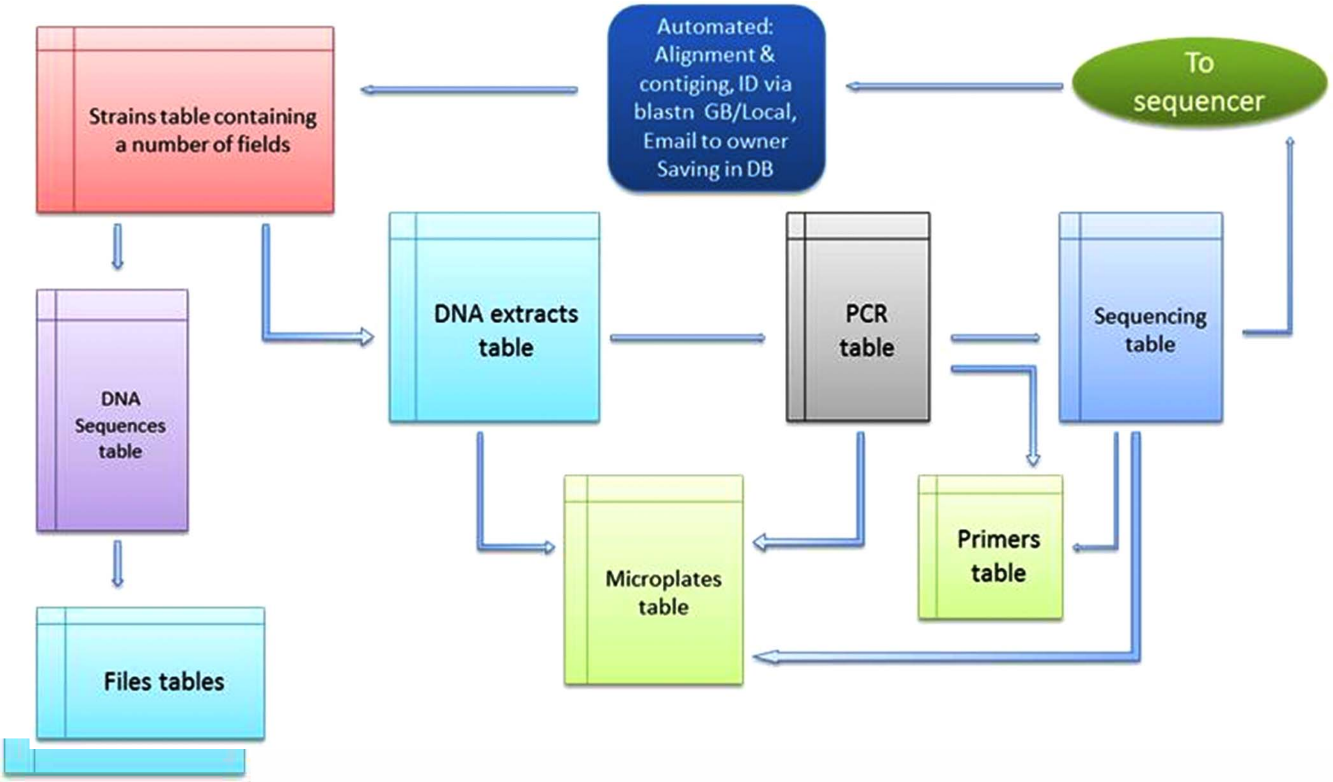

Strains table containing a number of fields
Automated: Alignment & contiging, ID via blastn GB/Local, Email to owner Saving in DB
To sequencer
DNA Sequences table
DNA extracts table
PCR table
Sequencing table
Files tables
Microplates table
Primers table

Fig. (2). Database structure of the LIMS module related to Sanger sequencing.

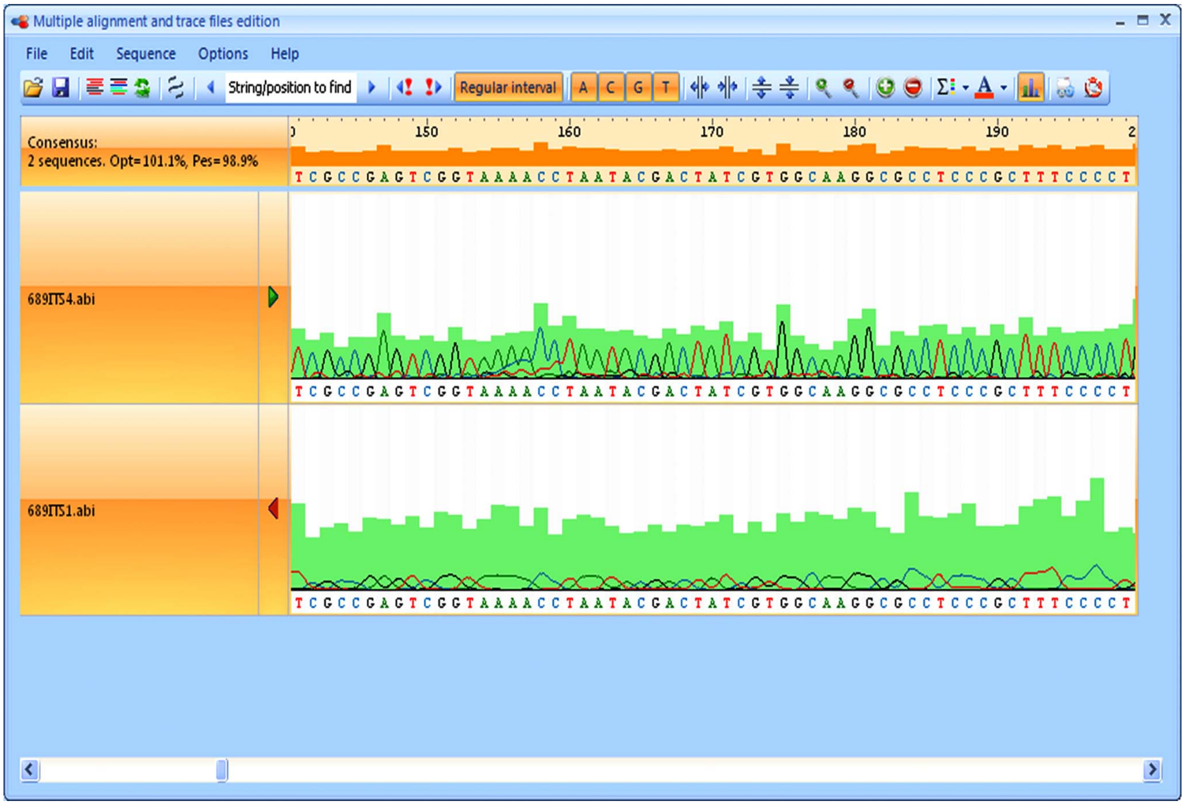

Multiple alignment and trace files edition
File
Edit
Sequence
Options
Help
String/position to find
Regular interval
A
C
G
T
Consensus:
2 sequences. Opt=101.1%, Pes=98.9%
150
160
170
180
190
TCGCCGAGTCGGTAAAACCTAATACGACTATCGTGGCAAGGCGCCTCCCGCTTTCCCCT
689ITS4.abi
TCGCCGAGTCGGTAAAACCTAATACGACTATCGTGGCAAGGCGCCTCCCGCTTTCCCCT
689ITS1.abi
TCGCCGAGTCGGTAAAACCTAATACGACTATCGTGGCAAGGCGCCTCCCGCTTTCCCCT

Fig. (3). Multiple alignment and trace files editions.

reaction and the final product is send to the sequencer. Trace files (ABI files) are produced by the sequencer for both strands (forward and reverse) of the DNA molecule under study. The software automatically aligns both trace files (including the reverse complementing of one of the traces) and removes unreliable bases at both ends of the traces using several algorithms. Finally, it produces a consensus sequence (Fig. **3**). The latter is then compared against a local reference sequence database and/or against Genbank. A report is then saved in an identification table and an email is send to the owner of the task as an identification and quality report. The assembly of the trace files and the consensus are saved in what is called a "contig" file. Consensus sequences and the "contigs" are then saved into the sequence and file tables respectively along with other information on the sequence. Curators can then select "contigs" that require further attention and editing.

This LIMS tool is completely embedded in the software and all data associated with sequences are stored in the tables of the database and available at any time for quality control and other analyzes.

A number of tools applicable to sequences are also available, such as: batch pairwise sequence alignments against Genbank or local files, multiple sequences alignment tool, exportation and importation of sequences to/from Fasta format or to/from Genbank, statistics on sequences (length, number of bases and frequencies), DNA to protein conversion tool and searching tools to find short sequences.

Electrophoresis Data

In many laboratories, electrophoresis methods are still used to assess the similarities and differences between strains or individuals. Because they are cheap, fast and easy to implement, electrophoresis methods will continue to be used widely for a long time. However, it is usually quite difficult and tricky to compare results obtained from different experiments. We have therefore implemented a complete tool to semi-automatically analyze original pictures, edit the results and produce a list of molecular weights that can be compared to updatable reference databases (Fig. **4**). Algorithms for the automated detection of lanes and bands have been incorporated in the software allowing fast, precise and efficient data retrieval, even with difficult and "smiling" gels.

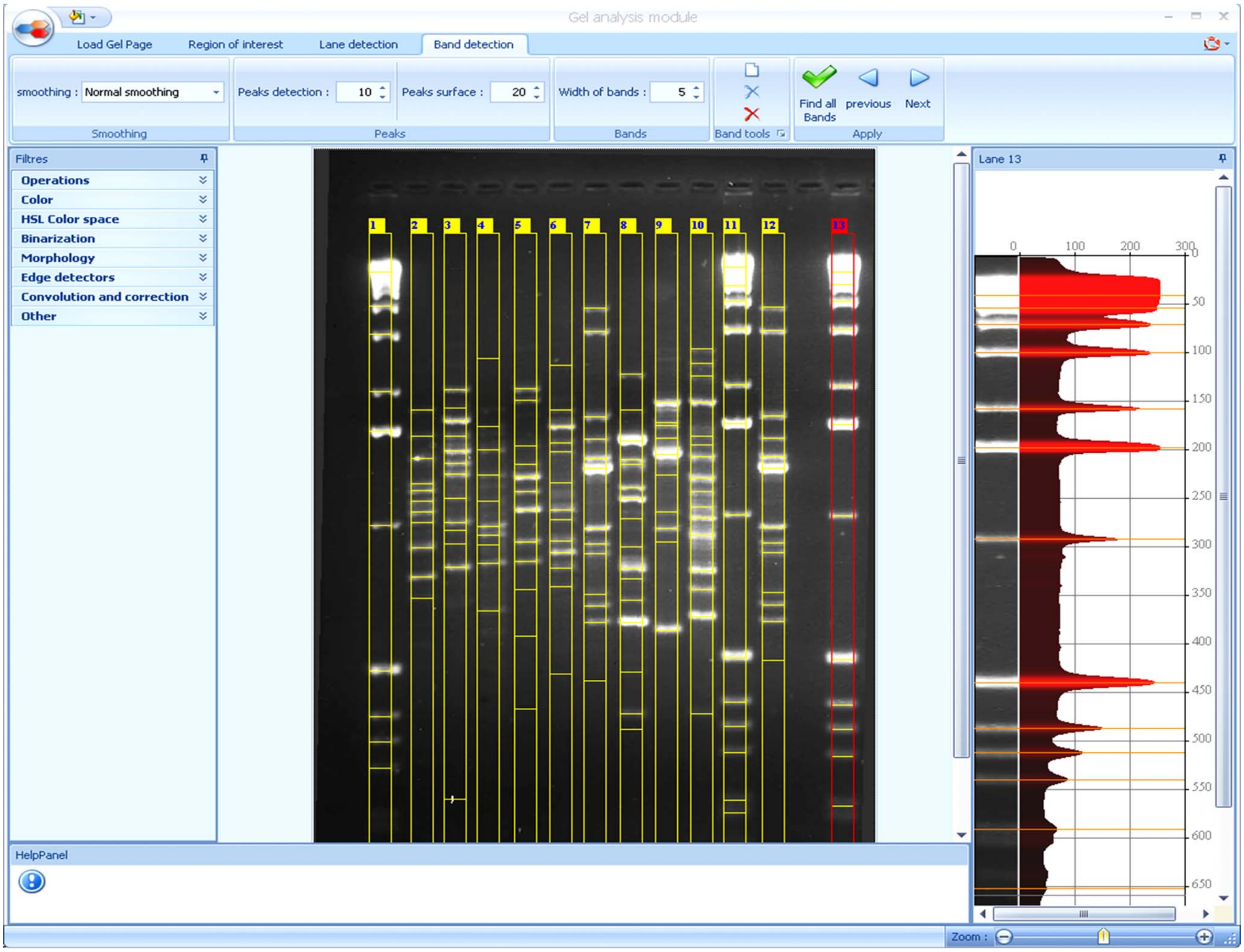
Morphological Data

Many taxonomists or researchers still need to measure structure directly from pictures taken under the microscope for example. We have created a tool allowing them to manually or automatically retrieve objects present in an image and measure their length, width, perimeter and surface area. A number of basic statistics such as minimum, maximum, average, median, percentiles, standard deviation, variance, kurtosis and skewness are automatically provided. These statistics or raw original values can then be easily saved in the database for further analyses.

Geographic Data

The geographic-data viewer allows maps (shape files and Google maps) to be displayed. Records can be plotted to create distribution maps that can be exported to the clipboard or to Google Earth.

A tool based on a web service provided by GeoNames (http://www.geonames.org/) was also implemented in the



smoothing : Normal smoothing

Peaks detection : 10

Peaks surface : 20

Width of bands : 5

Fig. (4). Module for automated detection of lanes and bands and their editing.

system to retrieve latitude and longitude data and associate them with records of strains or species, for example (Fig. **6**).

Programming the Software

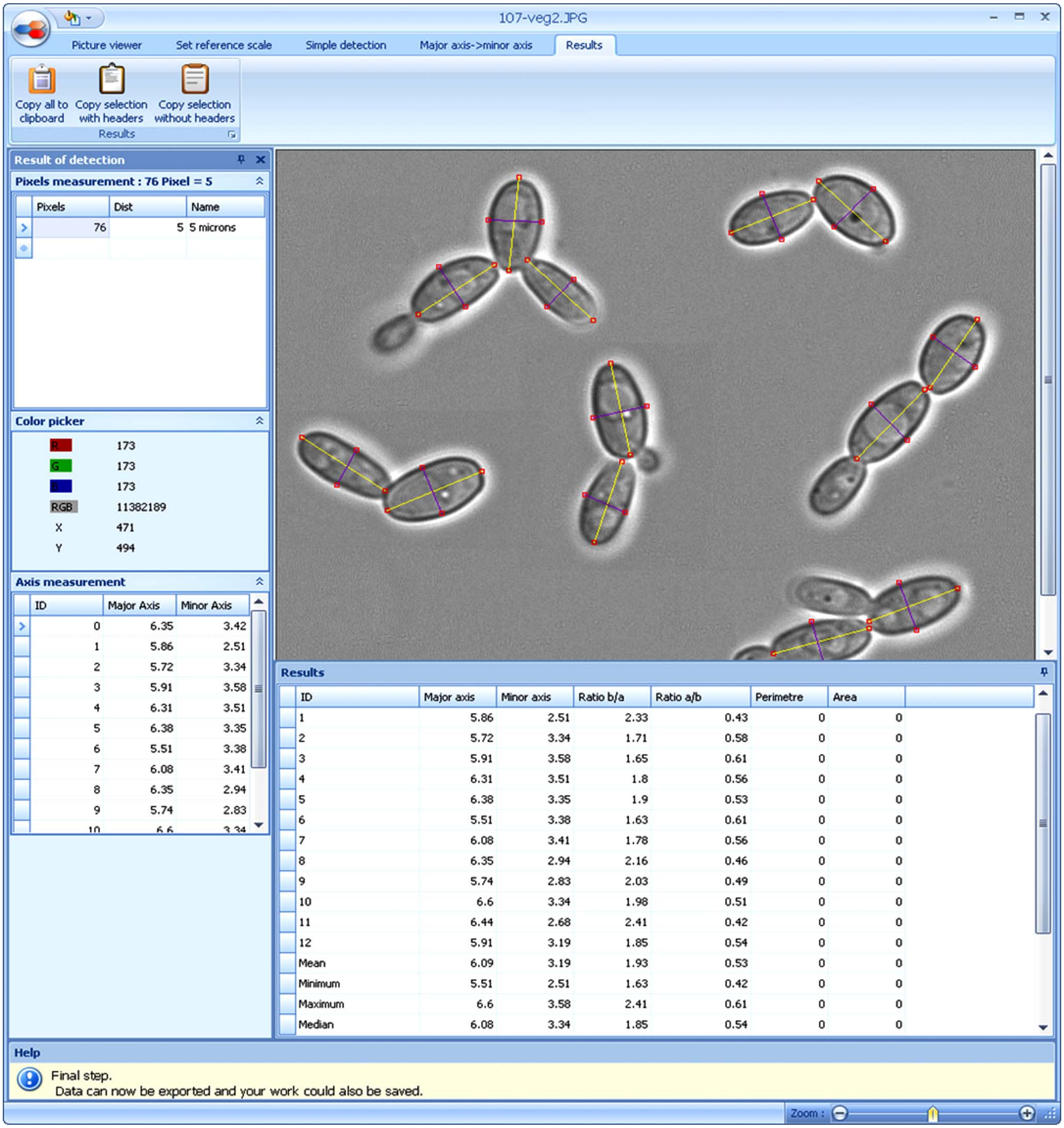
Because users of software always have very specific whishes that are impossible to foresee and since they often want to automate a number of actions related to their databases, a programming tool was implemented within BioloMICS (Fig. **7**). End-users can write scripts in Visual Basic.Net or in C Sharp.Net and include their code, algorithms and special functions in the menus of the main window of the software. The code will be compiled dynamically and executed as if it were hard coded in the software. Source code written in MS-Visual Studio 2008 or 2010 can be compiled. We also provide a complete source code editor including intellisense, code formatting, functions and parameters suggestions from the framework as well as specific functions from BioloMICS. A number of code examples are also provided showing how to access, modify or delete data and automate a number of actions.

Code written by end-users could be incorporated into the original code of the software if it is of general interest and, of course, with the agreement of the owner of the code.

Miscellaneous Tools

Of course, additional tools and functions are available and some are listed below:

- Importation of data module
- Templates creation module
- Numerous export functions
- Microplate reader connection and capture of absorbance values to produce growth curves (for example)
- Computation of basic statistics
- Advanced statistics module (under development)
- Multiple pictures viewer and pictures editor (Fig. **5**)
- Record compilation allowing merging several records into a single one



Result of detection

Pixels measurement : 76 Pixel = 5

Pixels	Dist	Name
76	5	5 microns

Color picker

R	173
G	173
B	173
RGB	11382189
X	471
Y	494

Axis measurement

ID	Major Axis	Minor Axis
0	6.35	3.42
1	5.86	2.51
2	5.72	3.34
3	5.91	3.58
4	6.31	3.51
5	6.38	3.35
6	5.51	3.38
7	6.08	3.41
8	6.35	2.94
9	5.74	2.83
10	6.6	3.34

Results

ID	Major axis	Minor axis	Ratio b/a	Ratio a/b	Perimetre	Area
1	5.86	2.51	2.33	0.43	0	0
2	5.72	3.34	1.71	0.58	0	0
3	5.91	3.58	1.65	0.61	0	0
4	6.31	3.51	1.8	0.56	0	0
5	6.38	3.35	1.9	0.53	0	0
6	5.51	3.38	1.63	0.61	0	0
7	6.08	3.41	1.78	0.56	0	0
8	6.35	2.94	2.16	0.46	0	0
9	5.74	2.83	2.03	0.49	0	0
10	6.6	3.34	1.98	0.51	0	0
11	6.44	2.68	2.41	0.42	0	0
12	5.91	3.19	1.85	0.54	0	0
Mean	6.09	3.19	1.93	0.53	0	0
Minimum	5.51	2.51	1.63	0.42	0	0
Maximum	6.6	3.58	2.41	0.61	0	0
Median	6.08	3.34	1.85	0.54	0	0

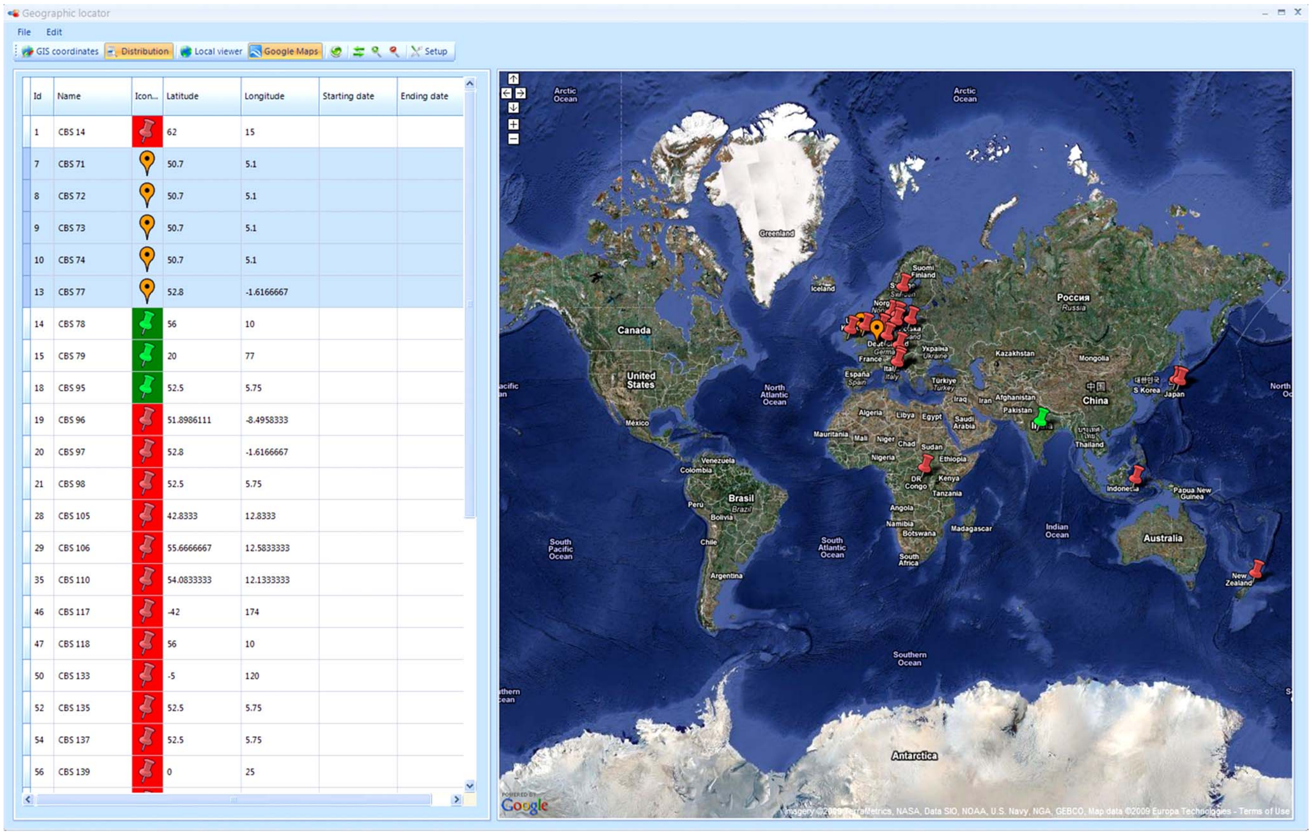
Help

Final step.
Data can now be exported and your work could also be saved.

Fig. (5). Module for automated and manual detection of objects and measurements.

SEARCHING AND DATA ANALYSIS

As discussed above, the analysis of stored data is of primary importance and should be possible using a variety of tools in response to a large panel of users having different goals in mind. Therefore, both basic searching tools and advanced multiple entry keys(MEK) have been developed to allow querying any type of field using the best possible method. For example, a field containing size data, such as from 5-10, could be searched in up to five different ways, while DNA sequences can be aligned using six different algorithms. This functionality allows users, for example, to select records with a given set of properties or to identify an unknown organism at the species level. Advanced and complex queries containing questions separated by AND, OR, NOT and using brackets can be performed.



Id	Name	Icon...	Latitude	Longitude	Starting date	Ending date
1	CBS 14	🟥	62	15		
7	CBS 71	🟧	50.7	5.1		
8	CBS 72	🟧	50.7	5.1		
9	CBS 73	🟧	50.7	5.1		
10	CBS 74	🟧	50.7	5.1		
13	CBS 77	🟧	52.8	-1.6166667		
14	CBS 78	🟩	56	10		
15	CBS 79	🟩	20	77		
18	CBS 95	🟩	52.5	5.75		
19	CBS 96	🟥	51.8986111	-8.4958333		
20	CBS 97	🟥	52.8	-1.6166667		
21	CBS 98	🟥	52.5	5.75		
28	CBS 105	🟥	42.8333	12.8333		
29	CBS 106	🟥	55.6666667	12.5833333		
35	CBS 110	🟥	54.0833333	12.1333333		
46	CBS 117	🟥	-42	174		
47	CBS 118	🟥	56	10		
50	CBS 133	🟥	-5	120		
52	CBS 135	🟥	52.5	5.75		
54	CBS 137	🟥	52.5	5.75		
56	CBS 139	🟥	0	25		

Fig. (6). Geographic distribution module using Google maps or shape files.

or fields at the same time using agglomerative clustering methods. Objects are compared with one another on the basis of a selection of characters. The obtained distance matrix is later represented as a phenotypic tree using one of the selected agglomerative clustering methods. The characters are first normalized and reduced, then correlated with one another to obtain a correlation matrix that can be displayed in a tree-like representation using one of the available agglomerative clustering methods. The result is displayed as a double tree (Fig. **9**). The first is a vertical tree showing the relationships between the objects/records. The second tree is horizontal and shows the groups of characters/fields that are positively (or negatively, depending on the display method used - 'positive' or 'negative' clustering) correlated. Between the two trees a colored "heat map" shows the states of the characters/fields for the different objects/records. With this method, it is easier to isolate characters or groups of characters that are associated with some groups of objects. It is also possible to infer possible relations between different types of criteria, for example the pathogenicity and a given physiological feature or the activity of a gene on a microarray.

Computation of Global Similarity Coefficient

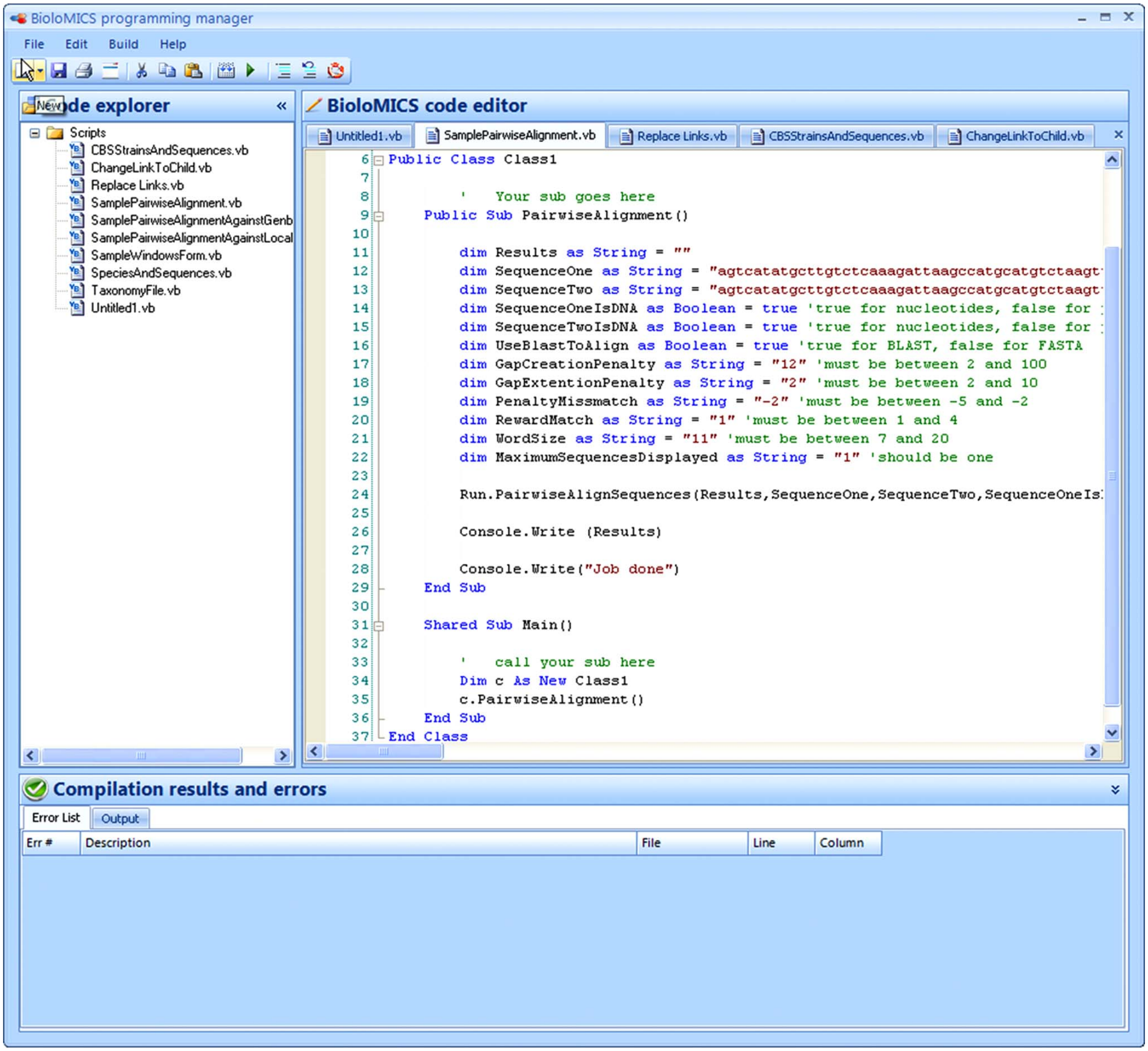
When comparing several records with one another, BioloMICS computes local similarity coefficients for the characters that have been included in the analysis (characters can be individually included or excluded from the analysis). Then a global similarity coefficient is computed by the following formula (Gower [12]):

$$S_{jk} = \frac{1}{\sum_{i=1}^{n} W_i} \sum_{i=1}^{n} S_i \, W_i$$

where, S_{jk} is the global similarity coefficient for the comparison of records j and k; S_i is the local similarity coefficient for character I; W_i is the weight of character I; n is the number of characters accounted in the similarity comparison.

PUBLICATION OF DATA

Scientists creating large biological databases want to use them as an archival tool or to keep track of all data related to their experiments or objects of studies (strains, species, etc). They also want to use them for a number of investigations and comparisons. Finally, they also want to publish their data *via* the Internet; the Web version of BioloMICS can do that for them. Again, they don't need to do any programming nor require anyone to do it for them. They just need to define the tables, fields and records that can be viewed, and modified on their Website. They can use the Windows based version of BioloMICS as a dynamic portal engine and content-management system. This means that they can create, modify and delete web pages and menus for their website and communicate with Internet users of their database(s).



```
Public Class Class1

        '   Your sub goes here
    Public Sub PairwiseAlignment()

        dim Results as String = ""
        dim SequenceOne as String = "agtcatatgcttgtctcaaagattaagccatgcatgtctaagt
        dim SequenceTwo as String = "agtcatatgcttgtctcaaagattaagccatgcatgtctaagt
        dim SequenceOneIsDNA as Boolean = true 'true for nucleotides, false for
        dim SequenceTwoIsDNA as Boolean = true 'true for nucleotides, false for
        dim UseBlastToAlign as Boolean = true 'true for BLAST, false for FASTA
        dim GapCreationPenalty as String = "12" 'must be between 2 and 100
        dim GapExtentionPenalty as String = "2" 'must be between 2 and 10
        dim PenaltyMissmatch as String = "-2" 'must be between -5 and -2
        dim RewardMatch as String = "1" 'must be between 1 and 4
        dim WordSize as String = "11" 'must be between 7 and 20
        dim MaximumSequencesDisplayed as String = "1" 'should be one

        Run.PairwiseAlignSequences(Results,SequenceOne,SequenceTwo,SequenceOneIs

        Console.Write (Results)

        Console.Write("Job done")
    End Sub

    Shared Sub Main()

        '   call your sub here
        Dim c As New Class1
        c.PairwiseAlignment()
    End Sub
End Class
```

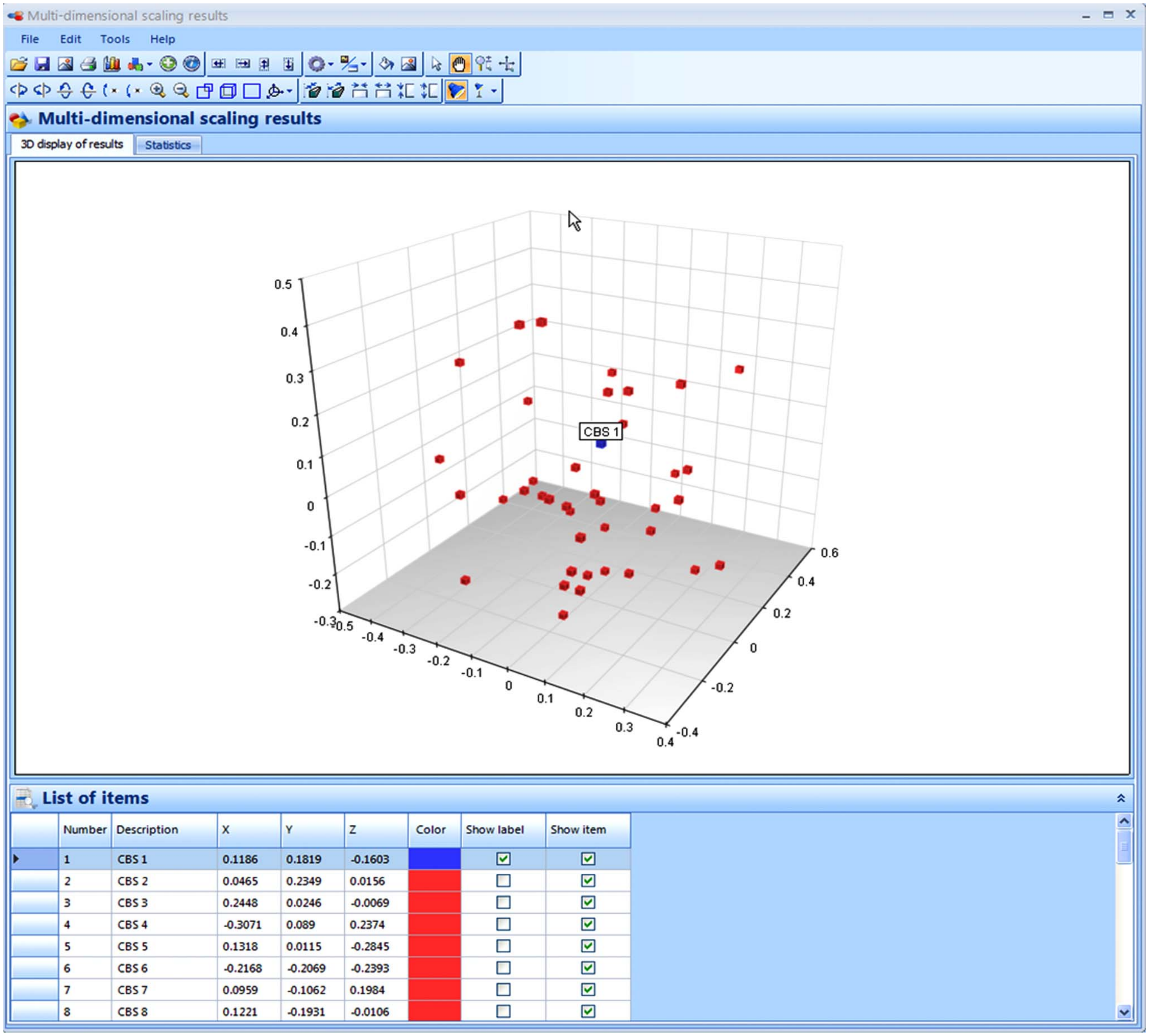
Fig. (7). Visual Basic.Net and C #.Net programming module allowing users to personalize the software and include additional self-designed functions.

Polyphasic identification and classification modules are available and permit a wide range of comparisons. Several two- or three-dimensional graphical displays are available to represent the results of the combination of methods based on similarity or correlation and clustering techniques (see Figs. **8**, **9** as examples). Comparison methods can be used singly or in combination to take the best of the different techniques. For example, one can first perform a preliminary selection of records based on a MEK query, then compare or identify an unknown against the selected set of records, then apply a divisive clustering method on the best matching records and finally analyze by agglomerative clustering some of the groups obtained in the previous steps and draw dendrogram or 3D display as shown in Fig. (**8**).

Another type of comparison available is what we call "Functional Analysis". This is a method allowing the grouping of objects or records and of their characters

As default options, BioloMICS provides a number of tools, such as:

- Basic and advanced tools to search the databases
- Deposit and online modification of data
- Cart system allowing Internet users to buy items online and obtain feedbacks from the sales manager of the database(s)
- Pairwise sequence alignment tool against local databases



	Number	Description	X	Y	Z	Color	Show label	Show item
▶	1	CBS 1	0.1186	0.1819	-0.1603	🟦	☑	☑
	2	CBS 2	0.0465	0.2349	0.0156	🟥	□	☑
	3	CBS 3	0.2448	0.0246	-0.0069	🟥	□	☑
	4	CBS 4	-0.3071	0.089	0.2374	🟥	□	☑
	5	CBS 5	0.1318	0.0115	-0.2845	🟥	□	☑
	6	CBS 6	-0.2168	-0.2069	-0.2393	🟥	□	☑
	7	CBS 7	0.0959	-0.1062	0.1984	🟥	□	☑
	8	CBS 8	0.1221	-0.1931	-0.0106	🟥	□	☑

Fig. (8). Three dimensional display of the relative positions of objects (could be strains or species for example) after a multiple dimensional scaling analysis.

- Polyphasic identification tool including any combination of morphological, physiological, ecological or molecular (and many more). Tree-building algorithms showing the relative position of the object to be identified

This Internet publication tool is a major resource not only for the end-users who can electronically access, use and compare data but also for the creators of biological databases. The latter can show their work better and communicate with the outside allowing "marketing" their know-how and attracting attention, collaborators and funding for future research projects (see a few published examples based on the previous or the new version of the software: Crous *et al*. [13]; Mostert [7]; O'Donnell *et al*. [8]; Robert *et al*. [5]; Samson and Frisvad [6]; (Fig. **10**)).

FUTURE CHALLENGES

Virtually everything in bioinformatics is challenging. Bioinformatics is a relatively new "science" that is a mixture of biology and informatics which means that it is the marriage of water and fire.

Second and third generation sequencing will bring incredible challenges to bioinformaticians. Not only, will we have to handle, store and analyze complete genomes, which is not completely new, but we will have to do that for very large numbers of records. As obtaining complete genomes, even for eukaryotes, becomes cheaper and easy to obtain and

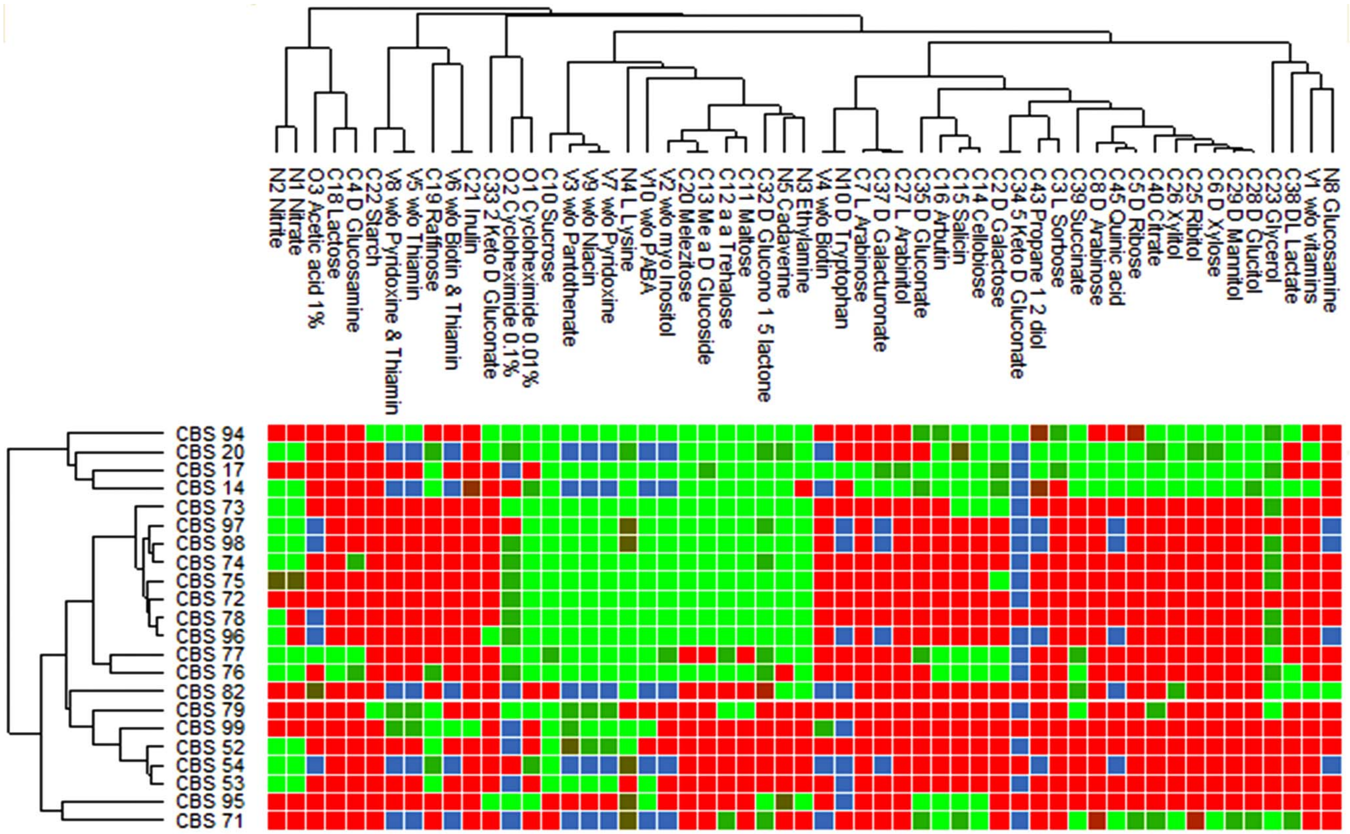


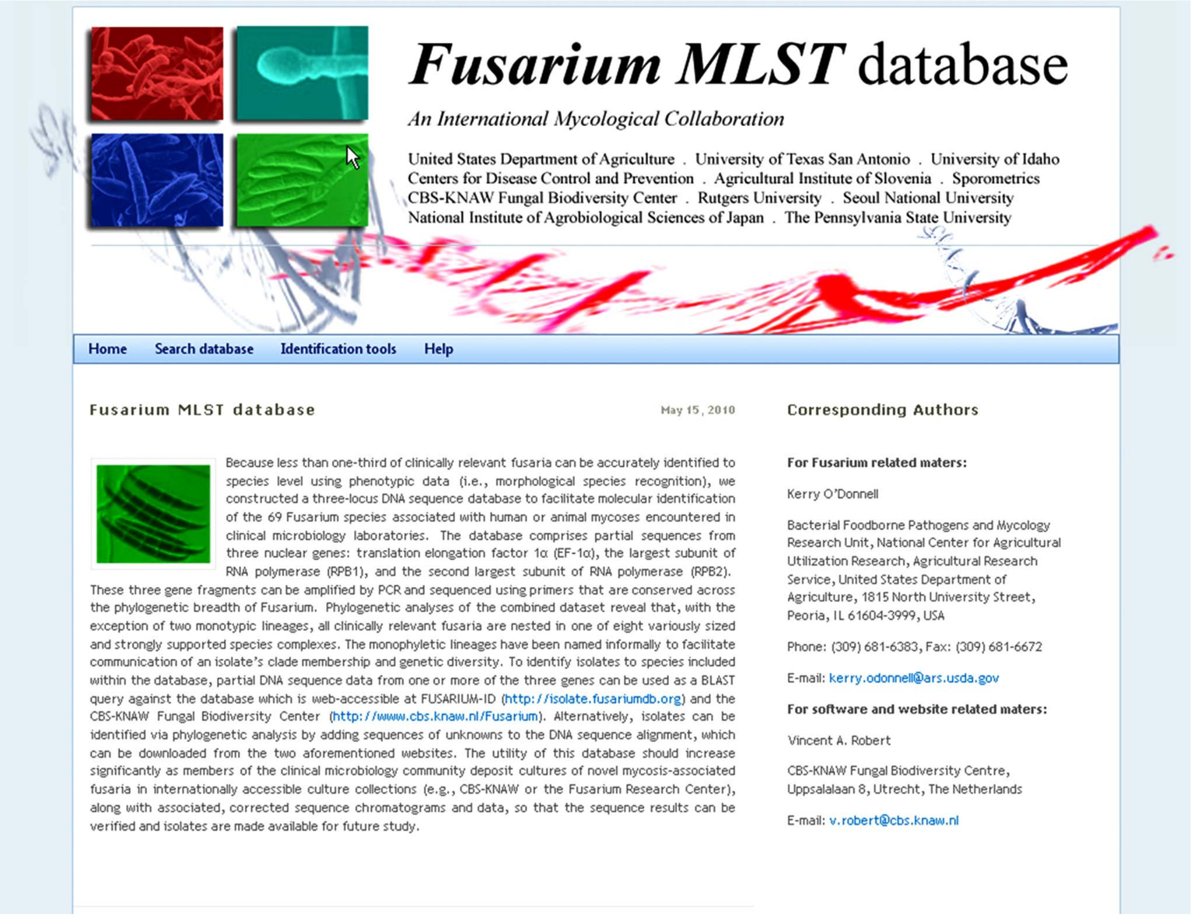
Fig. (9). Functional analysis of a series of strains of yeast (UPGMA left-vertical tree) based on a set of physiological features (UPGMA top horizontal tree). Normalized and reduced states of the above characters are displayed in clear red (negative result or absence of activity), in clear green (positive result or presence of activity), in blue (unknown result) or in an intermediate color between red and green for intermediate results.

even if we are "still a long way away from the 15-minute, $100 genome" (from http://scienceblogs.com/geneticfuture/2010/02/pacific_biosciences_session_at.php), we have to get ready for Tsunamis of data.

Normal computers will not be able to store and analyze such massive databases. We'll have to move to cloud computing and parallelization will become an absolute necessity for many applications. We are therefore vigorously investigating all these issues and trying to find efficient solutions that are both user friendly, fast and accurate. Solutions must be found at all levels of the problem: storage and databases, hardware and algorithms. Due to the probable complexity of future systems and the extent of the needed resources, it is likely that many laboratories will have to share such facilities. Companies or groups specialized in such infrastructures will probably flourish.

Another major issue will be the training of taxonomists and other end-users of such systems that can become giant black boxes that can be misused. Scientists of the future will of course have to master bioinformatics like never before. Multidisciplinarity and collaboration between groups of different nature are going to be major factors for successful research projects.

With most research projects, funding is provided by national or international agencies for periods ranging from a few months to a few years, usually 3 to 5. Large hosted bioinformatics and database infrastructures are expansive to establish but must also be permanently maintained and further developed to reflect new needs and standards. Classical funding schemes are most of the time inappropriate for large public and online databases that are heavily used by the scientific community. Permanent funding of all initiatives producing large databases, bioinformatics tools and websites is unlikely and not a good idea. Like in any research project, people tend to lose their dynamism and flexibility when heavily and easily supported. One of the solutions could be shared public and private funding for major and important databases and websites with regular evaluations. Financial contributions from end-users are also a partial financing solution and could ensure direct feedback to the administrators of the databases and websites. Usefulness, user-friendliness and enhancements of the system could therefore be improved. However, the reluctance of scientists to pay for software and associated services could be a major hurdle.



Fusarium MLST database

An International Mycological Collaboration

United States Department of Agriculture . University of Texas San Antonio . University of Idaho
Centers for Disease Control and Prevention . Agricultural Institute of Slovenia . Sporometrics
CBS-KNAW Fungal Biodiversity Center . Rutgers University . Seoul National University
National Institute of Agrobiological Sciences of Japan . The Pennsylvania State University

Fusarium MLST database

May 15, 2010

Because less than one-third of clinically relevant fusaria can be accurately identified to species level using phenotypic data (i.e., morphological species recognition), we constructed a three-locus DNA sequence database to facilitate molecular identification of the 69 Fusarium species associated with human or animal mycoses encountered in clinical microbiology laboratories. The database comprises partial sequences from three nuclear genes: translation elongation factor 1α (EF-1α), the largest subunit of RNA polymerase (RPB1), and the second largest subunit of RNA polymerase (RPB2). These three gene fragments can be amplified by PCR and sequenced using primers that are conserved across the phylogenetic breadth of Fusarium. Phylogenetic analyses of the combined dataset reveal that, with the exception of two monotypic lineages, all clinically relevant fusaria are nested in one of eight variously sized and strongly supported species complexes. The monophyletic lineages have been named informally to facilitate communication of an isolate's clade membership and genetic diversity. To identify isolates to species included within the database, partial DNA sequence data from one or more of the three genes can be used as a BLAST query against the database which is web-accessible at FUSARIUM-ID (http://isolate.fusariumdb.org) and the CBS-KNAW Fungal Biodiversity Center (http://www.cbs.knaw.nl/Fusarium). Alternatively, isolates can be identified via phylogenetic analysis by adding sequences of unknowns to the DNA sequence alignment, which can be downloaded from the two aforementioned websites. The utility of this database should increase significantly as members of the clinical microbiology community deposit cultures of novel mycosis-associated fusaria in internationally accessible culture collections (e.g., CBS-KNAW or the Fusarium Research Center), along with associated, corrected sequence chromatograms and data, so that the sequence results can be verified and isolates are made available for future study.

Corresponding Authors

For Fusarium related maters:

Kerry O'Donnell

Bacterial Foodborne Pathogens and Mycology
Research Unit, National Center for Agricultural
Utilization Research, Agricultural Research
Service, United States Department of
Agriculture, 1815 North University Street,
Peoria, IL 61604-3999, USA

Phone: (309) 681-6383, Fax: (309) 681-6672

E-mail: kerry.odonnell@ars.usda.gov

For software and website related maters:

Vincent A. Robert

CBS-KNAW Fungal Biodiversity Centre,
Uppsalalaan 8, Utrecht, The Netherlands

E-mail: v.robert@cbs.knaw.nl

Fig. (10). Example of a website managed using the web version of the BioloMICS software. *Fusarium* MLST database website (http://www.cbs.knaw.nl/Fusarium; O'Donnell *et al*. [8]).

ACKNOWLEDGEMENT

None declared.

CONFLICT OF INTEREST

None declared.

REFERENCES

[1] V. Robert, J.-E. de Bien, and G.L. Hennebert, "ALLEV, a new program for computer-assisted identification of yeasts", *Taxon*, vol. 43, pp. 433-439, 1994.

[2] V. Robert, P. Evrard, and G.L. Hennebert, "BCCM/Allev 2.00 an automated system for the identification of yeasts", *Mycotaxon*, vol. 64, pp. 455-463, 1997.

[3] V. Robert, and S. Szoke, "BioloMICS, Biological Manager for Identification, Classification and Statistics", Version 6.2, BioAware, Hannut, Belgium, 2003.

[4] V. Robert, "BioloMICSWeb software for online publication and polyphasic identification of biological data", Version 1, BioAware, Hannut, Belgium, 2000.

[5] V. Robert, W. Epping, T. Boekhout, M.Th. Smith, G. Poot, and J.A. Stalpers, "CBS Yeasts database", Centraalbureau voor Schimmelcultures, Utrecht, The Netherlands, 2004.

[6] R.A. Samson, and J.C. Frisvad, "Penicillium subgenus Penicillium: new taxonomic schemes, mycotoxins and other extrolites", *Studies in Mycology,* vol. 49, p. 257, 2004.

[7] L. Mostert, J.Z. Groenewald, R.C. Summerbell, V. Robert, D.A. Sutton, A.A. Padhye, and P.W. Crous, "Species of Phaeoacremonium associated with human infections and environmental reservoirs in infected woody plants", *J. Clin. Microbiol.,* vol. 43, pp. 1752-1767, 2005.

[8] K. O'Donnell, D.A. Sutton, M.G. Rinaldi, B.A.J. Sarver, S.A. Balajee, H.-J. Schroers, R.C. Summerbell, V. Robert, P.W. Crous, N. Zhang, T. Aoki, K. Jung, J. Park, Y.-H. Lee, S. Kang, B. Park, and D.M. Geiser, "An internet-accessible DNA sequence database for identifying fusaria from human and animal infections", *J. Clin. Microbiol.,* In Press, 2010.

[9] J.W. Grzymala-Busse, "Data reduction: discretization of numerical attributes" In *Handbook of data mining and knowledge discovery*, Publisher Oxford University Press, Inc. New York, NY, USA, 2002, pp. 218 - 225.

[10] P. Legendre, L. Legendre, "*Numerical Ecology*", *Elsevier*, Amsterdam, The Netherlands, 1998.

[11] V. Robert, "*Data Management and Bioinformatics*". In Yeasts and Food, T. Boekhout and V. Robert, Eds., Behr's Verlag Hamburg, Germany, 2003, pp. 139-170.

[12] J.C. Gower, "A general coefficient of similarity and some of its properties", *Biometrics*, vol. 27, pp. 857-874, 1971.

[13] P.W. Crous, W. Gams, J.A. Stalpers, V. Robert, G. Stegehuis, "MycoBank: an online initiative to launch mycology into the 21st century", *Studies in Mycology*, vol. 50, pp. 19-22, 2004.

Received: November 14, 2010 Revised: December 20, 2010 Accepted: April 22, 2011