IDENTAX

bacterial identifier

User’s manual

Free libraries used:
Prefuse information visualization toolkit (http://prefuse.org)
POI by Apache Software Foundation (http://poi.apache.org/)
Java CSV (http://sourceforge.net/projects/javacsv/)
L2FProd.com Common Components (http://www.l2fprod.com)

Manual version 1.0 – Printed on 12/12/08 21:44

Table of contents

INTRODUCTION 4

GETTING STARTED 5

PROGRAM USAGE 6

FIRST RUN 6

MAIN SCREEN 7

MATRIX LIBRARY 10

MATRIX LIBRARY SCREEN 10

MANAGING MATRICES 12

IDENTIFICATION WORKSPACE 13

CREATING A NEW WORKSPACE 13

WORKSPACE SETTINGS 14

WORKSPACE SCREEN 16

WORKSPACE SUMMARY 19

MANAGING WORKSPACES 20

WORKING WITH DICHOTOMYCAL TREES 21

SETTING-UP THE TREE GENERATOR 22

EXPORT TREES 23

ALGORITHMS AND THEORY 25

IDENTIFICATION ALGORITHM 25

SUGGESTING UNDONE TESTS 28

MATRIX CREATION 30

.XLS FILE (MICROSOFT EXCEL) 30

.CSV FILE (COMMA SEPARATED VALUES FILE) 32

WORKSPACE SUMMARY CUSTOMIZATION 34

Introduction

IDENTAX bacterial identifier is a free tool for the taxonomical identification of bacteria using phenotypical characteristics. It is adressed to the microbiologist community, and also to any field with a need for cheap and fast classification of unknown strains of bacteria.

In order to perform the identification process, *IDENTAX* uses the existing knowledge of various scientific workgroups, which generously yield matrices containing the prior probabilities of a taxa set of response (in a positive or negative way) to a specific test.

This software is developed using Sun Microsystems's Java Technology and runs perfectly on almost all existing operating systems. It is distributed under the LGPL license, allowing free use and modification of *IDENTAX* to all users. The complete text of this license is included with all *IDENTAX* distributions, and also available through the *IDENTAX* website www.identax.org.

Getting started

Note: Before proceed with the next steps, remember that is highly recommended check if there are any updates or news available on *IDENTAX* website http://www.identax.org

Install Java Runtime Environment

As this program runs over the *Java Virtual Machine* (JVM), the only requirement to execute *IDENTAX* is having a *Java Runtime Environment* (JRE) version 6 or newer.

JRE is bundled in some operative systems, like Mac Os X, but it has to be installed by the users in other ones, like Windows or Linux.

Due to license terms, the *Java Runtime Environment* cannot be included with *IDENTAX* and has to be downloaded separately on Sun Microsystems's website http://www.java.com.

Run *IDENTAX* build

IDENTAX is supplied as pure platform-independent .jar package or as customized packages for different operating systems. Operating-system-customized packages will be installed and executed as the other programs present in your system.

If you choose the .jar package and you have installed *Java Web Start*, you will be able to launch the *.jar* package, without need of any installation, you just click it. Else, you can launch the .jar file using the following command on your system terminal (remember to replace <*version*> with the current version string of the filename):

```
java –jar Identax_<version>.jar
```

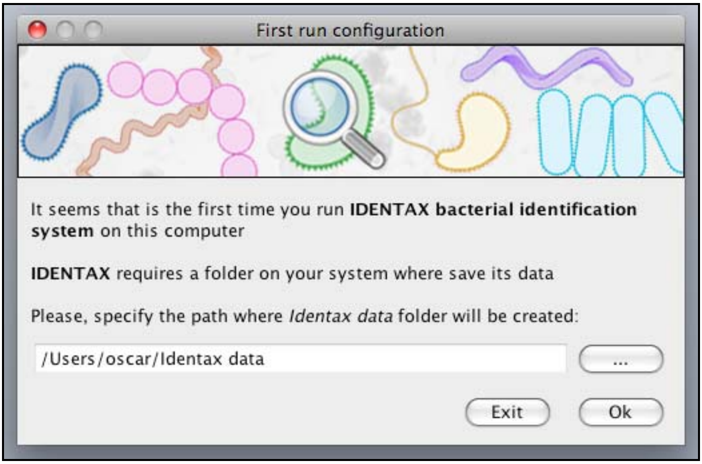
The installation and launch of operating-system-customized packages is the same use with other system programs. Additional information about these process could be available on installation screen or in supplied *README* files.

PROGRAM USAGE

Note: The screenshots in this manual are from the Mac Os X version of IDENTAX and may look different on other systems.

First run

The first time that you run the software, it will prompt you to choose a directory where save its data.



It seems that is the first time you run **IDENTAX bacterial identification system** on this computer

IDENTAX requires a folder on your system where save its data

Please, specify the path where *Identax data* folder will be created:

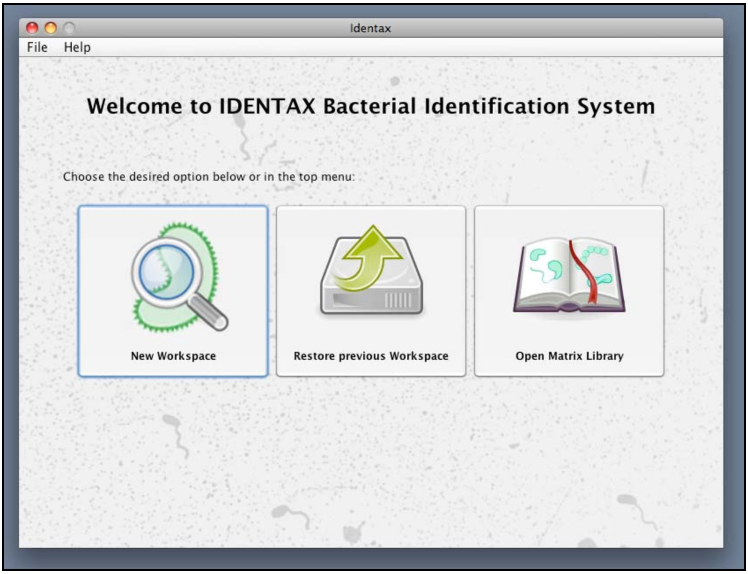
/Users/oscar/Identax data

By default, a folder called “Identax data” will be created in your user directory. The new directory will contain the saved workspaces, the imported matrices and other files used by the program.

If this directory is removed, *IDENTAX* will prompt to create it at the next launch. You can safely remove it if you do not plan use *IDENTAX* anymore.

Warning: If you delete “Identax data” folder, all saved matrices, workspaces and customized summaries will be also deleted.

Main screen



Welcome to IDENTAX Bacterial Identification System

Choose the desired option below or in the top menu:

The main screen contains three areas:

- Title bar
- Menu bar
- Work area

Title bar

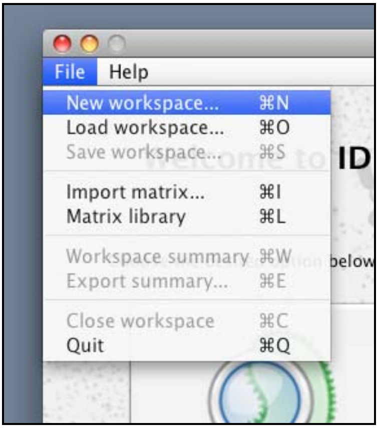
Title bar is the area on the top of the window. Its look and its behavior are system-dependent but there are always two buttons to close and minimize the program.

IDENTAX don't allow resizing its main window, so maximize or zoom buttons will be disabled.

If you try to close the main window and unsaved changes are present in the current view, a message will ask you to confirm this action.

Menu bar

Menu bar is placed just under the Title bar. It's a normal menu as many programs use.



The options presented are accessible thought shortcut keys. These shortcuts are a combination of a letter plus your default command key (CTRL in Windows, on Mac Os X... symbol ♦ is used here to designate general command key).

- ***File***
  - **New workspace** ♦ + N : Same as Welcome screen *New workspace* option (see page 13)
  - **Load workspace** ♦ + O : Same as Welcome screen *Load workspace* option (see page 20)
  - **Save workspace** ♦ + S : When in the Workspace screen, this option saves the active workspace to disk (see page 20)
  - **Import matrix** ♦ + I : Same as Matrix Library option *Import* (see page 12)
  - **Matrix library** ♦ + L : Same as Welcome screen *Open Matrix Library* option (see page 10)

- **Workspace summary** ♦ + W : When in the Workspace screen, this option do the same that View Summary button (see page 19)
- **Export summary** ♦ + E : When in the Workspace screen, this option allow export the workspace summary as text file (see page 19)
- **Close workspace** ♦ + C : When in the Workspace screen, this option returns to Welcome screen.
- **Quit** ♦ + Q : Same as close button on Title bar (see page 7)

- ***Help***
  - **Go to IDENTAX website:** This option will open a browser window pointing at *IDENTAX* website, where you could find the latest news and help.
  - **Check for updates:** This option will connect you to the *IDENTAX* website to check if any update is available.
  - **About:** This option will show the version and credits of *IDENTAX*.

Work area

The work area is the main area of the window. It changes to show the required information and control elements needed in every moment.

There are two main sections in *IDENTAX*

- Matrix library
- Identification workspace

Both will be deeply commented in the following chapters.

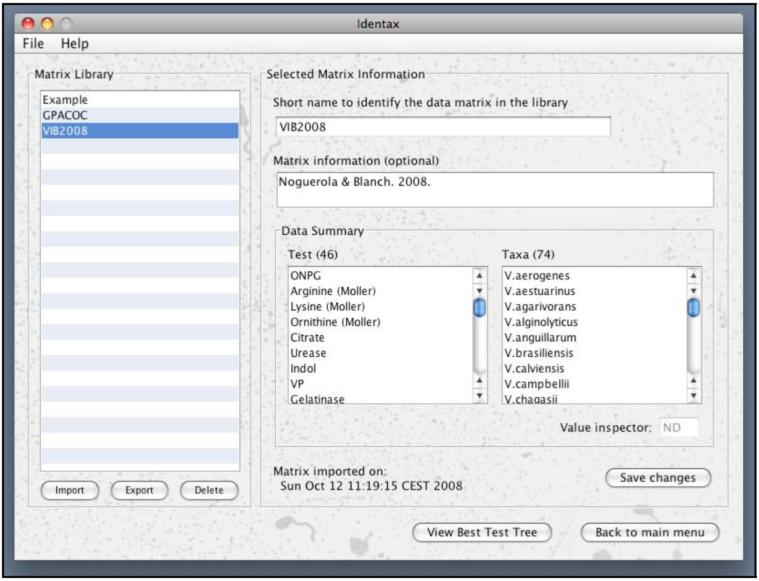
Matrix library

The approach of *IDENTAX* is, knowing how a set of taxa "uses to" response at certain tests (this is called *prior probabilities*), discover, for a concrete test results set, which is the closest taxon expected to give that results.

These prior probabilities are studied by the biologists, who represent it in two-dimension matrix, containing on one axis all the taxa studied and, on the other axis, all the tests used in the study.

Maintaining this nomenclature, *IDENTAX* calls its data files *Matrices*, and the set of all imported matrices in the system is called *Matrix library*.

Matrix library screen



Matrix Library:
- Example
- GPACOC
- VIB2008

Short name to identify the data matrix in the library: VIB2008

Matrix information (optional): Noguerola & Blanch. 2008.

Test (46)	Taxa (74)
ONPG	V.aerogenes
Arginine (Moller)	V.aestuarinus
Lysine (Moller)	V.agarivorans
Ornithine (Moller)	V.alginolyticus
Citrate	V.anguillarum
Urease	V.brasiliensis
Indol	V.calviensis
VP	V.campbellii
Gelatinase	V.chagasii

Value inspector: ND

Matrix imported on: Sun Oct 12 11:19:15 CEST 2008

Matrix library screen shows two main frames.

The first of them is labeled as *Matrix Library* and it contains a list with all imported matrices in our system plus three buttons to manage them. Detailed information about managing matrices is available on the next section.

This matrix list is also available on other parts of the program, like when creating a new Workspace. Before using a reference matrix it has to be imported through *Matrix library* or through the corresponding option in the *File* menu

If a Matrix is selected in the list, the second frame, labeled *Selected Matrix Information* will show the data of the selected matrix.

The data available is the following:

- **Matrix's short name:** This name is used to reference the Matrix on the lists and on the file system. It has to be a unique name, or otherwise, it will overwrite the previous Matrix with that name.
- **Matrix information:** Contains information about the matrix, like its authors, a brief description or anything else.
- **Data Summary:** Has two lists representing the both axis of the matrix, one for the Tests and another one for the Taxa. If you select a pair test-taxon the *Value inspector* will show the probability introduced in the matrix (in %) of selected taxon that gives positive result on the selected test. Value *ND* means Non-Deterministic value, as it can be missed or 50% value.

If you would know more about what values mean and how *IDENTAX* process it see the Algorithms and theory chapter (page *25*).

The *View Best Tree* button on the bottom of the screen is explained in the Working with dichotomycal trees chapter (page 21)

Managing matrices

Matrix management is realized using the buttons under the Matrix list:

- **Import:** Will show a dialog to select a file and import it.
- **Export:** Will show a dialog to save the exported file outside the *IDENTAX* environment. At this moment only .imat format (IDENTAX matrix) is available for exportation.
- **Delete:** Will delete the selected matrix.

Warning: Delete a matrix will also delete all the saved Workspaces referencing it.

IDENTAX can import matrices from various source formats. The supported formats list till the moment is:

- **.IMAT (Identax Matrix files)**
- **.XLS (Microsoft Excel 97-2004 files)**
- **.CSV (Comma Separated Values files)**
- **.MAT (Bacterial Identifier files)**[1]

Note: New Microsoft Excel 2007 format .XLSX it's not supported currently.

If you want to know more about format files or how to create new matrices with your own data, all information needed is available on Matrix creation (page **¡Error!Marcador no definido.**)

[1] Bacterial Identifier (© Blackwell Scientific Publications Ltd, Oxford, 1991) is a old MS-DOS software that also used his own file format.

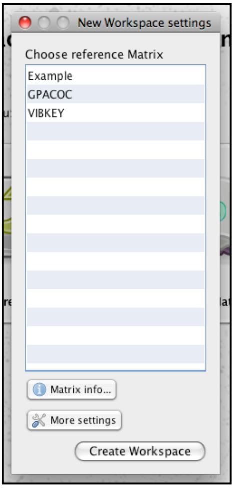
Identification Workspace

IDENTAX calls *Workspace* the combination of a reference matrix, a subset of realized tests and concrete settings.

The target of an identification job is identifying an unknown strain of bacteria realizing some tests on it. If we could guarantee that this unknown strain is likely to be "that one" and not "any of these other" with a high degree of confidence, we will say that identification has been realized successfully.

> **Note:** Notice that the goodness of IDENTAX conclusions depends on the reliability of the supplied data via Matrix.
>
> Think about IDENTAX as a calculator, it will do perfect calculations with the given numbers, but maybe the source numbers are not correct …

Creating a new Workspace

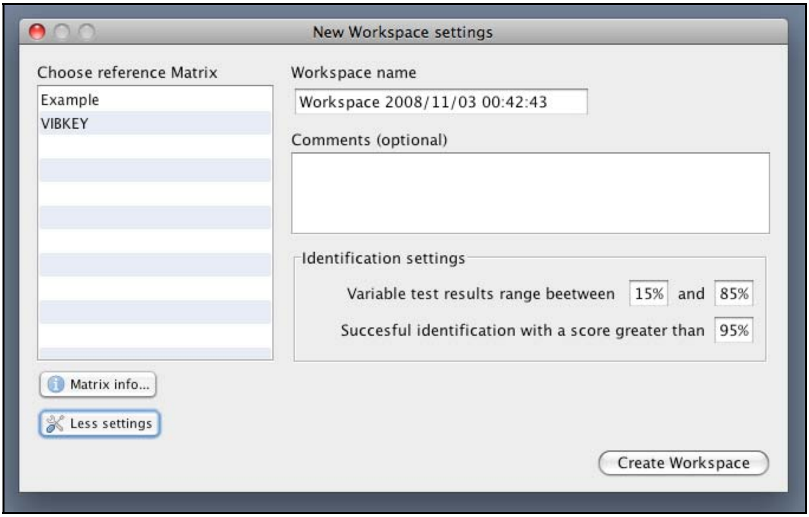


When you create a new Workspace there is only one thing the user has to decide: which reference Matrix use.

One list with the available matrix in the library will appear and will let you choose which matrix has to be used as a reference. You can see the matrix details clicking the *Matrix info* button at the bottom of the screen. This panel is like the *Selected Matrix Information* frame of Matrix library screen (page 10).

If you don't want to customize the identification settings, just click the *Create Workspace* button to go to the *Workspace screen*. Otherwise, click the *More settings* button to show the *Workspace settings* window (see next section).

Workspace settings



New Workspace settings

Choose reference Matrix

Example
VIBKEY

Workspace name: Workspace 2008/11/03 00:42:43

Comments (optional)

Identification settings

Variable test results range beetween 15% and 85%

Succesful identification with a score greater than 95%

Workspace name

This is the name used in the reports and in the file system to identify this Workspace. Can be changed when saving.

Comments

Optional comments about the Workspace can be added and it will appear at the reports.

Identification settings: Variable test results range

The algorithm needs to classify the probabilities in the matrix as *positive* or *negative*.

Sure that a 100% probability is positive and 0% one is negative, but intermediate values are not so clear. Values that are not mainly positive or mainly negative are considered *variable* (they could be “randomly” positive or negative).

You can adjust the variable probabilities range with the lower and upper limits where probabilities will be considered *variable*.

Values smaller than the lower limit will be considered *negative*, and values greater than the upper limit will be considered *positive*.

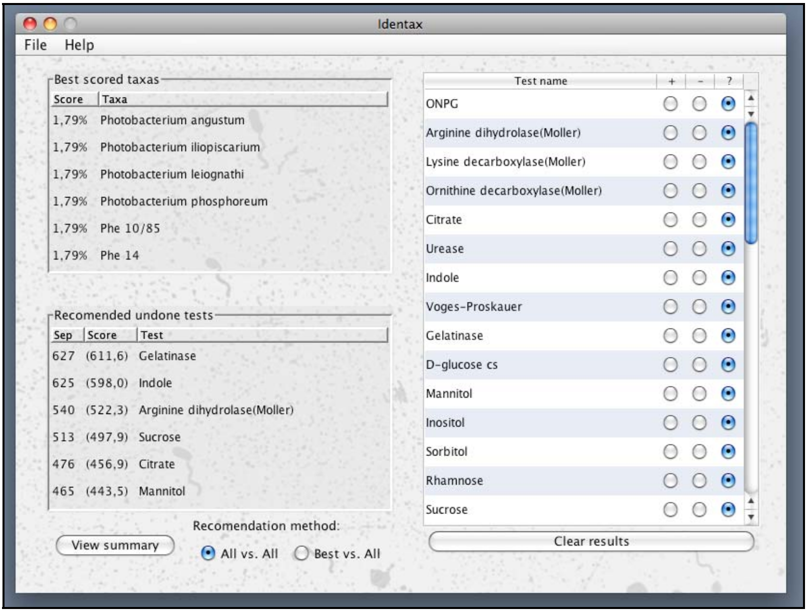
Identification settings: Identification threshold

Identification threshold is just a limit where the application will consider no further tests are needed and it will show the identified taxon.

This limit references the *relative or posteriori probability* of any taxon to belong to the same strain as the unknown specimen and not to another one.

For a better comprehension of these and other concepts, please take a look at the Algorithms and theory chapter (page 25).

Workspace screen



Identax

Best scored taxas

Score	Taxa
1,79%	Photobacterium angustum
1,79%	Photobacterium iliopiscarium
1,79%	Photobacterium leiognathi
1,79%	Photobacterium phosphoreum
1,79%	Phe 10/85
1,79%	Phe 14

Recomended undone tests

Sep	Score	Test
627	(611,6)	Gelatinase
625	(598,0)	Indole
540	(522,3)	Arginine dihydrolase(Moller)
513	(497,9)	Sucrose
476	(456,9)	Citrate
465	(443,5)	Mannitol

Recomendation method: ● All vs. All ○ Best vs. All

Test name	+	-	?
ONPG	○	○	●
Arginine dihydrolase(Moller)	○	○	●
Lysine decarboxylase(Moller)	○	○	●
Ornithine decarboxylase(Moller)	○	○	●
Citrate	○	○	●
Urease	○	○	●
Indole	○	○	●
Voges-Proskauer	○	○	●
Gelatinase	○	○	●
D-glucose cs	○	○	●
Mannitol	○	○	●
Inositol	○	○	●
Sorbitol	○	○	●
Rhamnose	○	○	●
Sucrose	○	○	●

The main area of the Workspace is divided into four sections.

Test results area

This is the main interactive area in this screen. It contains a list of every test in the matrix with 3 more columns where indicate if the test result is positive (+), negative (-) or missed (?).

You can enter the results either clicking the corresponding radio button or pressing one of the following keys:

- **+** or **1** to mark as positive
- **-** or **0** to mark as negative
- **?** or **m** to mark as missed

The tests maintain the order in which the author made the matrix (can be useful, in example, where most used tests appear on the top of the list). If you want to change the default order, you can click the header of the column to sort by new criteria.

If you want to clear all entered results, there is a button at the bottom of the list to do it.

Best scored taxa area

That list shows the six species with the most similar behaviour like the unknown specimen. The column *score* shows the *posteriori or relative probability* of every taxon as commented before.

Don't worry if the sum of all taxa scores is less than 100%, it is due discarded or negligible taxa that don't appear in the list.

> **Note:** It is possible that any taxon could be identified with the given test sequence. In that case a message will appear.

If a concrete taxon surpasses the identification limit set in the Workspace configuration, a panel with the title *Taxon identified!* will automatically appear.

Recommended undone tests area

That list shows, on the run, the best six undone tests. A test is "best" than another if it allows separate more taxa.[2]

There are two methods of recommendation (switch between them with the two on-screen radio buttons):

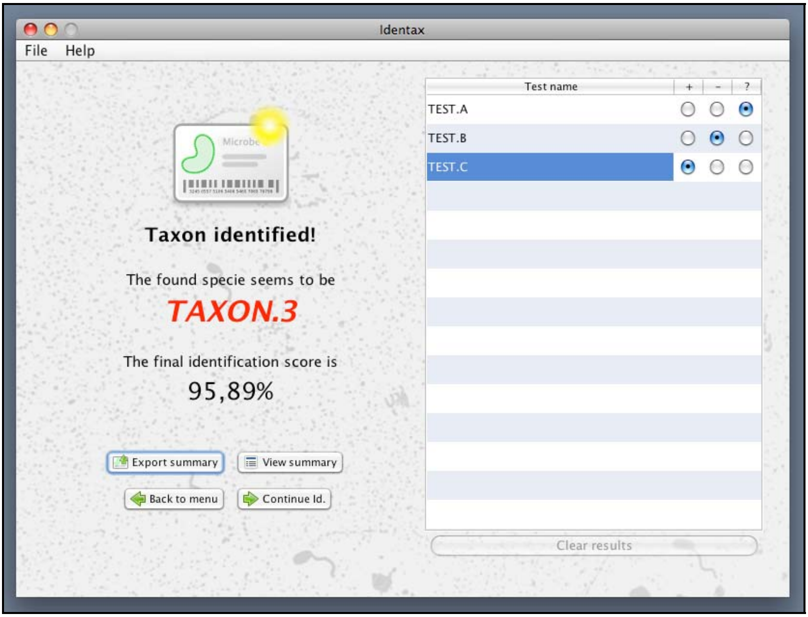
- **All vs. All**: It calculates for each taxon and for each undone test if that test separates the current taxon from the other candidates.
- **Best vs. All**: It calculates for every undone test if that test separate the best taxon from the other candidates

[2] See Algorithms and theory chapter (page 28) for more information

The first method will recommend tests that separate as much taxa as possible among themselves, and the second will try to separate the best-scored taxon from the others.

The column *Separations* on the recommended tests list shows how many pairs of taxa are separated by these tests.

Taxon identified panel



Taxon identified!

The found specie seems to be

TAXON.3

The final identification score is

95,89%

Test name	+	-	?
TEST.A	○	○	●
TEST.B	○	●	○
TEST.C	●	○	○

If a taxon has been identified, a panel with the basic actions and information is shown. These actions are available through the *File* menu too, but easier to access. If you choose to *Continue Id.* this panel will not reappear.

Workspace summary

In a Workspace there is much more information that the one needed to perform the basic work. At any moment of the identification a *Workspace summary* is available for view or export.

You can view the summary clicking the button besides the *Recommendation method* switch, via *File* menu or pressing the shortcut keys (see page 8). Export summary option is available either in the two last ways.

The information present in the summary is variable depending the state of the identification. It is also customizable via a very easy template system.

It can contain all the basic information about the current workspace, and also more sophisticated things, like:

- **Atypical results for the best scored taxa**: if a result was expected to be positive for any of the best scored taxa and supplied result is negative (or vice versa)
- **Absolute and relative probabilities for the best scored taxa**
- **Expected test results for the identified taxon**

The generated summary has plain text format for an easier manipulation and print using any plain text editor of your choice (for example Notepad in Windows, TextEdit in Mac Os X, Kate or GEdit in Linux…)

To view all variables available in the summary and/or how to customize your summaries, check Workspace summary (page 34).

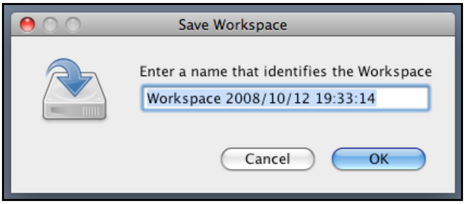
Managing workspaces

You have all the commands concerning Workspaces on the *File* menu or via shortcut keys (see page 8).

They include create, close, save and load a workspace.

Saving and Loading workspaces

To save a workspace you just need to click the file menu or its associated shortcut key. A dialog will appear asking you for the name to assign at the saved Workspace



Save Workspace

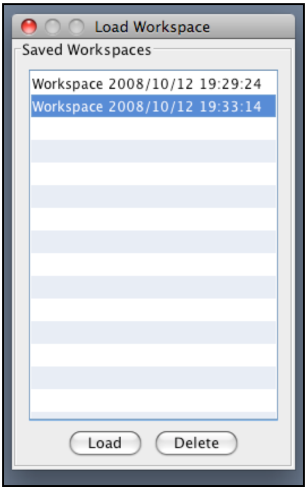
Enter a name that identifies the Workspace

Workspace 2008/10/12 19:33:14

Warning: Each saved workspace has to have unique name. If you try to save a new workspace with an existing name *IDENTAX* will prompt to overwrite it.

To load a previously saved Workspace, select that option and a list with all saved Workspaces will appear. Just click the Workspace name you want on the list and then the *Load* button.

To delete a previously saved Workspace, just click its name on the list that appear while loading, and click the *Delete* button



Load Workspace

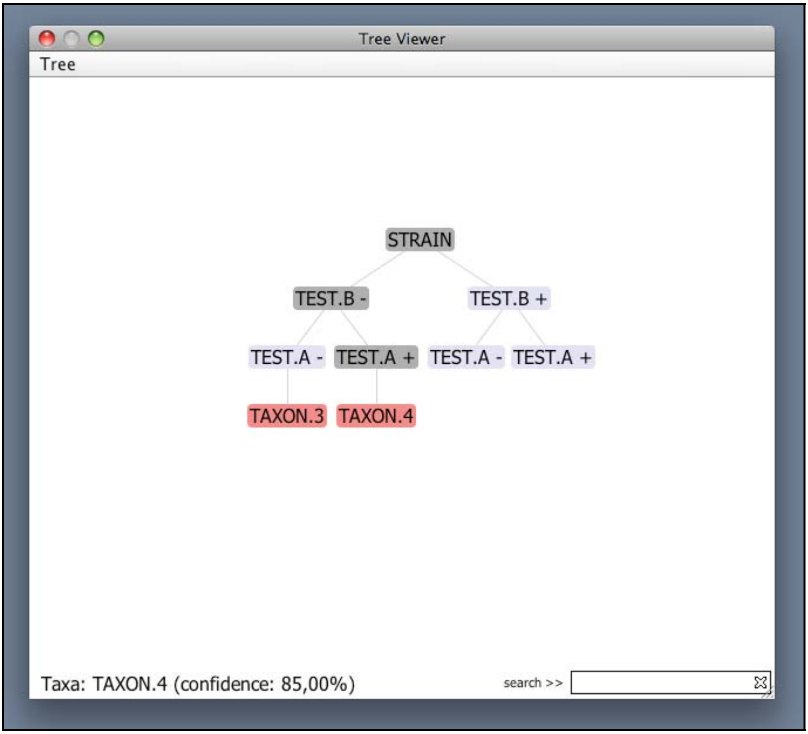
Saved Workspaces

Workspace 2008/10/12 19:29:24
Workspace 2008/10/12 19:33:14

Working with dichotomycal trees

Another of the features of IDENTAX is the generation of dichotomycal trees[3] to perform the identification. That could be useful when you work without the presence of a computer or you have a visual idea of the relevance of all the tests.

The nodes near tree root are the most discriminative tests in the matrix and they go expanding until the leaves, where is the specie that is expected to give that chain of results.



Taxa: TAXON.4 (confidence: 85,00%)

[3] *Tree* in this context means a directed graph without loops

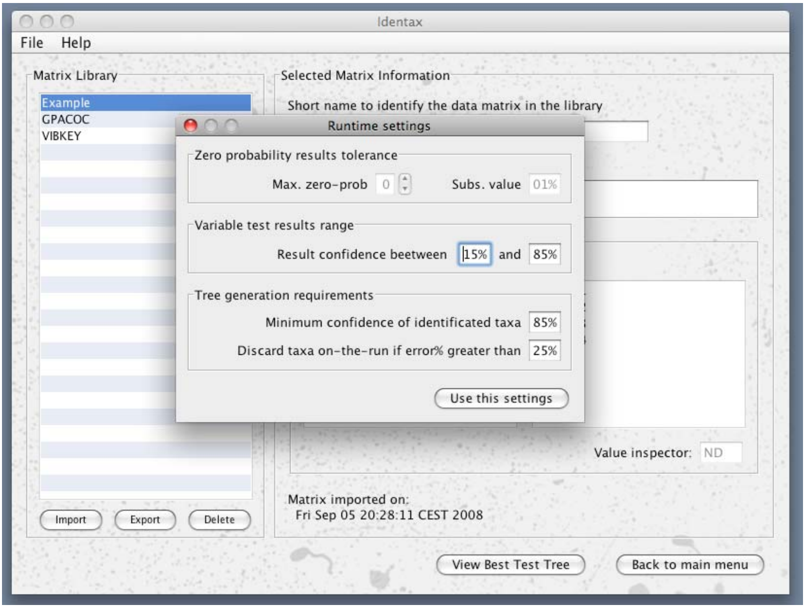
In the tree image, it can be seen that if we obtain a negative "B" test result and a positive "A" test result, the unknown strain is probable to be a member of the "TAXON.4" specie (with a 85% confidence, as it's indicated at the bottom of screen on moving the mouse cursor over a cell).

The integrated tree visualization is based on the Prefuse Library (http://prefuse.org) and has a few limitations (you cannot view the full tree, only one branch can be expanded at same time…).

To allow a better flexibility and tree representation, IDENTAX offers the possibility to export the generated tree to third parties software. More about this in the

Export trees section.

Setting-up the tree generator



Identax

Matrix Library

Example
GPACOC
VIBKEY

Selected Matrix Information

Short name to identify the data matrix in the library

Runtime settings

Zero probability results tolerance

Max. zero-prob 0 Subs. value 01%

Variable test results range

Result confidence beetween 15% and 85%

Tree generation requirements

Minimum confidence of identificated taxa 85%

Discard taxa on-the-run if error% greater than 25%

Value inspector: ND

Matrix imported on:
Fri Sep 05 20:28:11 CEST 2008

Tree generation is available in the Matrix Library (see page 10), when selecting a matrix and the *View Best Test Tree* button is clicked.

A dialog similar to the Workspace settings (see page 14) will appear.

The *Minimum confidence of identified taxa* parameter indicates which has to be the minimum confidence allowable for a taxon to be in a leaf, if the taxon has a lower score than the entered in this field, it won't appear at the tree. In other words, that value says the global confidence of the tree.

A tree with an overall confidence greater than 95% means that a taxon present on this tree should have to react as is expected at least on 95% of times.

Export trees

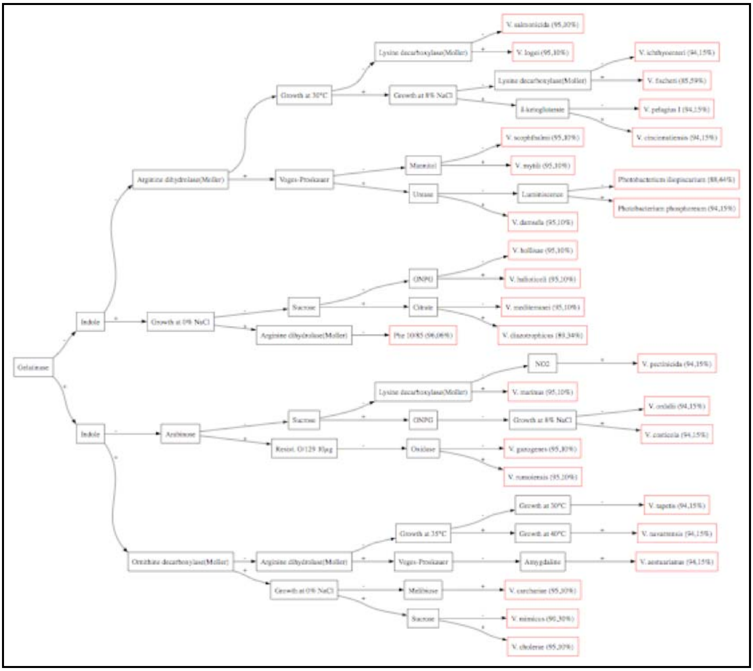
Trees can be exported selecting one of the following options in the *Tree* menu at the top of the *Tree Viewer* window or pressing the associated shortcuts as follows:

- **Save as image** ♦ + **S** : Exports the current window image (as you see it) in one of this image formats:
    - Bitmap Image (*.bmp)
    - JPEG Image (*.jpg)
    - PNG Image (*.png)
- **Export as file** ♦ + **E** : Exports as a text file the current tree structure. The generated files are not directly useful by a normal user, but they can be passed to a visualization program for a better customization of the tree. The available formats are:
    - **TreeML file** (*.tml): That is a standard format to represent trees in XML format.

- **IDENTAX XML file** (*.xml): A custom implementation of xml format to represent the dichotomycal trees in a human readable text format.

- **Plain text graph file** (*.dot): Used by Graphviz[4], one of the best graph visualization software nowadays, it is maybe the most useful export format you can work.

Below, there is an example of a complex tree generated using GraphViz Software:



[4] http://www.graphviz.org

Algorithms and theory

The following Chapter is about how internal processes of IDENTAX work and the basic theoretical aspects that make them possible.

The main algorithms are two: identify an unknown isolate starting from the test results and suggest the best-undone test at any moment.

Identification algorithm

The essential process in IDENTAX is evaluating the probability of each taxon present in the matrix to mach the given pattern of test results.

The information about which is the expected result for a concrete taxon-test pair is contained in the matrices. In statistical terms, this is called *prior probabilities* (what is known about something in absence of any evidence).

But that is not what we are looking for. We want to know, giving a concrete pattern of test results, which taxon is the most expected to give that results. This is called *posteriori probabilities* (the conditional probability that is assigned after some relevant evidence is taken into account).

Bayes theorem says how prior and posteriori probabilities are related:

$$P\left(t_i \mid R\right) = \frac{P\left(R \mid t_i\right) P\left(t_i\right)}{\sum_{i \in taxa} P\left(R \mid t_i\right) P\left(t_i\right)}$$

- ***$P(t_i|R)$*** is the (posteriori) probability that an unknown isolate can be really a member of taxon t_i, given the pattern of tests results R.

- ***P(R|t***

- ***$P(R|t_i)$*** is the (prior) probability that the unknown has a pattern R considering that it is a member of taxon t_i.
- ***$P(t_i)$*** would be the expected frequency of incidence of strains in the matrix, but in our bacterial environment it can be set to 1 and excluded from the above formula.
- The denominator contains the summation for all the taxa of the above probabilities, and it is used to give a normalized probability value between 0 and 1.

We can restate the formula to the following one:

$$Score(t_i) = \frac{P_{prior}(t_i)}{\sum P_{prior}(t)}$$

So, the final *score* (the posteriori probability) of the taxon t_i is its probability of being consistent with the given results, but normalized using all other taxa probabilities.

NOTE: In some parts of this manual or the program, the words *absolute* or *relative* probabilities are used to designate the *prior* or *posteriori* probabilities respectively.

Identification example

Below there is an example of the identification process. We will use the following matrix as a reference matrix:

	Test 1	Test 2	Test 3	Test 4
Taxon A	0	30	100	90
Taxon B	90	0	100	0
Taxon C	100	10	85	100

The values in the matrix represent the chances of a positive response at the tests by the taxa in percentage.

Negative response probabilities are complementary to the positive ones (i.e., 30% positive chance is 70% negative chance: 1 – 0.3 = 0.7)

As an example, we have the following test pattern:

Test 1	Test 2	Test 3	Test 4
+	-	+	MISS.

IDENTAX will calculate the prior probabilities for each taxon as the product of the sequence of these test result chances:

$$P_{prior}(t_i) = \prod_{j \in tests} P_{i,j}$$

Taxa		Prior probabilities
A	0.00 * (1.00-0.30) * 1.00 =	0.00
B	0.90 * (1.00-0.00) * 1.00 =	0.90
C	1.00 * (1.00-0.10) * 0.85 =	0.765
	SUM =	1.665

NOTE: As you can see, Taxon A has a 0% probability due to the zero present in the Test 1. On the implemented algorithm, a 0 probability value is translated into a *very low* value to keep information that allows detect unexpected test results.

Normalizing these values gives:

Taxa		Identification score
A	0.000 / 1.665	0.00
B	0.900 / 1.665	0.54
C	0.765 / 1.665	0.46
	SUM =	1.00

In this example, identification cannot be performed as we have two taxa with high prior probabilities both. Additional tests will have to be performed in order to distinguish better one from the other ones.

The following section talks about what to do in this case. That is a small example, but if many tests remain undone, which of them is the best we can do?

Suggesting undone tests

We can consider one test better than another one if the first allows separating more taxa than the second. We understand that two taxa are *separated* if one is expected to give a positive result for a concrete test and the other is expected to give negative for the same test.

A small example:

Test	Taxa probs.				Expected taxa results				Taxa separated
	A	B	C	D	A	B	C	D	
1	.85	.15	.15	.85	+	-	-	+	4
2	1.0	1.0	.70	0.0	+	+	v	-	2

NOTE: The *Expected taxa results* depend on the *Variable Test Result Range* threshold set at the Workspace (see page 14)

We can see that the following taxa pairs are separated by the tests (note that values inside *Variable value range* are ignored):

Test 1	Test 2
A <> B	A <> D
A <> C	B <> D
B <> D	
C <> D	

NOTE: The option "**All vs. All**" or "**Best vs. All**" on the Workspace screen (see page 17) changes separations count.

"**All vs. All**" is the way the example shows, there are considered all possible pairs of taxa.

"**Best vs. All**" only consider the best taxon in this moment. Suppose that this option is selected and A is the best scored taxa, then only pairs containing the taxon A will be counted resulting on two separations for test 1 and only one for test 2

In case of two or more tests with the same number of separated taxa, an additional refinement consisting of taking into account the value of the probabilities is considered to choose the best.

A separation with 100 and 0 probabilities is better that one with 85 and 15 probabilities, then the test with the greatest sum of absolute difference between all its separated pairs will be suggested first.

Consideration about missing data

As you can see, the missing tests results are not a problem as long as they are skipped and the user always can mark some more tests (if any available) as done.

But, a conceptual problem happens with missed information in the matrices. If a pair test-taxon is missed, the Bayes' Theorem cannot manage missing data directly.

The IDENTAX approach to the problem consists on treat missing values as 50% probability, as all alternatives have been studied and we think this is the less intrusive way to handle missing data.

Matrix creation

Some advanced users (such as microbiologist) will appreciate create his or her matrix with the tests and taxa of his research field.

The process of creates a matrix and import it into *IDENTAX* is very easy. In short, you only have to create the base file using one of the available formats for importation and import it through the program.

As a general consideration, the values present in the matrix are always the a priori probabilities (expressed as percentage) of get a positive result of test X on the taxon Y. These values are expected to be between 0 and 100.

Below there are the formal aspects of every allowed format

.XLS file (Microsoft Excel)

Most users will prefer Microsoft's spreadsheet program as they are used at it, but it is not the only way to work with .xls file. There are available many free alternatives as OpenOffice.org[5].

Create a matrix file using excel is trivial. Just put the taxa names on the first column (called "A") and the tests names on the first row (called "1").

Note: As a normal table, cell A1 (first column, first row) will be empty (anything entered there will be ignored)

As IDENTAX uses Apache POI[6] library to handle Excel files and it has his own restrictions, the following indications must be checked:

[5] http://www.openoffice.org/

- Use General (a.k.a. Automatic) or Text cell type for names

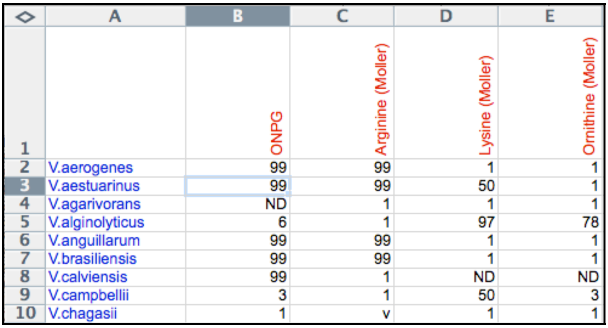
- Use General (a.k.a. Automatic) or Number cell type for probabilities

- Blank cells after the last column or after the last row must be removed.

Note: A *blank cell* is created if something is written in a cell and removed using backspace or delete key. The cell will appear as blank cell, but it is not null. IDENTAX will crash at the importation if a blank cell is found.

To avoid that, just use the contextual-menu Delete option when right-clicking a row or a cell if you have any problem.

An example of a basic excel matrix:



◇	A	B	C	D	E
1		ONPG	Arginine (Moller)	Lysine (Moller)	Ornithine (Moller)
2	V.aerogenes	99	99	1	1
3	V.aestuarinus	99	99	50	1
4	V.agarivorans	ND	1	1	1
5	V.alginolyticus	6	1	97	78
6	V.anguillarum	99	99	1	1
7	V.brasiliensis	99	99	1	1
8	V.calviensis	99	1	ND	ND
9	V.campbellii	3	1	50	3
10	V.chagasii	1	v	1	1

As you can see you can use any text styles (color, fonts, orientation) for the cells. The “A” column contain the taxa names and the “1” row contains the test names. Cell “A1” is empty and the other cells contain the probabilities in percentage (in the 0-100 range, without the “%” symbol) of a positive result by the taxon in the row if we do the test in the column.

[6] http://poi.apache.org/

There are some “ND” cells and another one with the “v” character. In fact, anything different of a number in the probabilities cells will be handled as a “ND”.

.CSV file (Comma Separated Values file)

CSV is a plain text format where each field is indicated using special characters to delimiter it. This file can be generated using a plain text editor, but most spreadsheet software also allows exporting of native program files to CSV.

The structure of this file is the same that the XLS, so the first line has to contain a first field that will be ignored (like the “A1” cell in XLS) and the tests names behind. The following lines must be containing the taxa names on the first field and the probability of a positive result (in percentage, without the “%” symbol) on the following fields.

You have use semicolon (**;**) character to separate fields and IDENTAX CSV parser (the JavaCSV library[7]) will detect if the quote (**"**) symbol delimits the text fields or not (don’t worry, it is not necessary use the quote symbol)

For example, the following matrix:

	Test A	Test B	Test C
Taxon 1	100	0	ND
Taxon 2	0	80	100
Taxon 3	100	100	0

will be as follows in CSV format

```
;Test 1;Test 2;Test 3
Taxa A;100;0;ND
Taxa B;0;100;100
Taxa C;100;100;0;
```

[7] http://sourceforge.net/projects/javacsv/

Note: Realize that the first field on the first line is empty; the beginning semicolon indicates the first field ends there and so it is null. Anything else can be putted in this field, but it will be ignored.

Note: Strange characters can appear at the imported data if there is a mess with the plain text file encoding. Not all plain text files have the same binary representation. By default ISO-Latin-1 encoding is used by Windows file system and UTF8 on other platforms. IDENTAX will try to guess the encoding of the given file, but it can easily fail.

If you experiment any trouble, just save it using the UTF-8 encoding (any plain text editor must let you choose the encoding of the text file when saving)

As final consideration on the Matrix creation, despite IDENTAX can import another matrix formats, they are not recommended to generate new matrices. The two explained formats are the easiest and most compatible way to create a matrix. The other formats are just present to allow importation of previous works on *IDENTAX*

Workspace summary customization

The generated summary can be easily customized just editing the summary template with a plain text editor.

The template is inside the *IDENTAX data* folder (its concrete location is specified the first time IDENTAX is launched) in a file named *summary_template.txt*

Default file contents can be changed as you like, but the line width should not have more than 80 characters for a good visualization.

Lines started with the **#** character will be removed from the output as they are considered commentaries only visible in the template.

The template can contain any text you want plus special names (variables) surrounded by the **%** character that will be replaced by IDENTAX for concrete values of the Workspace.

The list of all available variables and their meaning is the following:

Variable	Meaning
%WS_NAME%	Workspace name
%WS_INFO%	Workspace description/information
%MAT_NAME%	Reference matrix name
%MAT_INFO%	Reference matrix description/information
%MAT_TEST_N%	# of tests present in the matrix
%WS_DONE_TEST%	# of done tests in the current workspace
%MAT_TAXA_N%	# of taxa in the matrix
%WS_ID_RES%	Result of the identification (OK/FAIL)
%ID_NAME%	Name of the best scored taxon

%ID_ABS_PROB%	Absolute (prior) probability of the best scored taxon
%ID_REL_PROB%	Relative (posteriori) probability of the best scored taxon

The following variables are iterators, it means that each variable can contain many entries that will be printed on a different line each.

%I%	(for each iterator element) The count of the current element
%I_INC_TAXA_NAM%	(for each incoherent test result) The name of the taxa with incoherent result
%I_INC_TEST_NAM%	(for each incoherent test result) The name of the incoherent test
%I_INC_TEST_EXP%	(for each incoherent test result) The expected value of this test by this taxa
%I_INC_TEST_GIV%	(for each incoherent test result) The given result for this test
%I_INC_TEST_VAL%	(for each incoherent test result) The expected probability in % of this thest for this taxa
%I_TEST_NAME%	(for each test) Test name
%I_TEST_RES%	(for each test) Given test result (+/-/miss)
%I_ID_EXP_RES%	(for each test done) The expected result on the best scored taxon
%I_TAXA_NAME%	(for each top scored taxa) The name of the taxa
%I_TAXA_ABS_PROB%	(for each top scored taxa) The absolute (prior) probability of this taxon
%I_TAXA_REL_PROB%	(for each top scored taxa) The relative (posteriori) probability of this taxon