Microbiology and Artificial Intelligence

Partida Hanon, Angélica Inés

Madrid, March 2nd 2009

Abstract:
We are going to find out how important and interesting Artificial Intelligence in Biological Sciences can be. We are going to build a knowledge-base that is going to have all the dichotomous keys to identify a problem bacterium.

How can a machine be capable of doing the investigator's work in identifying a bacterium? From a simple way, we're going to ask this question in the following two ways: human reasoning by the expert system and image recognition by the neural network.

Keywords: *Expert Systems, Artificial Intelligence, Artificial Neural Networks, Microbiology. Bioinformatics.*

Introduction

Human reasoning is a complex process and sometimes very hard to explain. During the last few years, the digital era, science has made a great leap forward simulating this human attribute, using (among other tools) Expert Systems (ES), also considered in the Artificial Intelligence area.

There are a lot of programs with different purposes that are useful for getting results that cannot be reached using conventional methods in programming that determine a list of steps in a systematic order (algorithms). However, we are going to follow a different path, using a mechanism that simulates a human expert's mental reasoning called the heuristic method.

In fact, we have to assume yearly loss of valuable work that researchers, engineers and doctors have made, either because their contracts have been terminated, they may have retired and so on. One of the higher costs at the moment of contracting new employees references their training. If we can integrate all the content (in a knowledge-base) that the human experts have developed, and such information could be available for a large number of people, it would bring us exceptional results. Such information could be updated by other human experts.

Unquestionably, printing was one of the greatest inventions of mankind. We now we have mass reproductions in literature and science, and have been able take great steps forward technologically and scientifically. Certainly with ES and its proper development, implementation and maintenance, it could help lead to the continuing progress of the 21st century.

Throughout the investigation of the project, we are going to show that an ES is much more useful and flexible, at this point in time, as we have recourse to making a program capable of reaching a conclusion based on the use of dichotomous keys.

Why use an ES?

There are a lot of advantages such as:

- Quick results
- Reliable results
- No experts or people with years of experience in an area to solve the problem are necessary
- Worldwide access to the knowledge-base
- The necessity of consultation methods that can be uncomfortable, slow and costly (such as using extensive reference material during a diagnosis)
- There are few human experts in a determined area
- Cost reduction (It would be very expensive and difficult to get an expert botanist, doctor, physicist or biologist to do what an ES can do.)

What are the ES' that now exist and how are they used?

Dendral Biology and Chemistry, for molecular structures interpretation.

Dipmeter Advisor Oil exploration.

Mycin Medical diagnosis emulator.
CADUCEUS Medical use.
CLIPS Tool created by the NASA for creating ES.
Prolog Programming language, with applications and uses in artificial Intelligence.
LISP Language for programming ES, with applications and uses in artificial Intelligence.
Others Perform different functions in diverse areas such as economics, history, etc.

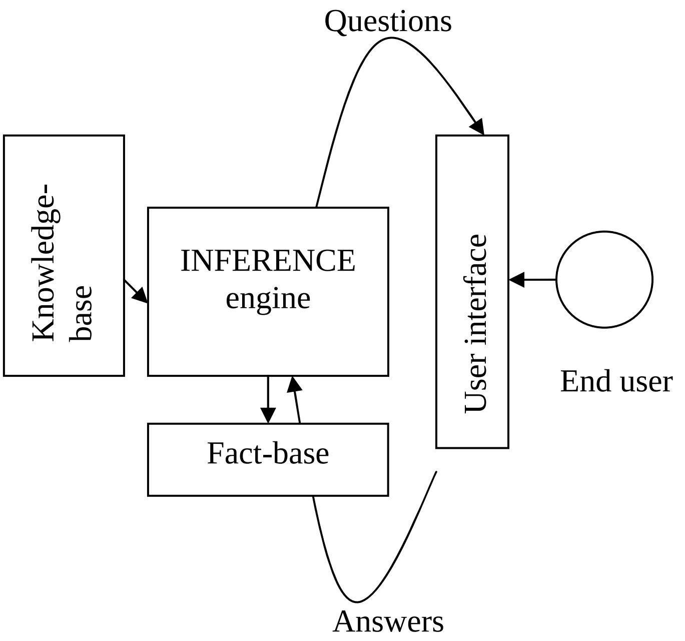


Figure 1: *ES architecture*

ES architecture:

Basically, the following elements are the components of an ES:

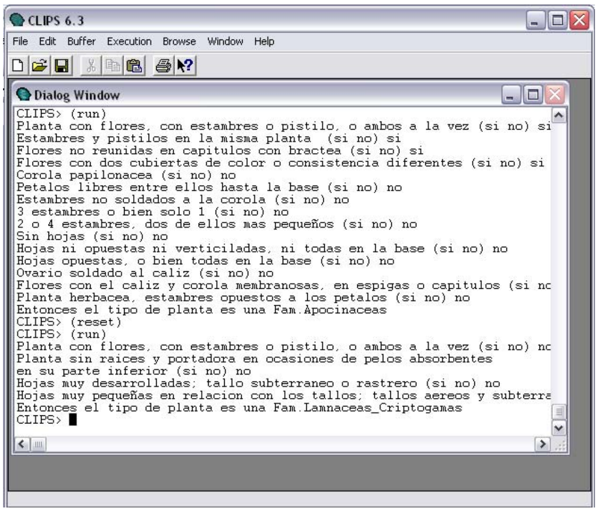
- Inference engine. - Set of instructions that make possible the ability of emulating human reasoning. It will obtain all the necessary data from the knowledge-base.
- Knowledge-base. - Contains the expert's knowledge and relates the information according to their characteristics.
- User interface. - The component that solves the interaction between the end user and the system.
- Fact-base. – Contains the intermediate instructions that were necessary to obtain the result.

Steps for developing an ES:

a) Approaching the problem.
b) Collecting information.
c) Design the Inference engine (We can already have a designed engine or shell).
d) User interface.

Methodology

We chose the CLIPS application, a free tool that lets us program an ES based on rules, developed by NASA and compatible with Windows, Linux and Mac. CLIPS also contains a shell that has been used as inference engine.



```
CLIPS> (run)
Planta con flores, con estambres o pistilo, o ambos a la vez (si no) si
Estambres y pistilos en la misma planta  (si no) si
Flores no reunidas en capitulos con bractea (si no) si
Flores con dos cubiertas de color o consistencia diferentes (si no) si
Corola papilonacea (si no) no
Petalos libres entre ellos hasta la base (si no) no
Estambres no soldados a la corola (si no) no
3 estambres o bien solo 1 (si no) no
2 o 4 estambres, dos de ellos mas pequeños (si no) no
Sin hojas (si no) no
Hojas ni opuestas ni verticiladas, ni todas en la base (si no) no
Hojas opuestas, o bien todas en la base (si no) no
Ovario soldado al caliz (si no) no
Flores con el caliz y corola membranosas, en espigas o capitulos (si nc
Planta herbacea, estambres opuestos a los petalos (si no) no
Entonces el tipo de planta es una Fam.Apocinaceas
CLIPS> (reset)
CLIPS> (run)
Planta con flores, con estambres o pistilo, o ambos a la vez (si no) nc
Planta sin raices y portadora en ocasiones de pelos absorbentes
en su parte inferior (si no) no
Hojas muy desarrolladas; tallo subterraneo o rastrero (si no) no
Hojas muy pequeñas en relacion con los tallos; tallos aereos y subterra
Entonces el tipo de planta es una Fam.Lamnaceas_Criptogamas
CLIPS> █
```

Figure 2: *CLIPS Graphical environment in Windows*

Our ES is established on a rules-based model, unlike the probabilistic model because it is the most appropriate way of identifying a problem bacterium. It is the method that the taxonomist uses, following a series of keys because both the ES and the taxonomist:

- Employ active rules (the presence of a character may be essential for choosing one way or another).
- The knowledge is acquired and updated by new rules.
- There is a chain back and forth (Kingdom, Phylum, Class, Order, Family, Genus and Species, or the presence of certain determinant characteristic).
- At the moment of identifying the species, it is recognized with 100% accuracy.

The logical topology that follows bacterium identification (or another taxonomic system) from the kingdom to the genus identification could be defined as a binary tree (Figure 3a), while from the genus to the species remains a binary-type topology (Figure 3b).

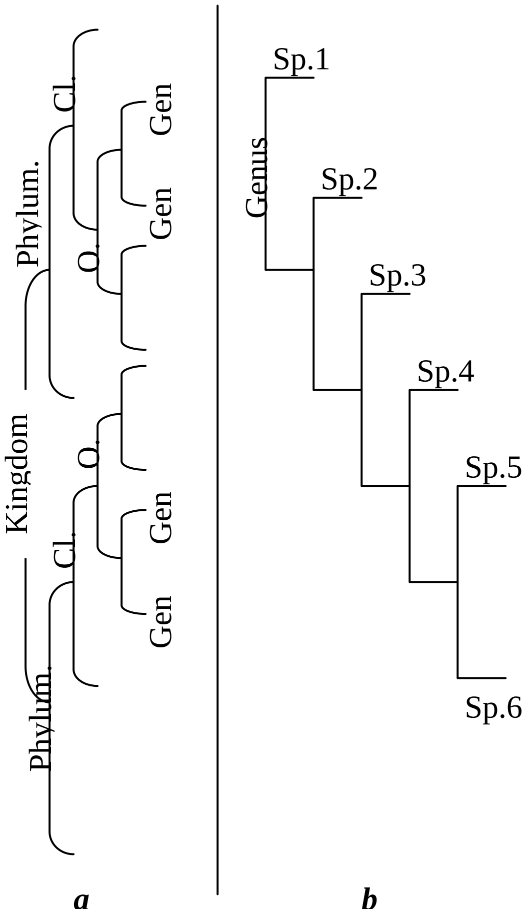


Figure 3: *ES Topology*

The following steps were defined for developing an ES:

- Define the work to be done
- Get the information from the human expert
- Design
- To choose the interaction level between the user and the system
- To choose the language
- Primary development (pseudocode)
- Testing
- Error correction, maintenance and knowledge enhancement

As an example of the ES functioning, we're going to identify the *Staphylococcus aureus* bacterium.

Results

The code is composed over 290 lines in the component of the knowledge-base, that, in some cases can identify the species and in some others the genus. There is also the possibility of adding new information about the bacterium so it can be identified.

At the beginning of the knowledge-base, we defined the goal and the only possible answers that support the system.

```
(goal is bacterium.type)
(legalanswers are yes no)
```

The basic structure of each module is the following set of instructions:

```
(rule (if previous.question is this.module and
question.name is yes)
 (then next.module is module.name))
(rule (if previous.question is this.module and
question.name is no)
 (then bacterium.type is bacterium.name))
 (question question.name is "Question")
```

Where:

- previous.question: refers the identifier of the question that relates to the actual module.
- this.module: refers to the actual module.
- question.name: refers to the identifier of the current question, the ID name should not be duplicated.
- next.module: if the system has to lead a next module, it mentions the successor module name that will have the same structure.
- module.name: not duplicable identifier of the successor module (depending on the case).
- bacterium.type: is the goal, once reached this point, the process have been finished.

As it is handled in the keys, the modules follow the same systematic, but with the most discriminating characters and shorter questions. For modules with three or more branches, it is necessary to create a new module for each branch, following the systematic with the legal answers "yes, no".

At the end of the consultation, all the followed steps are shown, which serve as the fact-base. To carry out the consultations, follow the next steps:
Load and select the location of the ES, reset and run:

```
CLIPS> (load "Bacteria.clp")
CLIPS> (reset)
CLIPS> (run)
Gram negative bacteria (yes no)
```

1st Experiment: Bacterium identification of *Staphylococcus aureus*

2nd Experiment: Bacterium identification of the Genus *Leuconostoc*.

3rd Experiment: Bacterium identification of the Genus *Salmonella*.

Fact-bases:
1st Experiment:
Gram negative bacteria (yes no) no
Cocci (yes no) yes
Catalase + (yes no) yes
Immobile (yes no) yes
Oxidative and fermentative metabolism (yes no) yes
Nitrates + (yes no) yes
Mannitol + (yes no) yes
Then, the bacterium type is Staphylococcus aureus

2nd Experiment:
Gram negative bacteria (yes no) no
Coccus (yes no) yes
Catalase + (yes no) yes
Glucose gas + (yes no) yes
Then, the bacterium type is Genus.Leuconostoc

3rd Experiment:
Gram negative bacteria (yes no) no
Cocci (yes no) no
Oxidase + (yes no) no
Lactose + (yes no) no
Methyl red + (yes no) yes
Gelatin + (yes no) no
Then, the bacterium type is Genus Salmonella

Artificial Neural Networks Annex

Methodology

The main idea we have demonstrated at the beginning of this project the usefulness of one area of AI, which is the use of expert systems. We have developed a system that identifies a bacterium according to the reasoning followed by an expert microbiologist, but until now, we have obtained a response based on a set of assumptions. Our new approaches are the following:

How can we use the bioinformatics methods in pattern recognition to identify a bacteria or colony morphology?

How can we ensure that our system is capable of learning to recognize new patterns?

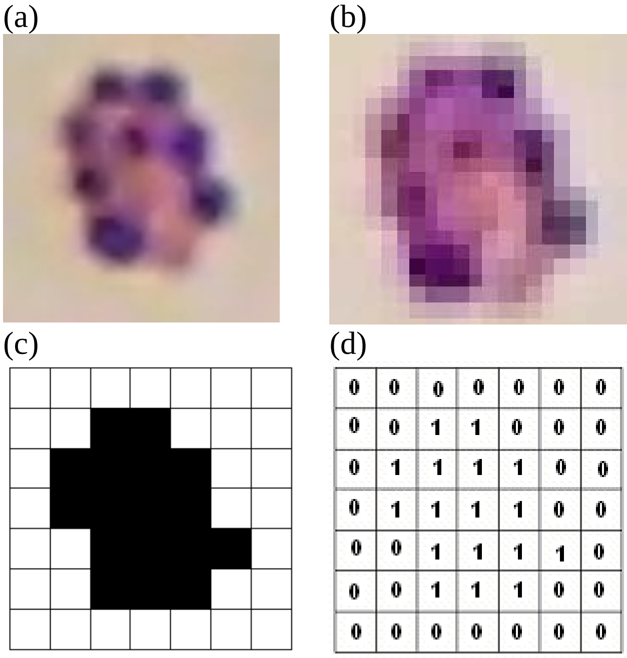
We were able to carry out a short abstract experiment showing how the learning process of an Artificial Neural Network (ANN) makes it capable of solving the above approaches.

In our example of an ANN, we define a model of two layers: one group of input neurons and another one of output neurons.

The input is responsible for receiving information, while the output shows the result based on the information processed by the input neurons (other ANN models may include an intermediate layer that is responsible for the processing of data). The ultimate goal is that, based on a series of input data, the network can be able to recognize the pattern without making mistakes.

Example:
If we have the image of a cluster of S*taphylococcus aureus;* after scanning the image and converting it into a data matrix that contains only the values 0 or 1; there may be the possibility of distortion the image, the ANN has to be able to recognize such distorted images because the network has been programmed to do so.

If, in the cluster, there is a part of a cocci that is not shown, the network must be able to recognize it as a cocci cluster and not as bacilli. **Figure 4** shows the process of transforming information to introduce its interpretation by the network:



0	0	0	0	0	0	0
0	0	1	1	0	0	0
0	1	1	1	1	0	0
0	1	1	1	1	0	0
0	0	1	1	1	1	0
0	0	1	1	1	0	0
0	0	0	0	0	0	0

Figure 4: **Steps for patterns recognizing**

The image a shows the photograph obtained from a cocci cluster of *S. aureus.* In **b**, the image is scanned and pixel; in **c**, it is translated into a matrix with a dial of 7 x 7 (the size of the dial may vary depending on the number of neurons available for recognition); in **d,** the values are translated into a binary code that will be introduced in the program.

Once given the values of the matrix, it’s time to show what we want the network to recognize.

Results

To run the network, open the program QBASIC.EXE, press Alt + A and then with the arrow keys, select “Open...” then press Tab to select the file RED.BAS. Now press Shift + F5 to run.

According to **Figure 4d,** we introduce a pattern with seven neurons that could memorize a row: the “input vector;” can be provided either from first to last. Subsequently, the network will request the input values. These values are the input to be recognized by the network depending on the parameters contained in the image that will be recognized.

4th Experiment: Recognition of the pattern in the third row of the matrix as shown in **Figure 4d.** The input value has the same values as the matrix base.

Result:
INPUT = 0 1 1 1 1 0 0
OUTPUT = 0 1 1 1 1 0 0
Pattern recognized because the output vector is the same as the recognition matrix.

5th Experiment: Distorted values are inserted, the network must be able to recognize it even with an error, giving an output value of 0 1 1 1 1 0 0

Result:
INPUT = 0 **0** **0** 1 1 0 0
OUTPUT = 0 **1** **1** 1 1 0 0
Again, pattern recognized because the output vector is the same as the recognition matrix.

6th Experiment: Another distorted value is inserted; the network must be able to recognize it again

Result:
INPUT = **1** **0** 1 1 1 0 0
OUTPUT = **0** **1** 1 1 1 0 0

7th Experiment: Rejection pattern, the data introduced represents the bacilli-type morphology.

Result:
INPUT = **1** 0 0 0 0 0 **1**
OUTPUT = 0 0 0 0 0 0 0
Rejection, is not coccus.

Conclusions

The usefulness of Artificial Intelligence in microbiology has been demonstrated.

At the end of the experiment, the taxonomic content has been digitized but the biggest significant difference is that it can expand knowledge. It can also merge two ES into one and have a wider knowledge-base. For example, a single system that determines plants, bacteria, protozoa and animals could compile the contents of a large number of books in a few bytes of information.

We must remember that the system does not understand the real meaning of the symbols nor the character chain. The ES does not need to show us all the data for giving us a result at the end of the experiment; it will show all the steps that have been followed to reach the conclusion.

Another advantage for applying this ES is that having done the first one, doing another system is increasingly simple and robust, plus anyone can add or correct data without having to alter the inference engine code.

Undoubtedly, since this began, we have expanded the system to new horizons in the biological sciences. We continue supporting the idea of asking professors and researchers to make an outline with the contents they have developed over their careers, and then add them into the base. This explains why some ES developers are universities and government administrations.

Using ANN, we can also demonstrate the great potential of pattern digitization for the recognition of the morphology of bacteria; this may also have an application in virology for morphological analysis.

Similarly, we have shown in experiments that ANNs are capable of detecting a type of bacteria recognized as cocci despite having some distortion at the time of inserting the input vector, no doubt we can eliminate interpretation errors, this is evident by the results of experiments 4, 5 and 6: cocci = 0111100.

Simultaneously we can work with the network and the expert system; the network interprets the information throughout pattern recognition, while the system is capable of giving results based on the information entered by the network. Likewise, the network could recognize the results of biochemical test and store them in a file once it gets the preliminary results; the system is going to be responsible for getting a final result.

A bigger development of these two technologies would be a great advance in science.

References

For further information, visit http://DNAngelica.com

PARTIDA HANON, A. (2009). *Sistema Experto para Determinar Plantas Vasculares.* In Reduca (Biology). Bioinformatic Series. 2 (1): 1-69.

LAHOZ-BELTRA, R. (2004). *Bioinformática. Simulación, Vida Artificial e Inteligencia Artificial.* Díaz de Santos Editions.

GAMAZO, C. (2005). *Manual Práctico de Microbiología.* Ed. Masson.

PRESCOTT, HARLEY, KLEIN. (2004). *Microbiología 5º ed.* Ed. Mc. Graw Hill.

MADIGAN, M., MARTINKO, J., PARKER, J. (2004). *Brock. Biología de los Microorganismos 10º Ed.* Ed.Pearson Prentice Hall.

CASTILLO, E., ÁLVAREZ, E. (1989). *Sistemas Expertos. Aprendizaje e incertidumbre.* Ed. Paraninfo.

ALONSO, J., LAHOZ-BELTRA, R., BAILADOR, A., LEVY, M., DIAZ-RUIZ, R. (1992). An Expert System to Classify Plant Viruses. In *Binary*, Vol. 4, 195-199)

CLIPS *A Tool for Building Expert Systems* http://clipsrules.sourceforge.net/