GENETIC NETWORK REPRESENTATION USING A SUITABLE IDE SOFTWARE TOOL TO THE BACILLUS *MYCOBACTERIUM TUBERCULOSIS*

Pablo del Moral, Sandra María Gómez, Jorge Navarro, Fernando Arroyo

Abstract: *Nowadays the new "network sciences" work together with the systems biology trying to provide computational tools for analyzing large information derived from bioinformatics and 'omics' science, such as enzymes, complex, and gene networks. The Aragon Institute of Health Sciences ICS has developed an attenuated vaccine that offers better protection against tuberculosis than BCG. In order to improve the study of this large number of genes, it is necessary to use network analysis and visualization programs. In this paper, we present the development of a new integrated IDE software tool to represent the Genetic Network of the bacillus tuberculosis. This tool provides visual, intuitive and relevant information, helping in the research and the study of the new vaccine.*

Keywords*: Systems Biology, Network Sciences, Genetic Network, IDE software Tool*

ACM Classification Keywords*: D.1.m Miscellaneous – Natural Computing*

Introduction

Biological Systems has appeared very recently (Kitano, 2001; Ideker, Galitski & Hood, 2001) and it is a discipline that seeks to integrate the vast information derived from the disciplines of bioinformatics and "omics" (genomics, proteomics,etc.) through computational models accompanied by a deep review of the classical theories of biological systems (physiology in particular). It aims to provide direction to the accumulation of data "omics" to facilitate a better understanding of biological processes in successive organizational levels of integration. Given the power of existing mathematical tools and the increasing computing power, this young discipline is broadening its own scope well. For example, already this discipline is applied in different models such as prokaryotic cell ("Virtual Cell" Project), metabolic networks, signaling and transcription in eukaryotes (Kitano, 2005) and also it starts to apply to infectious biological systems such as tuberculosis (Young et al., 2008).

Recently, important discoveries have shown a common pattern of self-organization that emerges again in different technological systems (Internet, mobile networks, the World-Wide-Web), biological (food chains, protein networks, genetic and metabolic), and social. Somehow, the interactions between the elements that make up these systems result in networks that share a common feature, large number of elements. These are known as Complex Networks (Boccaletti et al. 2006). The vast majority of relationships between biological compounds exhibit a network structure. Therefore, the new "network sciences" must cooperate with the systems biology view to provide further tools for network analysis of enzyme complexes, gene expression, and so on. Nowadays, we know the structure of these networks determines their function and vice versa. Models that simulate the effect of deletion (knockout) system components let you know what the consequences of these actions to system functionality and design laboratory experiments. In addition, the tools of network science to analyze the dynamics of tuberculosis at different scales: a cellular level, the network of proteins that is present in the phoP gene, at the population level, the dynamics of the different variants of the bacillus and finally social, how the bacillus spreads according to the existing socials interactions (Gomez-Gardeñes et al. 2008).

On the other hand, Tuberculosis, AIDS and Malaria remains being one of the most deadly infectious diseases on the planet. It is estimated that annually produces more than 500,000 children deaths and more than 2 million deaths in the adult population. One feature making the tuberculosis a disease with difficulties to control is the ability of *Mycobacterium tuberculosis* to process signals from the host and stay latent(about 1/3 of the world population), about one in 10 latent infections eventually progresses to active disease which, if left untreated, kills more than 50% of those infected hosts. The current vaccine against tuberculosis, called BCG (live attenuated strain derived from Mycobacterium bovis) remains the most widely used worldwide, but their degree of protection is variable and insufficient against the respiratory issues (Young & Die, 2006; Martin, 2006).

The Mycobacteria Genetics Team, Department of Microbiology and Public Health, University of Zaragoza has developed an attenuated vaccine that offers better protection against tuberculosis than BCG (Martin et al., 2006). This vaccine consists of a mutant strain of M. tuberculosis SO2 call, characterized by the phoP gene inactivation, which attenuates virulence markedly since the phoP gene regulates a large number of genes of tubercle bacilli, including complex membrane lipids related to virulence, as well as other factors and various cell cycle processes. The research project of the Institute of Health Sciences of Zaragoza has had as main objective expanding the network of transcription Balazsi proposed in 2008 (the most complete so far) [Marijuán, 2010].

The starting network (TR) was created using 3 different types of resources. The core of the network TR, 381 genes, consists of regulator-gene interactions documented in the literature, 222 of which have been collected in MtbRegList (Jacques et al, 2005). This first network was expanded, including 223 pairs of genes from M. tuberculosis which are relative to the TR network in E. coli (Babu et al, 2006). Finally, the research team increased the network based on the complete list of operons of M. tuberculosis (Roback et al, 2007), assuming that the transcription factor binding to the promoter region affects the expression of the genes within an operon.

The network nodes 783 TR correspond to genes of M. tuberculosis and their protein products, while the 937 to 45 links are transcription factors that directly regulate gene expression. It should be noted that 29 of these 45 transcription factors regulate their own expression, demonstrating the importance of self-regulation in prokaryotic gene networks (Thieffry et al, 1998).

The increase in the number of transcription factors per genome is translated into an increased genetic network connectivity, which in turn is correlated with a greater complexity of the organism (Levine & Tjian, 2003). Again, all this is illustrated in the new network, which contains 94 transcription factors.

In order to improve the study of this large number of genes, it is necessary to use network analysis and visualization programs. However, there aren't specific tools for this purpose, so it is necessary to develop tools capable of interpreting these networks and their characteristics. This work consists to develop a software integrated environment to support these purposes.

State of Art

Nowadays there are several tools available for gene networks representation. Among the tools used by ICS researchers there are two in particular that fit better due to network and software characteristics: Pajek and GraphViz. We are studied both tools in order to extract their characteristics as we define below.

- *Pajek:* Pajek (Slovene word for Spider) is a program, for Windows, for analysis and visualization of large networks. It is freely available, for noncommercial use. Pajek started development in November 1996 and is developed in Delphi (Pascal). Some procedures were contributed by Matjaz Zaver Snik.

  The main motivation for development of Pajek is that currently there are several forms that represent large networks. Pajek should provide tools for analysis and visualization of such networks: collaboration

networks, organic molecules, like proteins in receptor interaction networks, genealogies, Internet networks, the spread (AIDS, news) networks, etc.

The design of Pajek is based on previous experience earned in the development graph data structures and algorithms libraries Graph graphics and X-Graph, and visualization programs such as Stran, RelCalc, Draw, Energ, and SGML-based NetML.

- *GraphViz:* is open source graph visualization software. The Graphviz layout programs take descriptions of graphs in a simple text language, and make diagrams in useful formats, such as images and SVG for web pages, PDF or Postscript for inclusion in other documents; or display in an interactive graph browser. (Graphviz also supports GXL, an XML dialect.) Moreover, these layout programs are based on descriptions of pictures in plain text, allowing them to be edited by users and you don't need an additional program for that. The diagrams are made in various formats: images (jpg or png), SVG (Scalable Vector Graphics, two-dimensional vector graphics) to web pages, Postscript for inclusion in PDF or other documents, or they can be represented in an interactive browser , where the user can edit (Graphviz also supports GXL, Graph eXchange Language).

  Graphviz has many features to customize the diagrams such as options for labels, colors, fonts, tabular layouts, line styles, links and forms. In practice, the graphics are typically generated based on external data sources, but can also be done manually, by editing a plain text file in DOT language or using a graphical editor.

  This software has many useful features for concrete diagrams, such as options for colors, fonts, tabular node layouts, line styles, hyperlinks, rolland custom shapes. Grappa is a Java graph drawing package that simplifies the inclusion of graph display and manipulation capabilities within Java applications and applets. It has a good number of useful features built into it, but is also extensible.

None of these tools are capable of represent properly the genetic network because we don´t have the correct format files. Therefore it is necessary to develop tools that could parser properly the scientific documents.

Software architecture defining the IDE

The development phase has been carried out with a main purpose, to bring useful tool to the research team. The chosen programming language is Java, using Grappa Java package provided by Graphviz.

Java is designed to support applications that will be executed in the most varied network environments, from Unix to Windows NT, and Mac via workstations on different architectures and different operating systems. To accommodate such diverse performance requirements, the compiler generates Java byte codes: an intermediate format indifferent to architecture, designed to transport code efficiently to multiple hardware platforms and software. The remaining problems are solved the Java interpreter.

The indifference to architecture represents only part of its portability. In addition, Java specifies the sizes of its basic data types and behavior of the arithmetic operators, so that programs are the same on all platforms. These last two features are known as Java Virtual Machine (JVM).

Grappa mainly provides:

- Construct Methods, manipulate and display the graphs of the nodes, edges and subgraphs.
- Each graphic element may be associated with a number of arbitrary name and a value.
- Grappa includes methods for reading and writing graphs using the text format. Dot.

Some Java libraries have been added, such as Java Swing, which handle input files and output files. The Figure 1 shows the software architecture defined for our IDE.

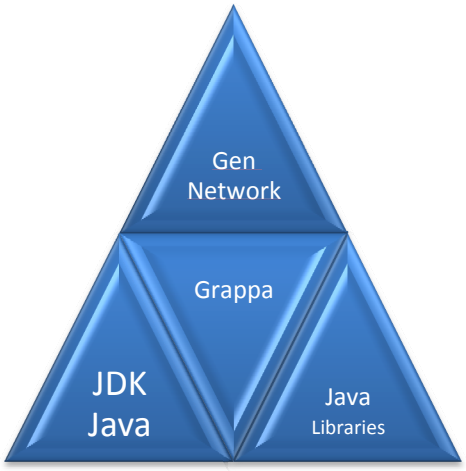


Figure 1. *Software Architecture for IDE proposed.*

Class design

The analysis and specification of classes has been made to create a conceptual design of information will be handled in the system and components to take charge of performing the relevant operations.

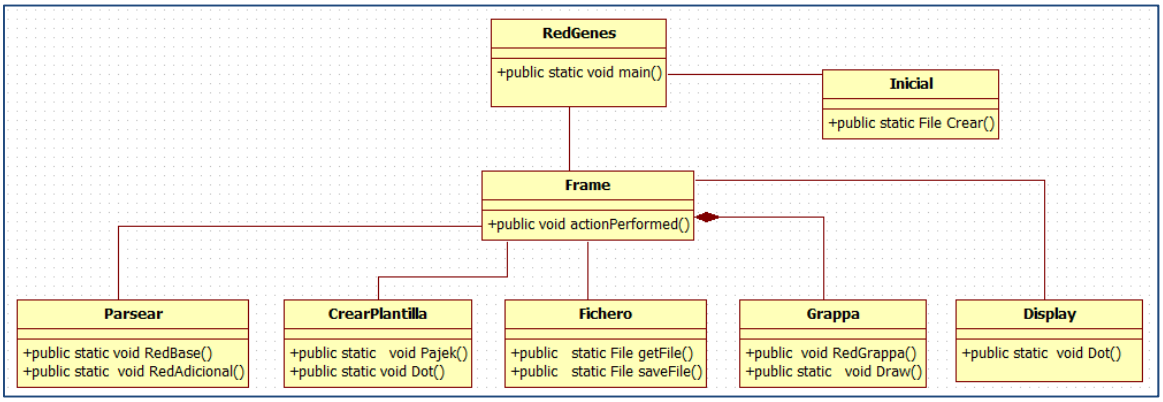


Figure 2. *Diagram of high-level classes*

As explained above, one of the reasons for choosing this work environment has been that Graphviz provides a library for making Java applications.

The Frame class is a child class of JFrame, and uses methods ActionListener (extend JFrame, implements ActionListener). Frame class will be responsible for interfacing with the user and invoke methods of other classes implemented (Parse, CrearPlantilla, File, Display) to achieve the desired functions.

As can be seen Grappa class has a composition relationship with the Frame class, this is because in this way can display the graph created by the methods of the screen Grappa library generated by the Frame class.

Grappa class, use the methods provided by the development team Graphviz, showing graphs implemented by our tool. RedGenes is the main class responsible for invoking the main frame, using a default graph generated at the initial class.

The graphical interface has been designed according to the requirements set by the research team, providing a usable and intuitive graphical interface being very simple as is illustrated in the next figure.

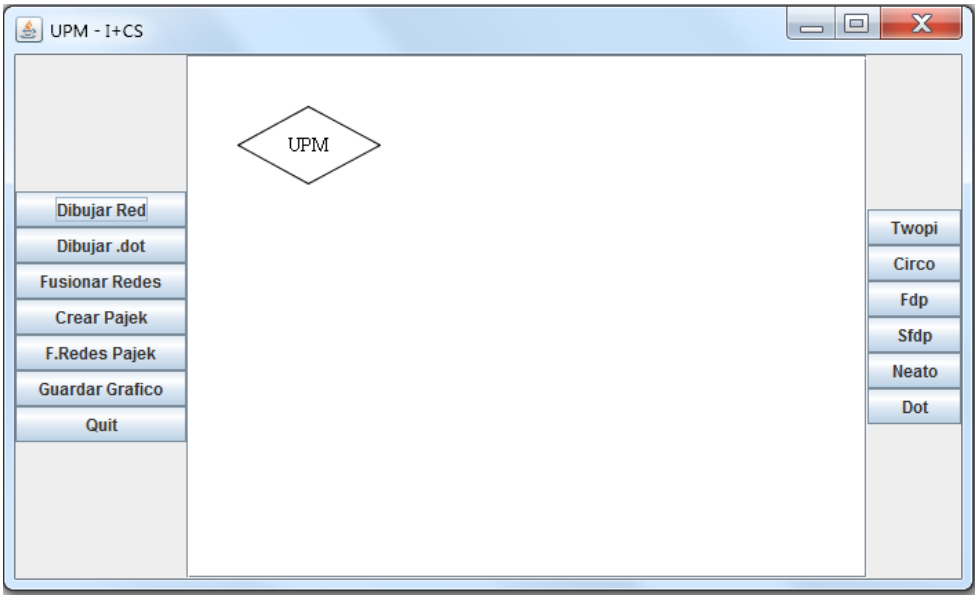


Figure 3. *The main window of the program.*

For the correct behavior of the application, it has been determined that the input file, which represents the network to draw, has a strict data format.Txt file that contains two columns separated by spaces and / or tabs, which represents the relationship between genes.

Input files:

- Txt-document that represents the genetic network;
- Dot- Graphviz supported format.

Output File:

- Dot- Graphviz supported format;
- Net- Pajek supported format;
- Jpg- Genetic network image.

The main features of the program:

1. Draw Network (input file provided by the research team). Using the .txt file provided by the research team, the tool draws the network, adding the functional gen information. Also generated the Graphviz file.
2. Draw Network, GraphViz format input file. Using the .dot file provided at the point 1, the tool draws the network, with all the functional gen information.
3. Mix two Networks. After select two networks provided by the research team, the tool draws the mixed networks, giving different functional information for each genetic network.
4. Change Network Layout. You can choose between those differents layouts:
   - Twopi: Radial layouts. Nodes are placed on concentric circles depending their distance from a given root node;

- Circo: Circular layout. This is suitable for certain diagrams of multiple cyclic structures, such as certain telecommunications networks;
- Fdp: "Spring model" layouts similar to those of neato, but does this by reducing forces rather than working with energy;
- Sfdp: Multiscale version of fdp for the layout of large graphs;
- Neato: "Spring model" layouts. This is the default tool to use if the graph is not too large (about 100 nodes) and you don't know anything else about it. Neato attempts to minimize a global energy function, which is equivalent to statistical multi-dimensional scaling;
- Dot: "Hierarchical" or layered drawings of directed graphs. This is the default tool to use if edges have directionality.

5. Save Genetic Network as jpeg image.
6. Provide pajek format files (.net) with the same specifications. In order to serve all the research requirements, the tool also provides pajek support, giving the pajek files format with all the functional information included, even you can generated the file that has two networks mixed.

Example: representation of a genetic network for the bacillus *Mycobacterium tuberculosis*

The genetic network of the bacillus *Mycobacterium tuberculosis* is one of the main points in the research of the new Tuberculosis vaccine. This IDE software tool provides these main characteristics necessary to analyze and study the behavior of the network representing the Genetic information. The input nodes, which have unknown transcriptional regulator, are shown in blue. The transit nodes, with known transcription regulators, are shown in green. The white nodes represent output nodes, genes that encode proteins with no transcriptional activity. The triangular nodes self-regulate their own expression. When two Networks are mixed, the common genes are shown in red, and the new genes are shown in purple as is illustrated in the next figures.

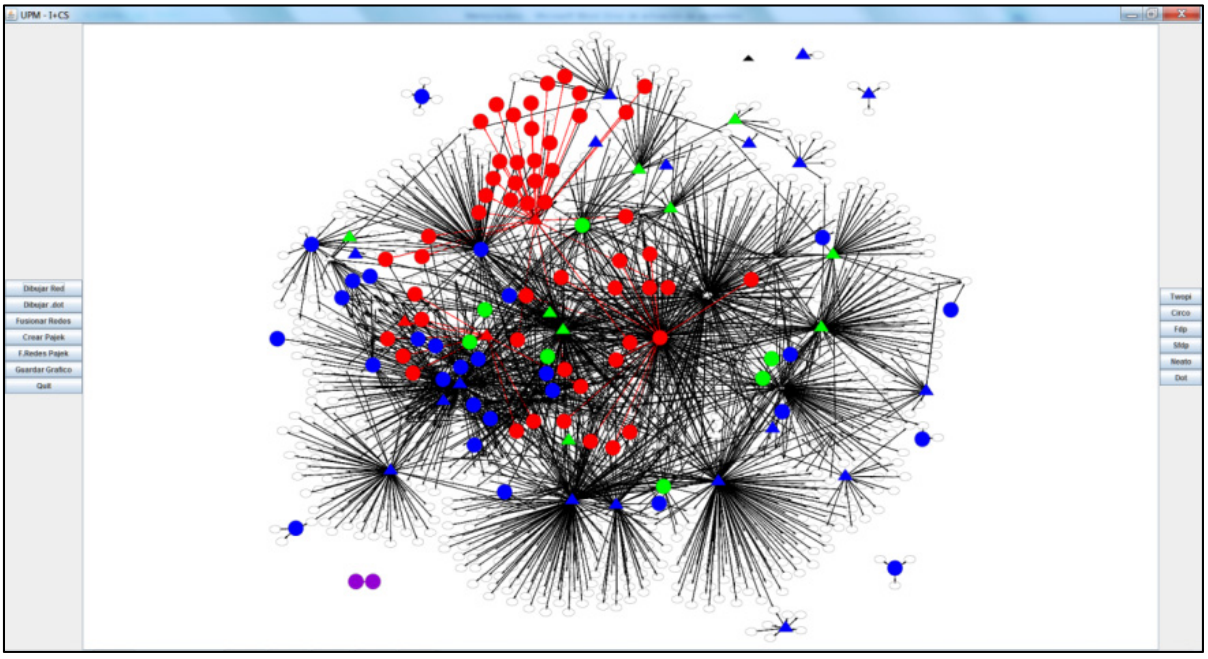


Figure 4. *Tuberculosis Network generated by the program*

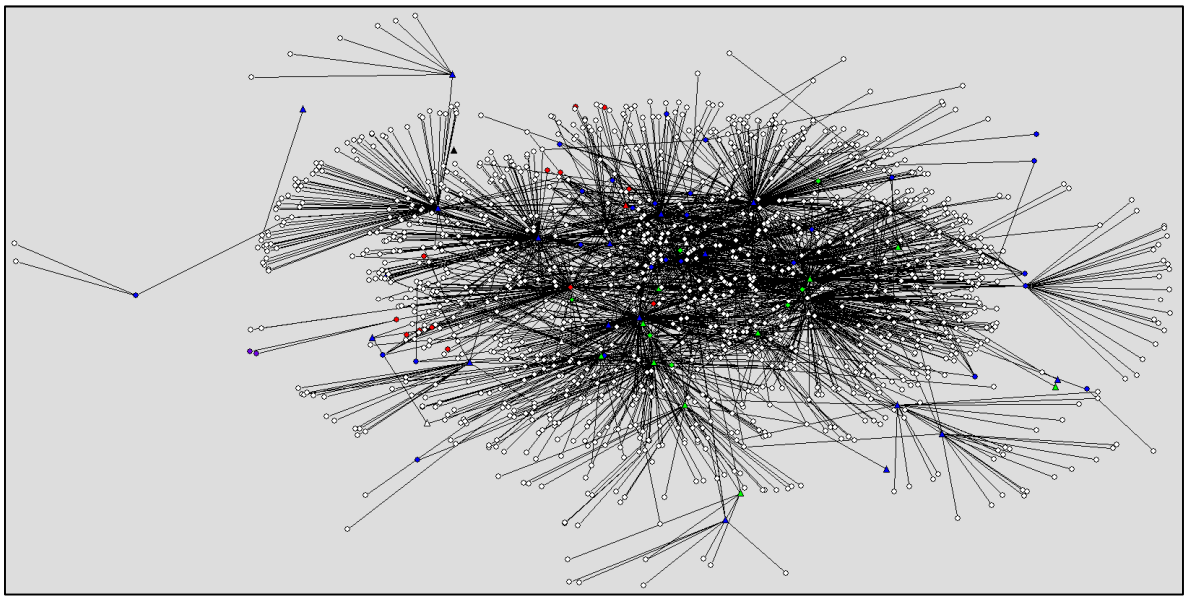


Figure 5. *Tuberculosis Network generated by Pajek, using the file provided by the program*

Conclusions and future work

Mostly, research conducted by the Institute of Health Sciences of Aragon, has allowed the expansion of the new network by increasing the number of genes by 79%, the number of links in a 145% and the number of transcription factors by 109% compared to the most comprehensive network to date (Balazsi et al., 2008), this network provides a much more dynamic integrated gene expression of M. tuberculosis.

The development of new vaccine based on phoP, with the support of this new tool, is in Clinical Trial, and it can be ready for 2016. In support of this study, we have implemented the visualization tool of Genetic Network, providing another tool for the advancement of this research.

The tool complies with the prospects of the research team; however there are still several ways to collaborate in this field. The collaboration possibilities between a research team and a developer team are very spacious and can provide tools in order to get the scores. For optimal collaboration, we consider that is necessary to get close to the research team, understanding their goals, and give a technical point of view.

Bibliography

[Babu, 2006] Babu M, Balaji S, Aravind L. General trends in the evolution of prokaryotic transcriptional regulatory networks in Genome Dynamics (Ed: Jean-Nicolas Volff), 3:66-80 (2006).

[Boccaletti, 2006] S. Boccalettia, V. Latorab, Y. Moreno, M. Chavez , D.Hwanga. Complex networks: Structure and dynamics. Elsevier Physics Reports 424 (2006) 175 – 308.

[Gomez-Gardeñes, 208] Gomez-Gardeñes J, Latora V, Moreno Y, Profumo E. Spreading of sexually transmitted diseases in heterosexual populations. PNAS 105: 1399-1405 (2008).

[Ideker, 2001] Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. Annu Rev enomics Hum Genet. 2: 343-72 (2001).

[Jacques, 2005] Jacques, P.-E., Gervais, A. L., Cantin, M., Lucier, J.-F., Dallaire, G., Drouin, G., Gaudreau, L., Goulet, J. & Brzezinski, R. MtbRegList, a database dedicated to the analysis of transcriptional regulation in Mycobacterium tuberculosis. Bioinformatics 21 (2005), 2563–2565.

[Kitano, 2001] Kitano H. Systems Biology: Toward System Level. In foundations of Systems Biology, ed. By Kitano. The MIT Press, Cambridge (2001).

[Kitano, 2005] Kitano H. International alliances for quantitative modeling in systems biology. Mol Syst Biol.1: 2005.000 (2005).

[Levine & Tjian, 2003] Transcription regulation and animal diversity. Nature 424: 147-51 (2003).

[Marijuán, 2010] Marijuán P., Navarro J. & del Moral R. On prokaryotic intelligence: strategies for sensing the environment. Biosystems 99: 94-103 (2010).

[Martin, 2006] Martin C., Williams A., Hernández-Pandoc R., Cardona P., Gormley E., Bordat Y., Soto C., Clark S., Hatch G., Aguilar D., Ausina V., Gicquel B. The live Mycobacterium tuberculosis phoP mutant strain is more attenuated than BCG and confers protective immunity against tuberculosis in mice and guinea pigs. Elsevier Vaccine 24 (2006) 3408–3419.

[Roback, 2007] Roback P., Beard J., Baumann D., Gille C., Henry K., Krohn S., Wiste H., Voskuil M., Rainville C., and Rutherford R. A predicted operon map for Mycobacterium tuberculosis. Nucleic Acids Res. August; 35(15): 5085–5095. (2007)

[Thieffry, 1998] Thieffry, D., Salgado, H., Huerta, A. M. and Collado-Vides, J. Prediction of transcriptional regulatory sites in the complete genome sequence of Escherichia coli K-12. Bioinformatics 14, 391-400. (1998)

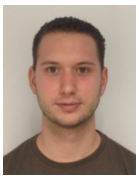
[Young, 2008] Young D, Stark J. & Kirschner, D. Systems biology of persistent infection: tuberculosis as a case study. Nature Reviews Microbiology 6: 520-8 (2008).

[Web references, 2012]:

Graphviz & Grappa Site. Website: http://pajek.imfm.si/doku.php.

Pajek Site. Website: http://www.graphviz.org/.

Authors' Information



Pablo del Moral Antón – Software Engineer of the Technical University of Madrid. Documentation Exchange Collaborative Business Network, Dokify. C\ Margarita Salas 12. Leganés. Madrid - Spain; email: pdelmoral@dokify.net

Sandra María Gómez Canaval *– Natural Computing Group, Department of Languages, Projects and Computer Systems, University College of Computer Sciences, Technical University of Madrid, Crta. de Valencia Km 7, 28031 Madrid – Spain; email: sgomez@eui.upm.es*

Jorge Navarro López*– Bioinformation and Systems Biology Group, Aragon Institute of Health Sciences (I+CS,) Zaragoza, Spain; e-mail: jnavarro.iacs@aragon.es*

Fernando Arroyo Montoro *– Natural Computing Group, Department of Languages, Projects and Computer Systems, University College of Computer Sciences, Technical University of Madrid, Crta. de Valencia Km 7, 28031 Madrid – Spain; email: farroyo@eui.upm.es*