A Review On Pathway Analysis Software Based On Microarray Data Interpretation

Abdul Hakim Mohamed Salleh[1], Mohd Saberi Mohamad[1*],
Safaai Deris[1] and Rosli Md. Illias[2]

[1]Artificial Intelligence and Bioinformatics Group,
Faculty of Computing, Universiti Teknologi Malaysia,
81310 UTM Skudai, Johor, Malaysia.
[2]Department of Bioprocess Engineering, Faculty of Chemical Engineering,
Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia.
abdhakim.utm@gmail.com, saberi@utm.my, safaai@utm.my, r-rosli@utm.my
**corresponding author*

Abstract

Recent advancement in microarray technologies and large high throughput data generated has made it very challenging to decipher and draw a feasible biological conclusion from current microarray experiments. The difficulty arises when the number of samples available for analysis is smaller than the huge numbers of genes that need to be considered. Currently, pathway analysis is a preferable tool in extracting and understanding the biological information obtained from high throughput experiments. It is essential to analyze microarray experiments along with their biological information to represent the underlying structure of the biological network. Currently, there are numerous software developed for pathway analysis available with the same goal of mining the information from the microarray experiments with biological relevance over the extensive amounts of data. This paper discusses the comparisons between pathway analysis software in terms of their performance, advantages and limitations as well as the available pathway databases in terms of their data availability and organization. The aim of this review is to provide a better understanding of the capabilities of these software and helps to select the tools most suited for a particular purpose.

Keywords: *Analysis software, biological pathway, microarray data, pathway analysis, pathway database*

1. Introduction

Systems biology aims at a system-level understanding of biological system. It is an approach to understand how different parts of the systems assemble dynamically to give rise to a unique metabolic response. Different from the traditional molecular biology approach, each part is isolated and analyzed individually without concerning the effects of the dynamic interactions. However, the terms “system-level understanding” is rather hard to be defined in this context because the system is not a tangible object.

Generally, it consists of four distinct phases that lead to a system level understanding of various levels of a biological system [1]. First, system structure identification enables us to understand the structure of the system which can identify both physical and interaction structures. Interaction structures are represented as pathways comprising of gene networks and biochemical networks that describe how components interact within and between cells.

Second, the dynamics of the systems with high dimensional space need to be understood due to its enormous complexity and chemical properties of each component. Third, a method to control the system has to be investigated so that it will be able to precisely predict and control the effect of the system. Finally, by designing the system, for instance by modifying and constructing biological systems with the design features for example bacteria may be redesigned to yield desired properties for particular drug production.

This understanding process requires a top down approach that essentially breaks down the system to gain insight into its compositional sub-systems and analyze how they operate at the same time [2]. One of the major fields in systems biology research is the pathway analysis of microarray data. Microarray technology can simultaneously identify genes involved in a particular process by measuring the expression of a large number of genes [3].

The common result obtained from microarray experiments are a long list of differentially expressed genes between two groups of samples compared (*e.g.*, normal and diseased) as well as the estimated expression changes between groups. The functionality of a cell of an organism is monitored by selective expression of the genes [4]. In order to get a better understanding of the biological significance of the data, it is relevant to include the available pathway information to analyze the genes under study. Pathway information provides insight into the biological processes underlying within microarray data. In this review, we discussed about the existing and commonly used tools and software for pathway analysis using microarray data mainly invented in the recent years.

We discuss these software tools designed to facilitate such analyses with the assumption that this information is useful for a better understanding and designing more focused experiments regarding the biological pathways to reveal and analyze their significance in a biological process, diseases and therapeutics [5].

Over recent years, researchers are focusing more on the single gene analysis technique that may give accurate identification and classification without taking consideration of the underlying structure of the global network. However, this single gene analysis technique cannot accurately interpret the biological meaning of the differentially expressed genes, as it does not consider the dynamic interaction of the genes within the complex network. Therefore, to address such limitation, a new approach of pathway analysis of gene expression data was developed. A biological pathway that comprises of coordinated sequence of biochemical reactions is a small segment of the overall biochemical network that contribute to a specific function.

However, a complete biological network is so huge and has a high complexity that could hide the key pathway that actually responsible to produce a specific biological response [6]. Therefore, an appropriate and effective model to extract and identify the pathway is needed and at the same time takes account of the biological features and interactions between the components of the pathway so that the real underlying structure of the system can be precisely obtained.

To overcome the challenges, many software were designed with the capability to detect consistent but subtle changes in gene expression by incorporating either pathway or functional annotations. In addition, to ensure the analysis is biologically plausible, these software utilized the publicly available databases to obtain the pathway information and data.

These databases provide visualization of the pathway in diagrams, which combine metabolic, signaling and regulatory network based on curation and literature. Here we discuss some of the widely used software tools for pathway analysis as well as the databases that are used to extract the biological information.

2. Pathway Analysis Software for Microarray Data

There are many software for pathway analysis available where each of them has unique features and advantages which is suitable for a particular experiment. Here we review some of the software and discussed each them in terms of functionality, method used and the output of the software. The basic information about the software is shown in Table 1.

2.1. ArrayXpath [7]

Table 1. Pathway Analysis Software

TOOL	Method	Script / Platform	URL
ArrayXPath	Fisher exact test; Multiple testing correction	Web	http://www.snubi.org/software/ArrayXPath/
Pathway Miner	Fisher exact test	Web	http://www.biorag.org/pathway.html
SPIA	p-value, false discovery rate	R	http://vortex.cs.wayne.edu/ontoexpress/
PathRanker	3M, HME3M	R	http://www.bic.kyoto-u.ac.jp/pathway/timhancock
MAPPFinder	Standardized difference score (z) from hypergeometric distribution.	Windows	http://www.GenMAPP.org

2.1.1. Functionalities: One of the comprehensive tools for pathway analysis and visualization is ArrayXpath that is a web-based tool for microarray gene expression profile mapping and visualization from biological pathway resources. The tool is suitable for a range of gene identifiers such as LocusLink [8], Swissprot [9], GenBank [10], and UniGene [11] and also includes pathway integration from the well-known publicly available databases as data source which includes KEGG [12], GenMaPP [13] and BioCarta.

2.1.2. Method: ArrayXpath can automatically recognize the probe identifiers from submitting data and maps onto the pathway database then calculate the statistical significance of the association of the two data. ArrayXpath utilize Fisher's exact test along with the False Discovery Rate method for multiple hypothesis testing correction as the statistical evaluation method to produce a list of highly relevant pathways of each cluster with statistical significant scores of non-random association.

2.1.3. Output: The visualization of the results is using JavaScript-enabled Scalable Vector Graphics (SVG) that enable users to analyze at each pathway node level by animation features with color coded expression level and cluster membership. The graphical display of the result also allow a few user controls such as zooming for better navigation and option to choose the particular experimental condition to be viewed. An improved version of ArrayXpath, ArrayXpath II [14] is available with several additional capabilities such as identification of disease related pathways and Gene Ontology (GO) based annotations.

2.2. Pathway Miner [15]

2.2.1. Functionalities: Another web based tool for pathway analysis is the Pathway Miner that able to extract and display pathways from pathway databases by using client application written using a Java swing API that run on Java run time environment version 1.4 or higher to gain access to databases including KEGG, BioCarta or GenMAPP. The tool provides a user-friendly environment that allows viewing, analyzing, interpreting and downloading pathway information and networks based on the association with the databases, suitable for high throughput analysis of gene expression data.

2.2.2. Method: Similar to ArrayXpath it uses Fisher's exact test to rank the pathways, but without correction for multiple hypothesis testing.

2.2.3. Output: The output is an organized pathway profiles extracted from each of the databases in HTML format, which is one of the advantages of Pathway Miner that allows a flexible organization of either to display sample specific pathway profiles or comparisons between different samples or mining the most frequent pathway associations or list of genes involved in a particular pathway or multiple pathways.

2.3. PathRanker [16]

2.3.1. Functionalities: PathRanker is a pathway analysis tool written in the R language for statistical computing and graphics visualization an available as an R package [17]. This R package provides a detailed pathway analysis of gene expression data with annotation from KEGG to identify the most significant biologically meaningful pathway within the huge global metabolic network. PathRanker allows user to process KEGG metabolic network, extract most relevant metabolic pathways as well as the key functional components of the particular pathways. The flexibility of the package also allows the analysis of pathways within the complete KEGG network of the specific species or more specific analysis of pathways between specific compounds. Another advantage of PathRanker is that it can significantly identify biologically meaningful pathways that correctly represent the underlying structure of metabolic networks.

2.3.2. Method: PathRanker uses 3 main techniques which are pathway ranking, clustering and classification with the implementation of 2 algorithms 3M Markov Mixture Model [18] and the extension of 3M, Hierarchical Mixture of Experts 3M (HME3M) [19].

2.3.3. Output: The visualization of the clustering results is presented in a heat map with the number of pathways associated for each clustering components. The heat map consists of the gene names associated with the pathways as well as the compound and reaction involved with the probability value that indicates the significance of the gene to the pathway. The results of the classification technique are presented in a directed pathway graph that allow user to display connected genes, compounds, reactions or pathways of desired biological response. The edge thickness represents its importance to the whole network. An example of results obtained using PathRanker can be seen in Figure 1.

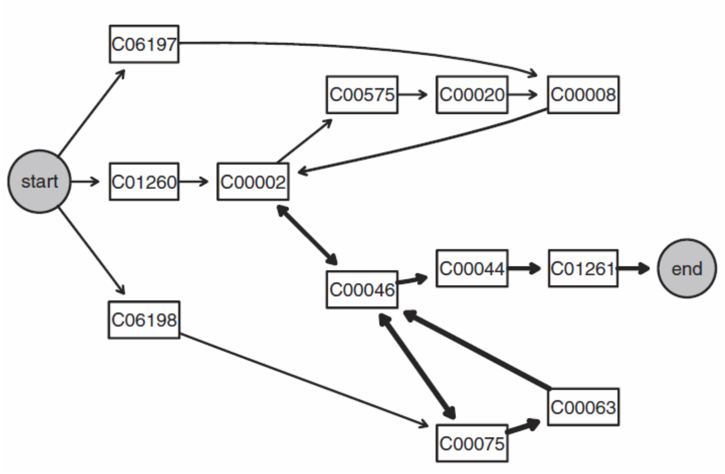

Figure 1. An Example of Result obtained by PathRanker which shows Connected Compounds that Made up a Particular Pathway [16]

2.4. SPIA [20]

2.4.1. Functionalities: Signaling pathway impact analysis (SPIA) is a tool to measure the actual perturbation on a particular pathway and improve specificity and sensitivity of several widely used pathway analysis method that is implemented as an R package. This package provides a technique for pathway analysis based on a combination of two types of evidence, the over-representation of differently expressed genes and the perturbation of the pathway as measured by expression changes.

2.4.2. Method: For each pathway, a p-value is calculated. With the assumption that the number of differentially expressed genes in a pathway (NDE) follows a hyper-geometric distribution, the probability of the number of differentially expressed genes in the given pathway more than observed (PNDE) is calculated. The second probability, PPERT, is calculated based on the estimated amount of perturbation in each pathway due to the differential expression of the input gene list.

2.4.3. Output: The output includes a table of signaling pathways containing at least one of the genes on the input list that summarizes the impact of the differentially expressed genes of each pathway based on PNDE and PPERT. SPIA package also produces a summary plot of those pathways that represent where the most impacted pathways lie between its (-log transformed) values for PNDE and PPERT with respect to two statistical thresholds, the Bonferroni or the family-wise error rate represents by oblique red line and false discovery rate represented by an oblique blue line.

2.5. MAPPFinder [21]

2.5.1. Functionalities: MAPPFinder is a tool capable to identify global biological insights in gene-expression profile with the integration of annotations from the Gene Ontology (GO) Consortium [22]. MAPPFinder allows user to define criteria for significant changes to each term in the GO hierarchy. This tool is an accessory for (Gene MicroArray Pathway Profiler)

GenMAPP that assist in the identification of important biological processes. The results are displayed in a browser, allowing users to explore and identify the color-coded GO terms with over-represented the numbers of gene expression changes.

2.5.2. Method: MAPPFinder assigns each of the genes in the dataset to the corresponding GO terms by using a relational database and the gene-association files from GO. It then calculates the percentage of genes that meet with the user desired criteria and a statistical z score for each GO term. The results are then produced as ranked list and displayed in GO browser. From the result, users can directly identify interesting GO terms with high levels of gene-expression changes. The methodology is summarized in Figure 2.

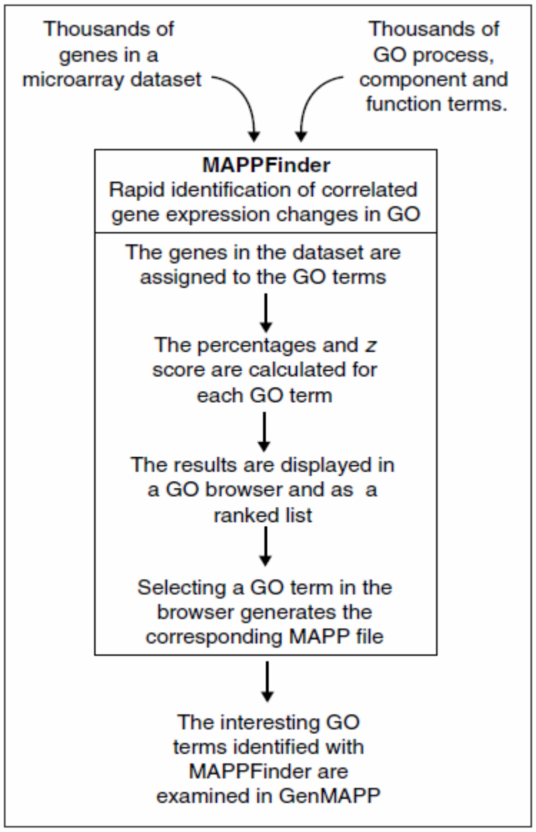
Thousands of genes in a microarray dataset
Thousands of GO process, component and function terms.
MAPPFinder
Rapid identification of correlated gene expression changes in GO
The genes in the dataset are assigned to the GO terms
The percentages and z score are calculated for each GO term
The results are displayed in a GO browser and as a ranked list
Selecting a GO term in the browser generates the corresponding MAPP file
The interesting GO terms identified with MAPPFinder are examined in GenMAPP

Figure 2. The Workflow of MAPPFinder [2]

2.5.3. Output: The results are displayed in a GO browser for analysis based on the GO hierarchy as well as tab delimited test files that can be exported into a spreadsheet program for further analysis and data manipulation. The GO terms are displayed in a color coding manner where each color represents certain threshold values that determine the membership of the gene in the array.

3. Pathway Databases as the Source for Biological Data

In order to extract or identify pathways that are statistically significant and biologically plausible it is very crucial to incorporate information from biological repositories such as the pathway databases.

However, the existences of diverse pathway databases are making it extremely difficult to carry out even a simple search across the databases. The inconsistencies in the representation of biological data and different data manipulation mechanism have created an uncertainty towards the accuracy and consistency of the pathway information [23]. For instance, there are no specific and consistent definitions of "MAPK Cascade" pathway from the literature hence raising the question of which genes to be included in the pathway.

Hence obtaining consistent pathway information from diverse data sources with different data formatting scheme is a very big challenge. Therefore, it is very critical to have a tool that enables the integration between these databases and standardizing the formatting scheme to obtain consistent pathway information for future references.

Over the past decades, the effort to document, organize and standardize the ever-expanding knowledge of biological pathways has brought to an increase in the number of pathway databases. Some of the databases provide a wide range of selection between organism while some are organisms-specific for example EcoCyc [24] which only covers the genome and biochemical pathways of E. coli MG1655 [25]. The available formats along with the capability of data manipulation also differ between databases. These available databases can generally be divided into three main categories based on the types of information provided which is metabolic pathway databases such as KEGG, BIOCYC and Metacyc [26]; signaling pathway databases and databases that covers both such as KEGG, BioCarta and Reactome [27].

One of the most widely used databases is the KEGG developed by researchers at Kyoto University back in 1995. The use of KEGG as the database source of interacting components within a cell has received a major attention from researchers due to the vast range of available information across different organisms and the ease of extracting information from the database itself. The KEGG's advantages lie in the facts that it collects annotated genomes across different species and provide a generic view of the pathway and the components involved thus enhancing the possibility to extend the software to cover a huge number of species. The rationale behind KEGG is to collect and organize all available information gathered from experimental observations in a computationally accessible format that enable researchers to develop new computational analysis for extracting, comparing and computing pathways.

Nevertheless, the existence of alternative databases such as Reactome, which provide more detailed information has started to attract researchers to a new data source rather than just one in particular. Reactome has a wider range of capabilities from pathway browser, pathway analysis, cross species comparison and expression analysis of either protein of encoded genes and enable cross-reference with NCBI Entrez Gene, Ensembl and UniProt databases, the UCSC and HapMap Genome Browsers, the KEGG Compound and Chevy small molecule databases, PubMed, and Gene Ontology.

4. Discussion and Conclusion

The enormous amount of information generated by high throughput profiling technologies has contributed to new challenges in data analysis. Several methods have been developed to overcome the challenges where these methods provide an integrated functional approach to microarray analysis. However, with the rapid growth in these methods, many software has been designed to solve the problem at a certain level and extract the desired information.

Pathway analysis of microarray data has become one of the most reliable tools in the interpretation of biological knowledge within the gene expression data. Most of the software tools are able to analyze pathways statistically by using annotation from available pathway databases such as KEGG, Biocarta and GenMAPP.

This review discussed a number of pathway analysis software in terms of the functionalities; methodology and the result that can be obtained using the software. The future of pathway analysis should also be focusing on the prediction of novel pathways in predicting orthologous pathways across different organism or predicting homologous pathways for an organism of interest [28]. This will be essential in field such as genetic engineering for strain improvement towards the production of desired products [29-31].

Data integration of diverse biological information ability will add tremendous value for the software that is going to be developed in the future. The used of diverse input data, including sequence data, expression data and known interaction data contain valuable information for prediction and analysis of the pathways [32]. This will helps to enhance pathway analysis capabilities and establish a standard for pathway structures and pathway database designing and organization.

Future development of pathway databases should also include pathway-level search engine rather than conventional gene, compound or reaction query to retrieve the significant pathways. Software tools that can retrieve a complete set of pathways with biological significance of a given query such as the starting and target compound, the particular genre of a biological response or a biochemical reaction of interacting compound will be essential in the new generation of pathway analysis.

New development of software tools for pathway analysis must address these challenges collectively to boost pathway analysis to a new level that is able to utilize the high throughput technologies in order to understand the huge biological systems and ensure more significant, relevant, and precise results can be obtained.

Acknowledgement

We would like to thank Malaysian Ministry of Higher Education for supporting this research by an Exploratory Research Grant Scheme (Grant number: R.J130000.7807.4L096) and a Fundamental Research Grant Scheme (Grant number: R.J130000.7807.4F190). This research is also funded by an e-science research grant (Grant number: 06-01-06-SF1029) from Malaysian Ministry of Science, Technology and Innovation.

References

[1] H. Kitano, “Introduction to Systems Biology”, Edited Choi, S, Humana Press Totowa, New Jersey, **(2007)**, pp. 3-13.

[2] Y. Ruolin and S. Bing, Current Bioinformatics, vol. 4, no. 3, **(2009)**, pp. 207-217.

[3] K. R. Sharma, “Bioinformatics: Sequence Alignment and Markov Models”, (1st ed.), McGraw-Hill, New York, **(2009)**.

[4] A. Fuente, P. Brazhni and P. Mendes, Trends in Genetics, vol. 18, no. 8, **(2002)**, pp. 395-398.

[5] Z. Wenguang, L. Jinquan, S. Rui and J. Wu, Current Bioinformatics, vol. 4, no. 3, **(2009)**, pp. 242-248.

[6] V. Hatzimanikatis, C. Li, J. A. Ionita, C. S. Henry, M. D. Jankowski and L. J. Broadbelt, Bioinformatics, vol. 21, no. 8, **(2005)**, pp. 1603-9.

[7] H. J. Chung, C. H. Park, M. R. Han, S. Lee, J. H. Ohn, J. Kim, J. Kim and J. H. Kim, Nucleic Acids Res., vol. 33(Web Server issue), **(2005)**, pp. W621-W626.

[8] K. D. Pruitt, K.S. Katz, H. Sicotte and D. R. Maglott, Trends Genet, vol. 16, no. 1, **(2000)**, pp. 44-47.

[9] C. O'Donovan, M. J. Martin, A. Gattiker, E.Gasteiger, A. Bairoch and R. Apweiler, Briefings in bioinformatics, vol. 3, no. 3, **(2002)**, pp. 275-284.

[10] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and D. L. Wheeler, Nucleic Acids Research, vol. 36, **(2008)**, pp. D25-D30.

[11] G. D. Schuler, J Mol Med., vol. 75, no. 10, **(1997)**, pp. 694-698.

[12] M. Kanehisa, M. Araki and S. Goto, Nucleic Acids Res, vol. 36, **(2008)**, pp. D480-484.

[13] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor and B. R. Conklin, Nat Genet, vol. 31, **(2002)**, pp. 19-20.

[14] H. J. Chung, C. H. Park, M. R. Han, S. Lee, J. H. Ohn, J. Kim, J. Kim and J. H. Kim, Nucleic Acids Research, vol. 32 (Web Server issue), **(2004)**, pp. W460-W464.

[15] R. Pandey, R. Guru and D. W. Moun, Bioinformatics, vol. 20, **(2004)**, pp. 2156-2158.

[16] T. Hancock, I. Takigawa and H. Mamitsuka, Bioinformatics, vol. 269170, **(2010)**, pp. 2128-2135.

[17] R. Gentleman, V. Carey, W. Huber, R. Irizarry and S. Dudoit, “Bioinformatics and Computational Biology Solutions Using R and Bioconductor”, Springer, **(2005)**.

[18] H. Mamitsuka Y. Okuno and A. Yamaguchi, “SIGKDD Explorations”, vol. 5, no. 2, **(2003)**, pp. 113-121.

[19] T. Hancock and H. Mamitsuka, “Workshop on Algorithms in Bioinformatics (WABI)”, **(2009)**, pp. 30-40.

[20] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J. Kim, C. J. Kim, J. P. Kusanovic and R. Romero, System Biology, vol. 25, no. 1, **(2009)**, pp. 75-82.

[21] S. W. Doniger, N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor and B. R. Conklin, Genome Biology, vol. 4, no. 1, **(2003)**, R7.

[22] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight and J. T. Eppig. Nat Genet., vol. 25, **(2000)**, pp. 25-29.

[23] S. Donny, D. Difeng, G. Yike and W. Limsoon, BMC Bioinformatics, vol. 11, no. 449, **(2010)**, pp. 1-16.

[24] P. D. Karp, M. Riley, M. Saier, I. T. Paulsen, J. Collado-Vides, S. M. Paley, A. Pellegrini-Toole, C. Bonavides and S. Gama-Castro, Nucleic Acids Res., vol. 30, no. 1, **(2002)**, pp. 56-8.

[25] A. M. Feist, C. S. Henry, and J. L. Reed, Mol Syst Biol., vol. 3, no. 121, **(2007)**.

[26] P. D. Karp, M. Riley, S. M. Paley and A. Pellegrini-Toole, Nucleic Acids Res., vol. 30, no. 1, **(2002)**, pp. 59-61.

[27] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney and L. Stein, Nucleic Acids Res., vol. 33, **(2005)**, pp. D428-32.

[28] A. Cho, H. Yun, J. H. Park, S.Y. Lee and S. Park, BMC Syst Biol., vol. 4, no. 35, **(2010)**.

[29] J. L. Reed, R. S. Senger, M. R. Antoniewicz and J. D. Young, J Biomed Biotechnol, 207414, **(2010)**.

[30] H. Ma and A. P. Zeng, Bioinformatics, vol. 19, no. 2, **(2003)**, pp. 270-277.

[31] G. Stephanopoulos, Science, vol. 315, no. 5813, **(2007)**, pp. 801-804.

[32] M. W.Covert, E. M. Knight, J. L. Reed, M. J.Herrgard and B. Ø. Palsson, Nature, vol. 429, **(2004)**, pp. 92-96.