PROM-OOGLE – Data mining and Integration of On-line Databases to Discover Gene Promoters

Dean Cheng[1], John Sheldon[1], Marcelo Marcet-Palacios[2], Osmar R. Zaïane[1,*]

[1] Department of Computing Science, University of Alberta, Canada
{dcheng, sheldon, zaiane}@cs.ualberta.ca

[2] Department of Medicine, University of Alberta, Canada
marcelo@ualberta.ca

Abstract. The vast number of on-line biological and medical databases available can be a great resource for medical researchers. However, the different types of data and interfaces available can be overwhelming for many medical researchers to learn. Moreover, the available resources lack needed integration. Here we focus on an important task in medical research: to provide researchers with promoter analysis for a given gene. PROM-OOGLE is a web based data mining tool that provides a means for researchers to take a gene name of interest and obtain its promoter sequence in return after automatic integration of text databases. Additionally, the program is capable of returning multiple promoters from different genes allowing researchers to study how promoters regulate genes. This tool facilitates the process of acquiring information on a promoter and may lead to interesting discoveries.

1 Introduction

The permeation of computer science into all aspects of research is growing every day. None as much as biology where computers are aiding in the acquisition of meaningful results from data that was never thought possible. The field of bioinformatics has become a powerful area in which a tremendous potential for great research is being seen. As a result, the vast amount of on-line biological and medical databases available can be a great resource for medical researchers. Examples of such databases include sequence databases such as GenBank [1], literature databases such as MedLine [2], chemical databases such as PubChem [3], transcription factor databases such as TESS [4]. To fully take advantage of these databases, one has to have the proper biomedical background as well as familiarities with the different interfaces. This can be overwhelming for many medical researchers. Moreover, the interfaces of the different on-line databases are neither standardized nor designed for integration. When the output of a query into a first database or part of it is required as input for a search in a second database, the operation needs to be done manually. This process is

* Contact author

laborious, time consuming and prone to errors. To alleviate the above-mentioned difficulties, data mining tools can be used to automate tedious searching procedures.

Developing data mining tools for biological applications is a growing field with vast potential [5]. There are many types of analysis that can be done across multiple types of data and as a result, there are many types of tools being developed. Tools range across from whole genome sequence analysis such as Vista Genome Browser [6] to information extraction from text for gene-gene/gene-disease associations such as MedMiner [7] and MedGene [8]. Our system, PROM-OOGLE, is a web content mining tool that tries to integrate sequence data and gene-promoter associations in order to help medical researchers with the important task of finding promoters of a gene. The unique functionality that PROM-OOGLE has to offer is to compare promoter sequences within a given species and in order to improve our understanding of promoter structure and similarities among genes that are co-regulated in similar ways.

In essence, PROM-OOGLE is a data mining tool that can facilitate researchers move through tedious work more efficiently and with less uncertainty with the validity of their results. It allows researchers to generate meaningful biological hypotheses with greater ease that can be tested in laboratory experiments with the aim of understanding better how gene expressions work and to eventually develop methods to counteract health problems. It is data mining because the tool combines parts of individual database outputs from different resources in order to generate new information that is not explicitly present in any individual database and presents it in a manner to support understanding and discovery of new knowledge regarding genes and their promoters.

2 Motivation

2.1 Background Knowledge

As humans, we are essentially made up of cells. These cells contain 23 pairs of chromosomes, which are compact intertwined molecules called deoxyribonucleic acids (DNA). A DNA strand is composed of linearly linked nucleotides that are subsequently linked with one of four bases: adenine (A), thymine (T), guanine (G), and cytosine (C). On each strand of DNA, in specific positions, are genes, which contain hereditary information. Genes control the production of proteins, which are made up of amino acids. The production of proteins is controlled by a sequence of nucleotides at the very beginning of the gene, called a promoter sequence. We defined a promoter sequence as the 2000 nucleotides upstream of a gene sequence. When other proteins bind themselves to this section of nucleotides, this can either cause the gene to produce a protein or inhibit the gene from producing a protein. These proteins are often called transcription factors or promoters. Depending on which kinds of proteins attach themselves to the binding sites of the promoter sequence and depending on the order of which they attach themselves, a specific kind of protein is produced by the

gene. The human body is a complex machinery of biological molecules functioning together. Irregular levels of gene expressions of any gene could lead to the development of a disease. Therefore, understanding how gene expression is regulated is an important task for combating diseases. In other words, investigating how the promoters regulate gene expressions is one important task in medical research.

2.2 Current State of the Art

Facilitating the acquisition of binding site information for the promoter sequence of a gene for Homo sapiens (human) is paramount in medicine. This information could be acquired through the acquisition of the promoter sequence from the National Center for Biotechnology Information (NCBI) website [9] and then entering the sequence into the Transcription Element Search System (TESS) website [4]. Results in tabular form are returned from TESS describing hundreds of possible different binding sites (See Figure 1). From this list the researcher must locate the binding site of interest, which is not straightforward. The outputs of the different databases are heterogeneous, incompatible and with formats raging from simple text, tabulated text or sophisticated interactive visualizations with Java applets. This procedure is long and requires a large amount of time to learn and there is no guarantee of its success in finding something interesting to base further research on. Nonetheless, this procedure could yield interesting insights and new found medical understanding as shown in [10]. Using information derived from TESS, the findings in [10] suggest that NO and cGMP are necessary to up-regulate the expression of MMP-9, a gene that has been suggested to be related to pulmonary emphysema [11] among other lung conditions. This discovery could not have been made without the valuable time taken to learn the database systems and performing the numerous searches.

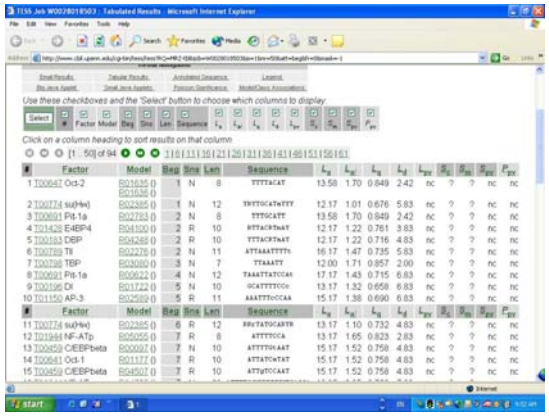

Fig. 1. Tabulated output from the Transcription Element Search System

Navigation through the NCBI website and the TESS website could be problematic if one does not have the proper background knowledge and training. For example, in order to obtain the promoter sequence from NCBI, one has to navigate through several web pages clicking the proper links and obtaining the correct data from texts or images. If this procedure could be automated so as to allow the researcher to gain results quickly then the researcher could proceed to find interesting results in hours instead of months. PROM-OOGLE implements this procedure, which can be easily run in a few minutes. The challenges are many and pertain not only to the complex integration of databases but also to subtle intricacies of the user interface. The dynamic aspect of the information returned to the user require a flexible system that could interpret the results returned by NCBI and TESS and provide a more useful display of the results. This tool would need to be complex enough to handle the data but simple enough for the user to feel confident in the results that were returned.

3 PROM-OOGLE: the proposed Solution

A web-based data mining tool, PROM-OOGLE, was designed and implemented to solve the time consuming and complex data integration problem stated above. PROM-OOGLE uses web content mining techniques to automate the procedure medical researchers typically do manually, which involves collecting the proper data and displaying the desired results. The tool consists of a set of Perl scripts that parse outputs from the on-line databases using static templates to map the structure and syntax of the output and extract specific fields and values, and a set of web-based user interfaces build with Javascript to allow valuable on-line interaction and manipulation of PROM-OOGLE's output assisting in knowledge discovery.

The application finds the corresponding gene sequence in the human genome database from NCBI [9] and subsequently retrieves a 2000 nucleotide sequence (later referred to as a promoter sequence) located upstream of the transcription (gene) start site. This promoter sequence is to be the input for the second tool TESS [4]. TESS then returns all the binding sites found in the promoter sequence and these binding sites are used along with the promoter sequence to produce an interactive report, which can be used by the researcher.

PROM-OOGLE takes the gene name provided by the researcher and automatically enters it into the NCBI database, extracts the promoter from the generated results and enters them into the TESS database. Results from the TESS database are then retrieved and parsed and are displayed in a convenient manner such that the researcher can easily investigate results with clarity. Figure 2.A shows the general flowchart of PROM-OOGLE.

3.1 Integrating and parsing on-line databases

PROM-OOGLE is a multipart tool with intricate process flow, briefly illustrated in Figure 2.B. It not only uses two existing on-line databases and integrates their results, but also carefully parses those results to automatically generate specific queries intended to those databases. This subtle process aims at simulating what a human researcher may do manually in order to discover existing knowledge about specific gene promoters. The first database used is NCBI Entrez[1], which is an integrated text-based search and retrieval system compiling a large collection of major bio-medical databases. From this system, different information is retrieved at three major steps of PROM-OOGLE search. First a gene map view is retrieved, then the coordinates of the gene within the chromosome is acquired and finally the nucleotides positioned before the gene in the chromosome are obtained. This is achieved by a series of automated parse-and-submit operations explained below. The second database used is TESS, which is a string search system for finding known binding sites in DNA sequences. The collection of all this data is used to create interactive web pages that organize and align sequences to help users better explore the available information and discover unforeseen arrangements.

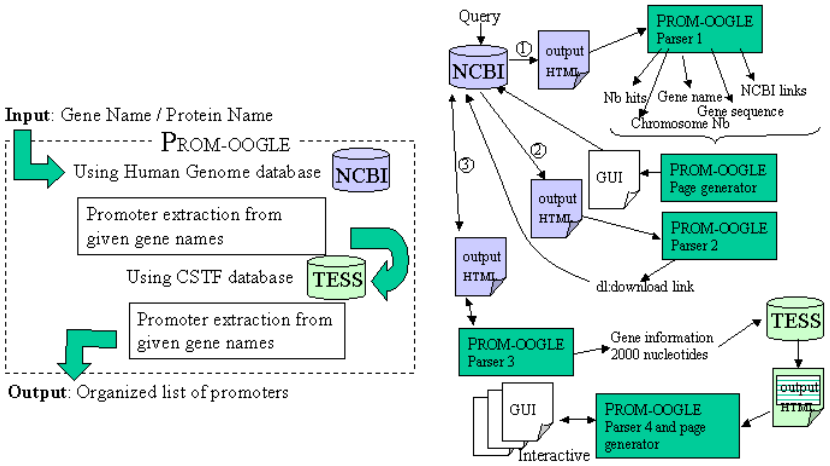

Fig. 2. A: General flowchart of PROM-OOGLE; B: PROM-OOGLE Process flow

3.2 Detailed sequence of events

With PROM-OOGLE, the user need not know the details of the NCBI and TESS databases and the semantics and syntax of their interfaces. Once a gene name is obtained

[1] NCBI-Entrez: http://www.ncbi.nlm.nih.gov/Database/

from the user the query string is submitted to NCBI Map Viewer (Homo sapiens)[2]. An example of output from NCBI using MMP9 as query is illustrated in Figure3. The page obtained from NCBI is never displayed but directly parsed by PROM-OOGLE. The parser targets the fields framed in red in Figure 3. Specifically, the page is parsed to obtain the number of hits, the chromosome number, making sure PROM-OOGLE is extracting information from the "reference" assembly, the gene name, its NCBI link and the "Genes sequence" link. This information is used to generate a summary page like the one in Figure 7-right. Upon the users choosing, PROM-OOGLE follows the "Genes sequence" link, parses the page and extracts the "download link" (dl) and follows it. This last hyperlink leads to a form as the one illustrated in Figure 4.

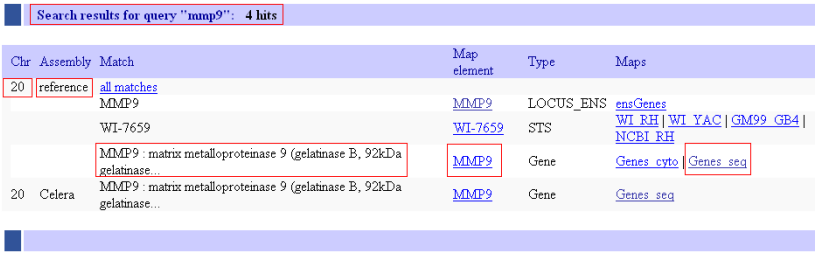

Fig. 3. Results from NCBI Map Viewer with framed targeted fields.

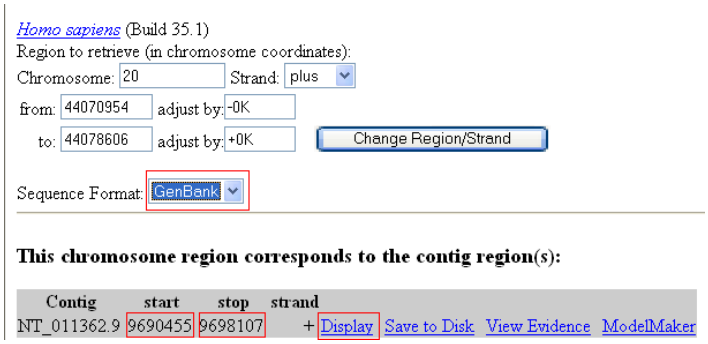

Fig. 4. Extracting information from the "dl" link.

Again, this form is not displayed but parsed and filled automatically by PROM-OOGLE. PROM-OOGLE extracts information from the "gene" feature. More specifically, the gene start and stop coordinates within the chromosome region are dug up. In addition, if the gene is on the complementary strand of the DNA, PROM-OOGLE adjusts the gene start and stop coordinates accordingly (Figure 5-left). Once PROM-OOGLE extracts enough information to know the coordinates of the gene within the chromosome's coordinate system, the "dl" page and form (Figure 4) are filled again to re-

[2] http://www.ncbi.nih.gov/map_search.cgi?taxid=9606

quest 2000 nucleotides before the gene start coordinate in FASTA format (Figure 5-right).

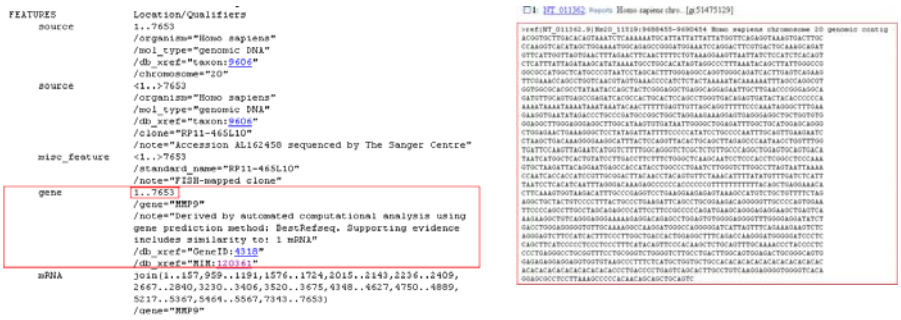

Fig. 5. Extracting gene information from the GenBank display (left). 2000 nucleotides in FASTA format (right)

These 2000 nucleotides are used to fill-in the TESS web form[3] and to obtain known protein bindings from the database. The series of obtained pages are parsed to extract relevant information to be organized and displayed by PROM-OOGLE. Figure 6 shows an example of table obtained from TESS. This table is parsed for the highlighted columns and placed into a form that can be easily used by the user. The gene is also recorded so that it can be used to reference the binding sites of the various factors.

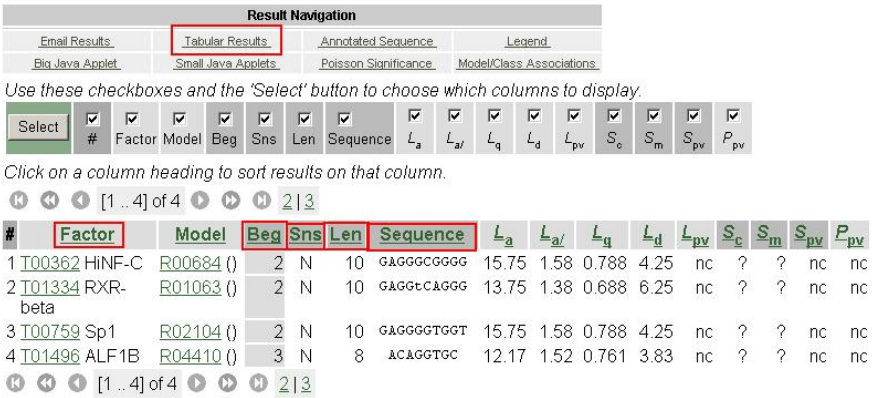

Fig. 6. Output from TESS database (the framed columns are targeted and extracted).

The next section explains the output organization by PROM-OOGLE.

[3] http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=SEA-FR-QueryS

3.3 Describing PROM-OOGLE user interface

As hinted to above, the output and input interfaces for NCBI and TES databases are completely transparent to the user who may ignore their existence. These interfaces are managed automatically by PROM-OOGLE and only its web-based GUI is presented to the user. Figure 7 shows a snapshot of the user interface of the web based mining tool. The user interface is designed to be user-friendly for medical researchers. Navigating through the PROM-OOGLE website is plain and simple while it still provides various functionalities.

Fig. 7. PROM-OOGLE initial search interface (left); First gene search results (right).

On the right of Figure 7, shows the display of returned gene matches that the user can select from after the first access to NCBI. The search for a gene name can return many hits and as such the user must ensure that PROM-OOGLE is searching the proper gene that the user has requested. The results of the gene search show the name of a possible subfamily of the gene, the short version of the name and the chromosome number. The user can then select a link to run PROM-OOGLE on the gene or can select the map element to locate the gene in the NCBI database if further information is needed pertaining to the gene. It is more important to verify that the promoter sequence returned by PROM-OOGLE is correct since wrong promoter sequences would definitely give rise to false results. During development and initial experiments, several gene names such as NOS2A, PI3K and CD8A have been verified to return biologically correct results with known correct promoter sequence.

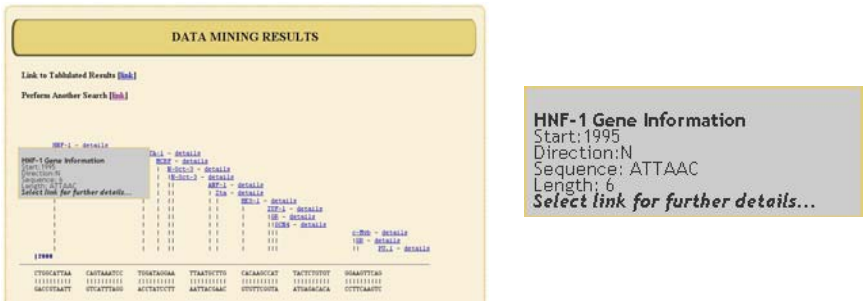

Fig. 8. Data mining results and the Basic gene information in pop down window.

Results from the data mining tool are returned in two formats: in the form of a gene sequence with the different bindings and directions; and in tabular form listing the promoters that can be sorted by different attributes interactively as the user sees fit. Figure 8 shows the promoter binding locations for the gene along with information pop down tables that provide more useful information regarding the gene and its binding site location (Figure 8-right). The link that can be taken directly from the name of the gene takes the researcher directly to the NCBI website which allows the user to access more information pertaining to that specific gene. This is very useful for the researcher because of the quick and simple access to pertinent information. Details of a promoter binding location, which include the sequence name, starting location, direction, length, and the sequence composition, can be displayed at will (Figure 9-left). It also includes the useful visual information pertaining to the exact placement of the sequence on the gene and its direction.

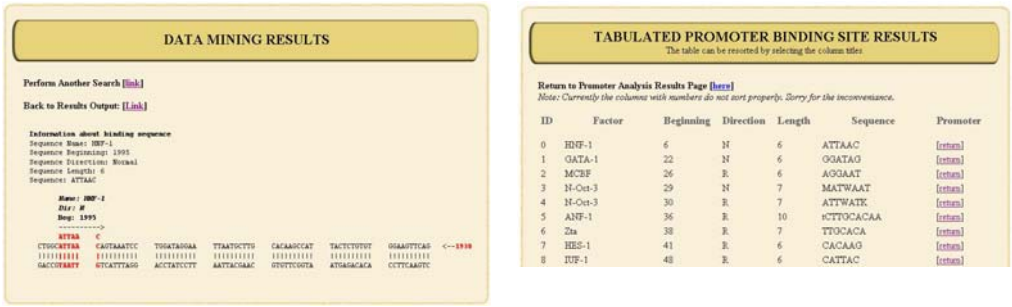

Fig. 9. Complete Gene details (left); Tabular form of PROM-OOGLE results (right)

The results in tabular form (Figure 9-right) can be sorted by any of the attributes by simply clicking on the heading. This provides the user with the ability to easily find information in the table. For example, if the user wished to find all sequences of a certain length, the user could click on "length", the list would be sorted by length, and then the user would simply need to scroll down the list to the point of where the sequence length is of the desired number.

4 Possible Extensions

Despite the functional correctness of our application, improvements to the tool are always possible and recommendations were collected after a first round of use by biomedical researchers. Possible improvements include user interface functionalities that would help the user navigate through the information provided and application functionalities that would aid the user in making more informed decisions. Other improvements are related more to over all correctness of our application and require extensive user testing to ensure.

Other extensions to the web-based tool include more user inputs to provide more information to reduce the number of transcription factor binding sites returned. This would allow the output only of the transcription factor binding sites of user's interests.

Another extension of PROM-OOGLE is to include a local database that could be used to include proteins of interest in different medical areas and alert users if the search results match some proteins in the local database. In addition, users would be able to save searches locally on their computers so that searches need not be repeated.

A final extension is the ability to run multiple searches simultaneously with results being stored for further review at a later time. This would also give the ability to perform multiple sequence alignment to find out if there are common transcription factors that are shared among a family of genes.

A concern of note is the running time of PROM-OOGLE. This is, however due to the running times of NCBI and TESS and as such we are limited to the time constraints they impose on our program.

5 Conclusion

PROM-OOGLE is a web content mining tool for promoter analysis that can facilitate researchers move through tedious work more efficiently and with less uncertainty with the validity of their results by combining query results from different existing on-line databases. It allows researchers to generate meaningful biological hypotheses with greater ease that can be tested in laboratory experiments with the aim of understanding better how gene expressions work and to eventually develop methods to counteract health problems.

Even though software tools like Vista Genome Browser [6] are capable of performing phylogenic analysis they lack the functionally of PROM-OOGLE to compare promoter sequences within a given species and improve our understanding of promoter structure and similarities among genes that are co-regulated in similar ways, therefore making PROM-OOGLE an exciting new tool for researchers.

This project is still in its infancy and there are numerous other functionalities that will be included in future developments. Preliminary feedbacks from bio-medical researchers have been very positive. This web-based tool is a promising aid to those who perform research in the area of biology and medicine.

References

1. GenBank : http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide
2. MEDLINE: http://medlineplus.gov/
3. PubChem: http://pubchem.ncbi.nlm.nih.gov/
4. Jonathan Schug and G. Christian Overton , TESS: Transcription Element Search Software on the WWW, Technical Report CBIL-TR-1997-1001-v0.0, Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania, 1997, URL: http://www.cbil.upenn.edu/tess
5. D. Page and M. Craven, (2003). Biological applications of multi-relational data mining. *SIGKDD Explorations*, **5**(1): 69-79.
6. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004 Jul 1;32 (Web Server issue):W273-9, URL: http://pipeline.lbl.gov/cgi-bin/gateway2
7. L. Tanabe, U. Scherf, L.H. Smith, J.K. Lee, L. Hunter and J.N. Weinstein, (1999). MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. BioTechniques Dec. **27**:1210-1217
8. Y. Hu, L.M. Hines, H. Weng, D. Zuo, M. Rivera, A. Richardson, and J. LaBaer, (2003). Analysis of genomic and proteomic data using advanced literature mining. Journal of Proteome Research, **2**: 405-412
9. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D.L., Khovayko, O., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Pontius, J.U., Pruitt, K.D., Schuler, G.D., Schriml, L.M., Sequeira, E., Sherry, S.T., Sirotkin, K., Starchenko, G., Suzek, T.O., Tatusov, R., Tatusova, T.A., Wagner, L. and Yaschenko, E. (2005) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 33 (Database issue):D39-45. NCBI Database: http://www.ncbi.nlm.nih.gov/
10. M. Marcet-Palacios, K. Graham, C. Cass, A.D. Befus, I. Mayers, and M.W. Radomski, (2003). Nitric Oxide and Cyclic GMP Increase the Expression of Matrix Metalloproteinase-9 in Vascular Smooth Muscle. **307**:429-436
11. Minematsu N, Nakamura H, Tateno H, Nakajima T, Yamaguchi K. (2001) Genetic polymorphism in matrix metalloproteinase-9 and pulmonary emphysema. Biochem Biophys Res Commun. Nov 23;289(1):116-9.