Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων
Department of Computer & Communication Engineering

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ

ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ ΚΑΙ ΔΙΚΤΥΩΝ

The Role of MicroRNAs in Diseases

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Κυριακής Κ. Πλωμαρίτου

ΤΟΜΕΑΣ :

Βιοπληροφορική

Επιβλέποντες καθηγητές : Μούντανος Ιωάννης
Αναπληρωτής Καθηγητής Π.Θ.
Χατζηγεωργίου Άρτεμις
Ερευνήτρια Β' ΕΚΕΒΕ Φλέμινγκ.

Βόλος, Σεπτέμβριος 2011

Στα παιδιά …

Summary

MicroRNAs (miRNAs) are small non-coding RNA molecules which normally function as post-transcriptional regulators of target mRNA expression, resulting in target mRNAs cleavage or translation inhibition. There is a lot of evidence associating miRNAs with many key biological processes, including cell growth and tissue differentiation. In fact, their role is critical and extremely important in the regulation of gene expression. As such, deregulation of microRNAs and their targets result in various diseases such as cancers and cardiovascular disease.

There are many studies that have produced a large number of microRNA-disease associations however, the acquisition of such information is becoming increasingly difficult due to the large volume and rapid growth of biomedical literature. Therefore developing Web services for helping users quickly and efficiently to search and retrieve useful information along with relevant publications have led to an increase in the number and quality of information provided for a bibliographic analysis which correlates miRNAs to diseases.

In this thesis, we automatically retrieve the associations of microRNAs with disease from literature. Using Perl scripts, all abstracts associated with microRNAs are retrieved from PubMed, the free Web literature search service of United States National Center for Biotechnology Information (NCBI). The abstracts are associated with microRNAs based on the presence of the name of the microRNA in the title or abstract of the publication.

The retrieved data are associated with Medical Subject Headings (MeSH), the controlled vocabulary of pre-defined terms of United States National Library of Medicine (NLM) and the information we extract contains microRNA names, MeSH terms, and the literature PubMed ID. This information was properly processed by the team of the DAINA Lab (DNA intelligent Analysis) of Biomedical Sciences Research Center "Alexander Fleming", so as to be used in the latest version of DIANA-microT Web server, a program for microRNA target prediction which is based on Artificial Neural Networks. It may be used to search for target

genes of miRNA sequences in order to assist biologists and researchers retrieving important data in an efficient way.

Περίληψη

Τα microRNAs (miRNAs) είναι μικρά μη-κωδικοποιητικά μόρια RNA τα οποία συνήθως λειτουργούν ως μετά-μεταγραφικοί ρυθμιστές της έκφρασης του mRNA-στόχου, με αποτέλεσμα τη διάσπαση των mRNA-στόχων ή την αναστολή της μετάφρασης. Υπάρχουν πολλά στοιχεία που συνδέουν τα microRNAs με πολλές βασικές βιολογικές διεργασίες, συμπεριλαμβανομένης της κυτταρικής ανάπτυξης και της διαφοροποίησης των ιστών. Στην πραγματικότητα, ο ρόλος τους είναι κρίσιμος και εξαιρετικά σημαντικός στη ρύθμιση της γονιδιακής έκφρασης. Ως εκ τούτου, η απορρύθμιση των microRNAs και των στόχων τους έχουν ως αποτέλεσμα τους σε διάφορες ασθένειες όπως ο καρκίνος και οι καρδιαγγειακές παθήσεις.

Υπάρχουν πολλές μελέτες που έχουν επιφέρει ένα μεγάλο αριθμό συσχετισμών των microRNAs με ασθένειες όμως, η απόκτηση αυτών των πληροφοριών γίνεται ολοένα και πιο δύσκολη λόγω του μεγάλου όγκου και της ταχείας ανάπτυξης της βιοϊατρικής βιβλιογραφίας. Επομένως η ανάπτυξη Διαδικτυακών υπηρεσιών ώστε να βοηθηθούν οι χρήστες γρήγορα και αποτελεσματικά να αναζητήσουν και να ανακτήσουν χρήσιμες πληροφορίες, μαζί με σχετικές δημοσιεύσεις, έχουν οδηγήσει σε αύξηση του αριθμού και της ποιότητας των πληροφοριών που παρέχονται για μια βιβλιογραφική ανάλυση η οποία στη συνέχεια συσχετίζει τα microRNAs με ασθένειες.

Σε αυτή τη διπλωματική εργασία, ανακτούμε αυτόματα τις συσχετίσεις των microRNAs με τις ασθένειες από τη βιβλιογραφία. Χρησιμοποιώντας Perl scripts, όλα τα αποσπάσματα που σχετίζονται με microRNAs ανακτώνται από τη PubMed, την ελεύθερη διαδικτυακή λογοτεχνία υπηρεσία αναζήτησης του Εθνικού Κέντρου Βιοτεχνολογικών Πληροφοριών των Ηνωμένων Πολιτειών. Τα αποσπάσματα σχετίζονται με microRNAs με βάση την παρουσία του ονόματος των microRNA στον τίτλο ή την περίληψη του δημοσιεύματος.

Τα ανακτημένα δεδομένα είναι συσχετισμένα με Medical Subject Headings (MeSH), το ελεγχόμενο λεξιλόγιο προκαθορισμένων όρων της Εθνικής Ιατρικής

Βιβλιοθήκης των Ηνωμένων Πολιτειών, και οι πληροφορίες που εξάγουμε περιέχουν τα ονόματα των microRNAs, όρους MeSH, και το PubMed ID βιβλιογραφίας. Αυτή η πληροφορία επεξεργάστηκε κατάλληλα από την ομάδα του DAINA Lab (DNA intelligent Analysis) του Ερευνητικού Κέντρου Βιοϊατρικών Επιστημών "Αλέξανδρος Φλέμινγκ", έτσι ώστε να χρησιμοποιηθεί στην τελευταία έκδοση του DIANA-microT Web server, το οποίο είναι ένα πρόγραμμα για την πρόβλεψη στόχου των microRNA που βασίζεται σε Τεχνητά Νευρωνικά Δίκτυα. Μπορεί να χρησιμοποιηθεί για την αναζήτηση γονιδίων-στόχων των αλληλουχιών miRNA προκειμένου να βοηθήσει τους βιολόγους και τους ερευνητές στην ανάκτηση σημαντικών στοιχείων με αποδοτικό τρόπο.

Ευχαριστίες

Με την ολοκλήρωση της διπλωματικής μου εργασίας, θα ήθελα να ευχαριστήσω τους επιβλέποντες καθηγητές μου κ. Μούντανο Ιωάννη και κα. Χατζηγεωργίου Άρτεμις καθώς και τον κ. Αλεξίου Παναγιώτη για την εμπιστοσύνη που έδειξαν στο πρόσωπό μου, την πολύτιμη βοήθεια στην υλοποίηση της διπλωματικής εργασίας, την καθοδήγηση και υποστήριξή τους όπως επίσης και για την ευκαιρία που μου έδωσαν να ασχοληθώ με ένα τόσο ενδιαφέρον αντικείμενο.

Ευχαριστώ θερμά τους γονείς μου Κώστα και Γεσθημανή και τις αδερφές μου , για την ηθική και υλική υποστήριξή τους όλα αυτά τα χρόνια, για τη στήριξη των επιλογών μου και για όλα όσα μου δίδαξαν και θα μου διδάξουν στη ζωή μου.

Τέλος, ένα μεγάλο ευχαριστώ στις φίλες μου και στον κ. Θεόφιλο Κορωνά για την αμέριστη συμπαράσταση και στήριξη που μου προσφέρουν, δίπλα μου, σε κάθε μου βήμα.

Table of contents

Appendix

Appendix

1

Introduction

1.1 Biological Research of microRNAs

This thesis is directly related to the science of Biology. The scope of Biology is the study and the extraction of conclusions concerning the structure, function, growth, origin and evolution of living organisms. Biology attempts to answer all the minor and big questions about the organisms' life so as to resolve problems that surround it. This science flourished since 1953, where Francis Harry Compton Crick and James D. Watson announced the revolutionary model of the structure of DNA. The knowledge gained through this discovery, led gradually to the development of valuable tools to fight many of the diseases today.

In 1993, Victor Ambros, Rosalind Lee and Rhonda Feinbaum discovered the microRNA biomolecules. This discovery was particularly important because it contradicted the "Central Dogma of Biology" on basis of which, biologists believed until then that protein synthesis takes place. Those biomolecules bind to the "messenger RNA" (mRNA) resulting from specific genes and prevent the production of the corresponding protein. The binding sites of microRNAs in mRNAs are called targets.

Through the mechanism described above, microRNAs function regulatory on the organism's cells and controlling protein synthesis. Many types of diseases are the result of the non-production of a particular protein. As such, the acquisition of knowledge about target prediction of microRNAs or not in a gene is particularly important in understanding various diseases. Biological experiments can reveal this information, but they cost much and it takes time for completion. Bioinformatics, in response to this problem uses computational methods for microRNA target prediction.

The team of the DAINA Lab (DNA intelligent Analysis) of Biomedical Sciences Research Center "Alexander Fleming" has developed a prediction algorithm so as to identify miRNA target genes whose results are presented by the latest version of DIANA-microT Web server (Diana microT v4.0).

1.2 The subject of this thesis

This relatively new academic field of Informatics applications in Biology and Biological research, Bioinformatics is a field full of promise for important inventions and discoveries through research in its major areas, which include the analysis of the gene expression as well as the analysis of its regulation.

The study of gene expression and its control at the post-transcriptional level from small non-coding RNA molecules called microRNAs helps us to understand their function so that we can extract information associating microRNAs to many various diseases.

The need for storing, retrieving, processing and analyzing these available data generated by the scientific community is becoming more urgent because the volume of information continues to grow exponentially.

The subject of this thesis is to extract knowledge about the associations of microRNAs with diseases from biological literature. Specifically we are developing the automatic search and retrieval of articles that are associated with names of microRNAs using for this procedure the appropriate MeSH terms, from PubMed. As well as the automatic extraction of the respective microRNAs, MeSH terms and the literature PubMed IDs, witch determines the associations of microRNAs with diseases.

The implementation of this thesis was developed in collaboration with the team of the DAINA Lab (DNA intelligent Analysis) of Biomedical Sciences Research Center "Alexander Fleming", so as to be used in the updated version of DIANA-microT Web server, a program for microRNA target prediction. It may be used to search for target genes of microRNA sequences in order to assist biologists and researchers retrieving significant information.

1.3 Text Organization

The content of this thesis is organized into the following chapters:

In Chapter 2 we present the theoretical background on which this thesis was based and some basic concepts used. Because of the direct relation of this thesis with Biology, some basic concepts of this field were presented. Moreover, some of the associations of microRNAs with diseases are mentioned.

In Chapter 3 we present all the basic knowledge needed for the implementation of this thesis. We introduce some concepts of computer science such as pattern matching and we briefly introduce some of the key elements used for the implementation of this thesis. Afterwards we present the implementation of this thesis, the Perl scripts used for the automatic search, retrieval and extraction of the data.

In Chapter 4 we present some of the basic functions and features of the DIANA-microT Web server, those in which are relevant with the implementation of this thesis.

In Chapter 5 is the conclusion of this thesis followed by the References.

.

2

Theoretical Background

This thesis is part of Bioinformatics, the scientific discipline that resulted from the collaboration of Molecular Biology and Computer Science. This thesis attempts to develop software so as to provide useful information in the research community of biologists. Therefore, it is appropriate to briefly mention some issues related to the field of Biology so that it is easier to understand the subject and scope of this thesis.

2.1 Introduction to Bioinformatics

Recently there has been a massive explosion in the amount of biological information generated by the scientific community. This enormous increase in available genomic information and biological data resulting from the major developments and advances in both molecular biology and in genomics technology has led to an absolute need to use databases. In these databases, the biological data are stored and organized in order to be analyzed by specialized tools. This wealth of biological information has given rise to a new field, bioinformatics, which combines methods and techniques of biology and computer science.

Bioinformatics is the science that provides tools and methods to meet the need of exploiting computing power and knowledge extraction from biological data. A definition for bioinformatics according to *What is Bioinformatics? A Proposed Definition and Overview of the field*, by N. M. Luscombe, D. Greenbaum and M. Gerstein : "*(Molecular)* **bio** – **informatics:** bioinformatics is conceptualising biology in terms of molecules (in the sense of Physical chemistry) and applying **"informatics techniques"** (derived from disciplines such as applied maths, computer science and statistics) to **understand** and **organize** the **information** associated with these molecules, on a **large scale.** In short, bioinformatics is a management information system for molecular biology and has many **practical applications.**"

Paulien Hogeweg, a Dutch theoretical biologist and Ben Hesper another biologist coined the term Bioinformatics in 1978, but the history of bioinformatics started long before that. Its basic use is in the area of genomics which includes large scale DNA sequencing. Bioinformatics deals with many aspects of computer science and information technology such as algorithms, databases and information systems, web technologies, information and computation theory, data mining, image processing, modeling and simulation, signal processing, discrete mathematics and statistics, in order to generate new knowledge of biology and medicine, and also to improve and discover new computational models and methods [1].

Bioinformatics manages to uncover the wealth of biological information hidden in the mass of data and obtain a clearer insight into the fundamental biology of organisms. The rationale for applying computational approaches to facilitate the understanding of various biological processes includes the development and application of computationally intensive techniques.

The production process of biological data, the resulting storage interpretation and analysis are entirely dependent computer tasks. Some of the distinguished software technologies used in this field are Java, XML, Perl, C, C++, Python, R, SQL and MatLab which incorporate computers into the research process.

The discovery of new biological concepts and insights as well as the development of a global perspective from which consolidative principles in biology can be distinguished is the ultimate goal of bioinformatics. One of the greatest challenges that molecular biology community is facing today is to interpret the wealth of data that has been generated by the genome sequencing projects.

The field of bioinformatics has evolved so as to involve the analysis and interpretation of sequence information. Thus resulting in Computational Biology. Two Important sub-disciplines within bioinformatics involving computational biology would include:

- the development and implementation of tools that enable efficient access and management of different types of information
- the development of new algorithms and statistics with which to assess relationships among members of large data sets [2].

The integration of information learned about these computational processes should allow us to achieve the long term goal of the complete understanding of the biology of organisms.

The science of Bioinformatics now refers to the creation and development of databases, algorithms computational and statistical techniques in order to find solutions in formal and practical problems arising from the analysis and manipulation of biological data. It is also essential to the use of genomic information in understanding human diseases and in the identification of new molecular targets for drug discovery. Many and varied areas could benefit from this new knowledge generated by Bioinformatics, such as human health, agriculture, the environment, energy and biotechnology.

2.1.1 Major Research Areas and Applications in Bioinformatics

Bioinformatics is a modern field of research and development and is being used in many areas of the scientific world, allowing extensive and important work to be done in the medical community as well.

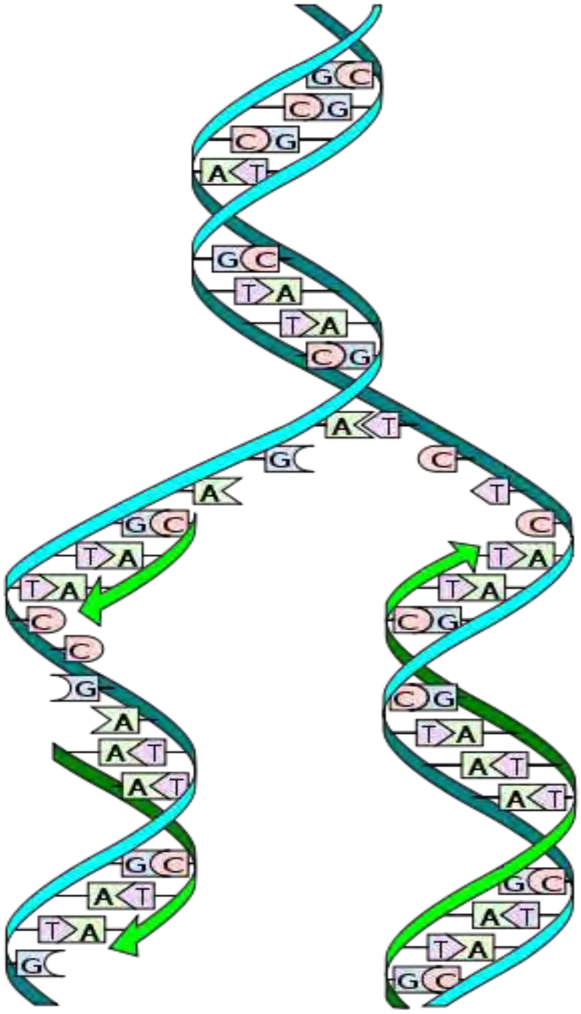
Researchers are making significant efforts in the area. Some of the traditional avenues include sequence analysis, genome annotation, literature analysis, analysis of gene expression, analysis of regulation, analysis of protein expression drug design, drug discovery and modeling of evolution. [1]

Bioinformatics has many different applications, the options are limitless. The modern day world has many beneficial utilities to gain by the science of Bioinformatics. These include the following: molecular medicine, preventative medicine, gene therapy, drug development, microbial genome applications, climate change, alternative energy sources, antibiotic resistance, bio-weapon creation, improve nutritional quality and development of drought resistance varieties [3].

2.2 Central Dogma of Molecular Biology

The "Central Dogma" of molecular biology refers to the flow of genetic information in biological systems. It was first proposed by Francis Crick in 1958 and re-stated in a paper published in 1970. The central dogma of molecular biology deals with all possible directions of information flow between DNA, RNA, and protein [4].

DNA (Deoxyribonucleic acid) is a double-stranded nucleic acid that contains the complete genetic information and defines the structure and function of all cellular forms of life and most viruses. Proteins, the molecules that carry out the specialized functions of the cell, are formed using the genetic code of the DNA. In other words specific regions of the genome called genes encode the information for the synthesis of a protein. There are many proteins that are absolutely essential for a wide range of cellular processes that allow organisms to function [5][6].

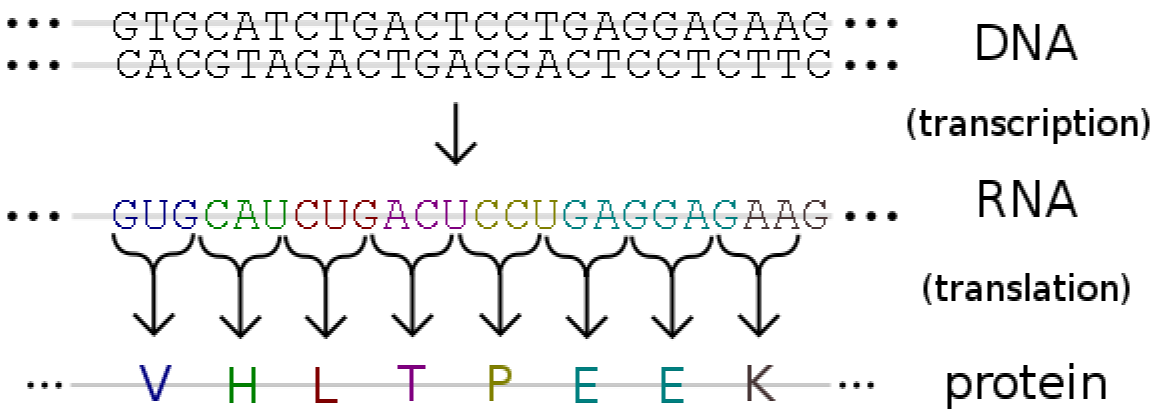


2.1 DNA Replication

In early use of the term, the researchers knew the following three possible ways for the flow of genetic information:

- **Replication of DNA** : a mechanism by which an exact copy of the genome is made. This process preserves and transfers genetic information from cell to cell.
- **Transcription** : describes the process where the DNA codes for the production of messenger RNA (mRNA). Before protein synthesis can occur, genes must be converted into RNA.
- **Translation** : where a messenger RNA molecule carries coded information to ribosomes and these, in turn, "decode" this information and use it for protein synthesis. As the protein is formed, the RNA sequence is translated into a sequence of amino acid.

Crick concluded that once information was transferred from nucleic acid (DNA or RNA) to protein it could not flow back to nucleic acids. In other words, once information has passed into protein, it cannot get out again.


••• GTGCATCTGACTCCTGAGGAGAAG •••
••• CACGTAGACTGAGGACTCCTCTTC •••
DNA
(transcription)
••• GUGCAUCUGACUCCUGAGGAGAAG •••
RNA
(translation)
... V H L T P E E K ...
protein

2.2 Central dogma of molecular biology as expressed by Francis Crick in 1958.

Nowadays, researchers in biology know that there are additional ways of genetic information flow. In particular, it has been an identification of some viruses whose genetic material is RNA and they are using a special enzyme, called reverse transcriptase, in order to use this RNA to synthesize DNA. Researchers also found that in some viruses RNA replication can occur [7].

Moreover, with the discovery of the key roles of small RNA molecules, called microRNAs, in gene expression that perspective changed drastically. Scientists estimate that microRNAs have the ability to regulate the expression of approximately one third of human genes. MicroRNAs are segments of RNA and they are not translated into proteins [9].

We will describe microRNAs, as well as some key aspects concerning their function in biological processes in the next paragraph of this chapter.

2.3 MicroRNAs

In 1993 researchers involved in the Human Genome Project mapped the first chromosome by isolating a gene. The human genome is estimated to contain 20,000-25,000 protein coding genes but less than 2% of the genome codes for proteins, while the rest does not produce proteins and it is consisted of non-coding RNA genes, regulatory sequences, introns and noncoding DNA (dismissively referred to as “junk DNA”) [10][11]. Noncoding RNAs (ncRNAs) are RNA molecules, a whole new class of genes which, unlike most genes, do not encode for protein. MicroRNAs consist such an example, of non–protein-coding RNA molecules [12][13].

Discovered in 1993 by Victor Ambros, Rosalind Lee and Rhonda Feinbaum, microRNAs (abbreviated miRNAs) are single-stranded RNA molecules found in all eukaryotic cells which are typically 21-23 nucleotides long and regulate gene expression, that have already been transcribed from DNA to RNA, by binding to the messenger RNA (mRNA) of protein coding genes. Even though microRNAs were characterized in the early 1990s it was in the early 2000s that were recognized as a distinct class of post-transcriptional regulators whose their main function is to downregulate gene expression[14][15].

In mammals, microRNAs are predicted to control the activity of approximately 30% of all protein-coding genes, and have been shown to participate in the regulation of almost every cellular process. They are also vital components in the progression or treatment of different diseases including cancer, cardiovascular disease, diabetes, mental disorders and viral infection. There have been several research results reported recently that present some of the important functional roles of miRNA [16].

2.3.1 Nomenclature of MicroRNAs

In this paragraph, we present the specific conventions which are used for naming microRNAs that have been experimentally confirmed before publication of their discovery. The regular expression used, in order to extract microRNAs names from biomedical literature, was constructed under this standard nomenclature system. This process will be presented in the next chapter.

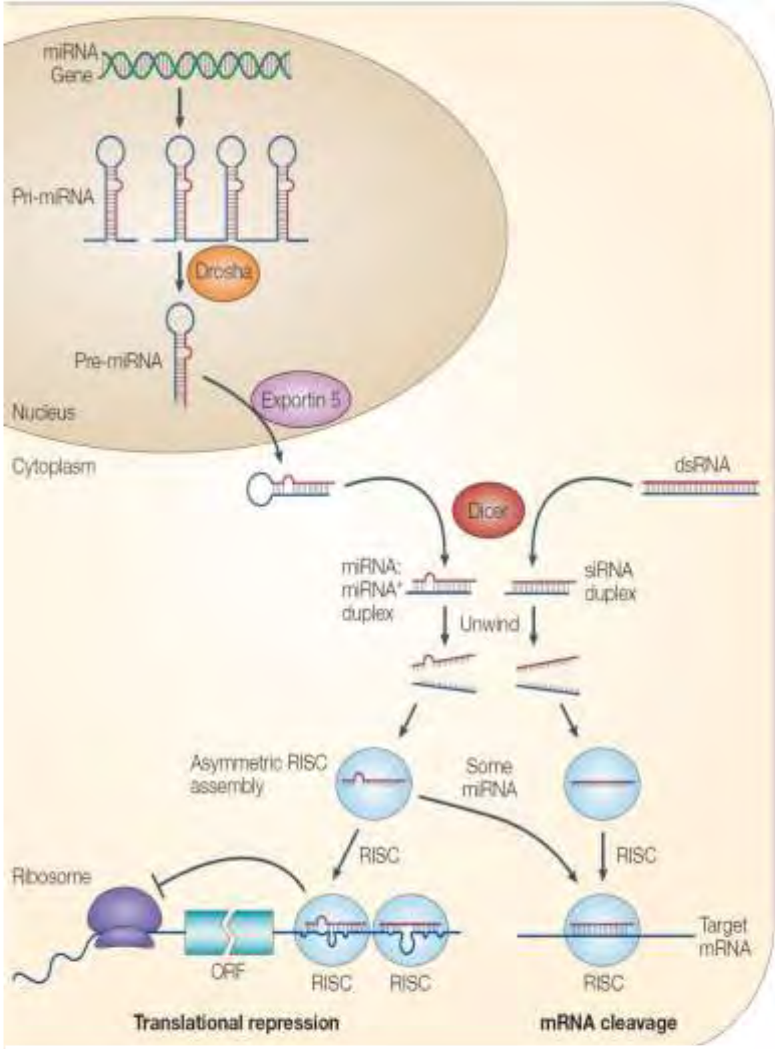
MicroRNAs are named using the three-letter prefix “mir” followed by a unique identifying number which is often designating order of naming since the numbering of miRNA genes is sequential. For example, mir-3 was named and likely discovered after mir-2.

The mature form of microRNAs is also designated using the same prefix, with capitalization. For example the mature miRNA is designated miR-122, whilst mir-122 refers to the pre-miRNA.

Lettered suffixes denote species of origin. For example hsa-miR-123 is a human (Homo sapiens) miRNA and oar-miR-123 is a sheep (Ovis aries) miRNA, with identical microRNAs having the same number, regardless of organism.

MicroRNAs with identical or nearly identical sequences within a species can also have the same number, with their genes distinguished by letter and/or numeral suffixes, according to the convention of the organism. For example mir-13a and mir-13b are slightly different in sequence bar one or two nucleotides, whereas those of hsa-mir-194-1 and hsa-mir-194-2 are identical. Also, these pre-miRNAs lead to an identical mature miRNA (hsa-miR-194) but are located in different places in the genome.

With a -3p or -5p suffix we denote two mature microRNAs originated from opposite arms of the same pre-miRNA. [16]



2.3 Function of MicroRNAs

2.3.2 MicroRNAs and Gene expression

In order to make a protein, a gene codes for a specific messenger RNA molecule. Each mRNA molecule contains a design for making a protein, the genetic information from DNA. A microRNA attaches to a piece of mRNA in a non-coding part at one end of the molecule. This acts so as to prevent translation of the mRNA into a protein.

These small molecules bind to complementary sequences in the 3'-UTR (untranslated region) of their target gene mRNA and interfere with translation usually causing gene silencing. In this way they repress protein production. MicroRNAs regulate their target genes through two main mechanisms: target mRNA cleavage and translational repression however the exact mechanism(s) are still unclear.

The regulation of gene expression through microRNAs and through other organisms is extremely important. Deregulation of gene expression can actually cause many diseases. One of those diseases that either cause or are caused by this process is cancer.

Biologists and researchers hope to identify how the regulation of gene expression through microRNAs fails in human diseases, by improving the understanding about how microRNAs control protein synthesis [17][18][19].

2.3.3 MicroRNA Target Prediction

The binding sites at which the molecules of microRNA are connected to messengers RNA (mRNA) are called targets. The knowledge of these targets is important for researchers - biologists, because in this way they can learn about the function of microRNA molecules.

There are two different techniques to identify the targets. The first is the experimental method. Unfortunately, this method is expensive and time consuming so it is not that easy to carry out. The second technique refers to the use of sophisticated algorithms from the field of informatics, which may predict the existence of microRNA targets in a mRNA sequence through a series of computational procedures. This process is called target prediction and is a collaboration point-which is rapidly evolving over the last years- between biology and computer science.

An algorithm for the prediction of such targets has been developed by the group of the Research Center of Biomedical Sciences "Alexander Fleming "[20].

2.4 MicroRNA and Diseases

Since the discovery of the strong impact of microRNAs on biological processes, it has been hypothesized that microRNAs may have a pathogenic role in human diseases. An association between microRNAs and human disease can be extracted from the fact that two genes that are known to be causative agents in unrelated congenital diseases encode essential miRNA components.[21]

2.4.1 MicroRNAs and Cancer

A large amount of evidence has already shown that miRNA expression is implicated in most forms of human cancer. A number of microRNA genes are located in genomic regions that are frequently amplified, deleted, or rearranged in cancer, providing evidence of some kind of a role for microRNAs in cancer pathogenesis.

Early clues linking microRNAs to cancer came from observations in chronic lymphocytic leukemia (CLL). Initial observations that miRNA genes were located on genomic instability and fragile sites lead to further analysis that showed deregulated miRNA expression profiles in different tumors.

MicroRNAs can target more than one target – gene. Therefore, a single miRNA can act as both an oncogene, if it targets a gene that prevents cell proliferation, and as a tumor suppressor, if it targets a gene that promotes the growth of a tumor. [22][23][24]

2.4.2. MicroRNAs and Heart Disease

The role of microRNAs in cardiac development has been well described in several studies. In response to pathological states, such as hypertension, ischemic myocardial injury, and valvular disease, the myocardium reacts with changes in the gene expression profile. This adaptation usually results in cardiac remodeling, which is characterized by severe structural alterations of myocardial tissue and reshaping of left ventricle geometry and performance.

Recently, several reports have revealed important roles of microRNAs in cardiac hypertrophic growth and heart failure. During the adaptive response of the heart to stress stimuli, microarray analyses have shown upregulated, downregulated, or unchanged miRNA expression, when compared with normal heart.

Although some of the microRNAs regulated during cardiovascular disease are not highly expressed outside the heart, a large number of such microRNAs are broadly expressed throughout the body. Inhibition of microRNAs that are specifically expressed, or highly enriched, in the cardiovascular system may circumvent side effects resulting from miRNA activity in additional organs. [25][26]

3

Implementation

This chapter is a description of the subject of this thesis, the necessary concepts and elements for the implementation and the programming part (code) of this thesis.

3.1 Techniques and Features in Computer Science

Next we will refer briefly to some basic concepts and techniques of computer science in order to make clear how this thesis was implemented. We will also refer to the database used (PubMed), the medical subject heading terms (MESH) and the Perl programming language which was used.

3.1.1 Pattern Matching

Manipulating text is one of the most basic and important computing tasks. Pattern matching is one of the key techniques used for the implementation of this thesis. The Perl programming language we used is built to support pattern matching.

In computer science, String Matching or Pattern Matching is the process in which we locate and identify all possible occurrences of a given string, word, pattern to another text (usually longer). The simplest case is finding a given word in a sentence or a text. It is used for accessing and acquiring information, and without doubt, nowadays many computers solve this problem as a simple and common operation in an application system.

According to FOLDOC, a computing dictionary supported by Imperial College Department of Computing, London :

“1. A function is defined to take arguments of a particular type, form or value. When applying the function to its actual arguments it is necessary to match the type, form or value of the actual arguments against the formal arguments in some definition. For example, the function

```
length []     = 0
length (x:xs) = 1 + length xs
```

uses pattern matching in its argument to distinguish a null list from a non-null one.

2. Descriptive of a type of language or utility such as AWK or Perl which is suited to searching for strings or patterns in input data, usually using some kind of regular expression.”

3.1.2 Regular Expressions

A regular expression is a model that describes a set of strings. It is used to describe a formal language: strings that match this model belong to the language, it describes if there isn’t a match strings don’t belong in that language. Regular expressions allow the definition of complex strings and are used to find patterns in texts. According to FOLDOC, a computing dictionary supported by Imperial College Department of Computing, London :

“Any description of a pattern composed from combinations of symbols and the three operators:

- Concatenation - pattern A concatenated with B matches a match for A followed by a match for B.
- Or - pattern A-or-B matches either a match for A or a match for B.
- Closure - zero or more matches for a pattern.”

Mathematics and in particular Keene Algebra, is the basis of regular expressions. They were named and introduced by Keene in 1950s. In computer science, we use a regular expression as a special string to describe a search pattern. [27]

We are using regular expressions because they are a short, distinct and easy way to express any combination of characters in order to match amounts of text. The article : Regular Expression in Wikipedia refers : “ A regular expression is written in a formal language that can be interpreted by a regular expression processor, which is a program that either serves as a parser generator or examines text and identifies parts that match the provided specification.”

Among the many programming languages which are using regular expressions is Perl, which includes tools for pattern matching or pattern recognition based on a variety of standards into Perl’s syntax.

Below in this chapter, we present Perl’s regular expression used according to the needs of the task.

3.2. Medical Subject Headings (MESH)

Medical Subject Headings (MeSH) is a controlled vocabulary of pre-defined terms used for indexing PubMed citations. Created and updated by the United States National Library of Medicine (NLM), it is used to search within databases such as MEDLINE/PubMed.

It is a unique feature of health and medical literature designed in a hierarchy of terms and phrases which allows us to retrieve information that may use different terms to describe the same concept. MeSH vocabulary is also designed as a thesaurus in order to facilitate searching. [28]

Use in Medline/PubMed

Usually between 10 and 15 MeSH terms are assigned to every journal article In MEDLINE/PubMed in order to describe its content. The database maps the search term(s) to the preferred terminology, the MeSH heading, used in the database to describe this topic. [29]

3.3 PubMed

PubMed is a free literature search service created and first became available in January 1996 by the United States National Center for Biotechnology Information (NCBI), an United States National Library of Medicine‘s (NLM) division. The NLM is located at the campus of the United States National Institute of Health (NIH).

PubMed provides access to MEDLINE database, which is the primary information source of PubMed, and is covering the fields of medicine, pediatrics, dentistry, veterinary medicine and pre-clinical research such as molecular biology (including also studies from NIH). [32]

It is designed so as to provide access to information of medical sciences with references and summaries from current bio-medical journals and to use hyper-links in order to access full text articles as published online by publishers. [31]

PubMed is part of a wider network search system, Entrez, which is a tool for cross-searching databases of NCBI ,and also the basic system to search and retrieve medical data in all NCBI's databases including PubMed.[33]

Basic Features of PubMed

- *Automatic Term Maping* : PubMed is using an Automatic Term Mapping (ATM) feature in order to satisfy the query (queries) submitted to the database by using keywords and in particular in order to facilitate phrase searching.
- *Mapping to MeSH terms* : PubMed has the ability to automatically map each article with MeSH terms and subheadings.
- *PubMed identifier* : A PMID (PubMed identifier or PubMed unique identifier) is a unique number assigned to each PubMed's article citation of biomedical or life sciences journal.[30]

While PubMed is an efficient and effective information search service of medical and life sciences, it has become increasingly difficult to quickly determine the information required for acquisition of expedient knowledge, mainly due to the massive amounts of biomedical literature which is more and more growing. The Entrez Programming Utilities come to answer this problem by making possible the development of useful and efficient applications from the research community.

3.4 Entrez Programming Utilities (E-utilities)

The implementation of automatic search and retrieval of the required information for associating microRNAs with diseases is using a set of web services, provided by the United States National Center for Biotechnology Information (NCBI), in order to form a query to PubMed, covering a wide range of biomedical data, such as protein and nucleotide sequences, genomes, three-dimensional molecular structures and biomedical literature. For accessing PubMed is being used a set of eight server-side programs (Entrez Programming Utilities, or E-utilities). E-utilities are using a fixed URL syntax, which translates a standard set of input parameters to the required values, which allow search and retrieval of data.

For interaction with E-utilities we developed scripts in Perl. In particular, the implementation of automatic search and retrieval communicates with the following E-utilities:

- **ESearch** : The interface responds to a text query with a list of UIDs, matching the required parameters of the submitted search query. UID identifiers (PMID for PubMed and UI for MEDLINE) are used to uniquely identify articles contained in the databases.

- **EFetch** : Returns the data corresponding to a list of UIDs.

The automatic search and retrieval combines both of these E-utilities and implements a data pipeline to retrieve records that meet the criteria we set. In particular, it produces an ESearch → EFetch pipeline to recover the titles and abstracts that match the list. Titles, abstracts, and PMIDs are encoded in XML format. [34]

3.5 Perl (Practical Extraction and Report Language)

In order to implement the automatic search and retrieval of all abstracts associated with microRNAs as well as the extraction of the respective microRNAs, MeSH terms and PMIDs, the Perl (Practical Extraction and Report Language) programming language has used. It is a high-level, general-purpose, programming language developed by Larry Wall in 1987.

This dynamic programming language has become widely popular amongst programmers due to its text processing capabilities as well as its parsing abilities. One of the main reasons behind this great ability in manipulating data is Perl's regular expressions. The patterns described by Perl's regular expressions are used to search, extract and replace desired parts of strings in an efficient and flexible way.

Perl is used in applications from many different fields and sciences such as finance, network programming, and bioinformatics due to its features including performance power and flexibility. [35]

3.6 Implementation of Automatic Search, Retrieval and Extraction of Information

In order to implement the automatic search and retrieval of all abstracts associated with microRNAs from PubMed, the following steps were used:

Initially, we searched the appropriate MeSH terms using the MeSH database in order to accomplish the required search in PubMed by using the combination of those MeSH terms. This combination is actually the input file, which along with some required parameters is the query, which will be searched in PubMed using the E-utilities.

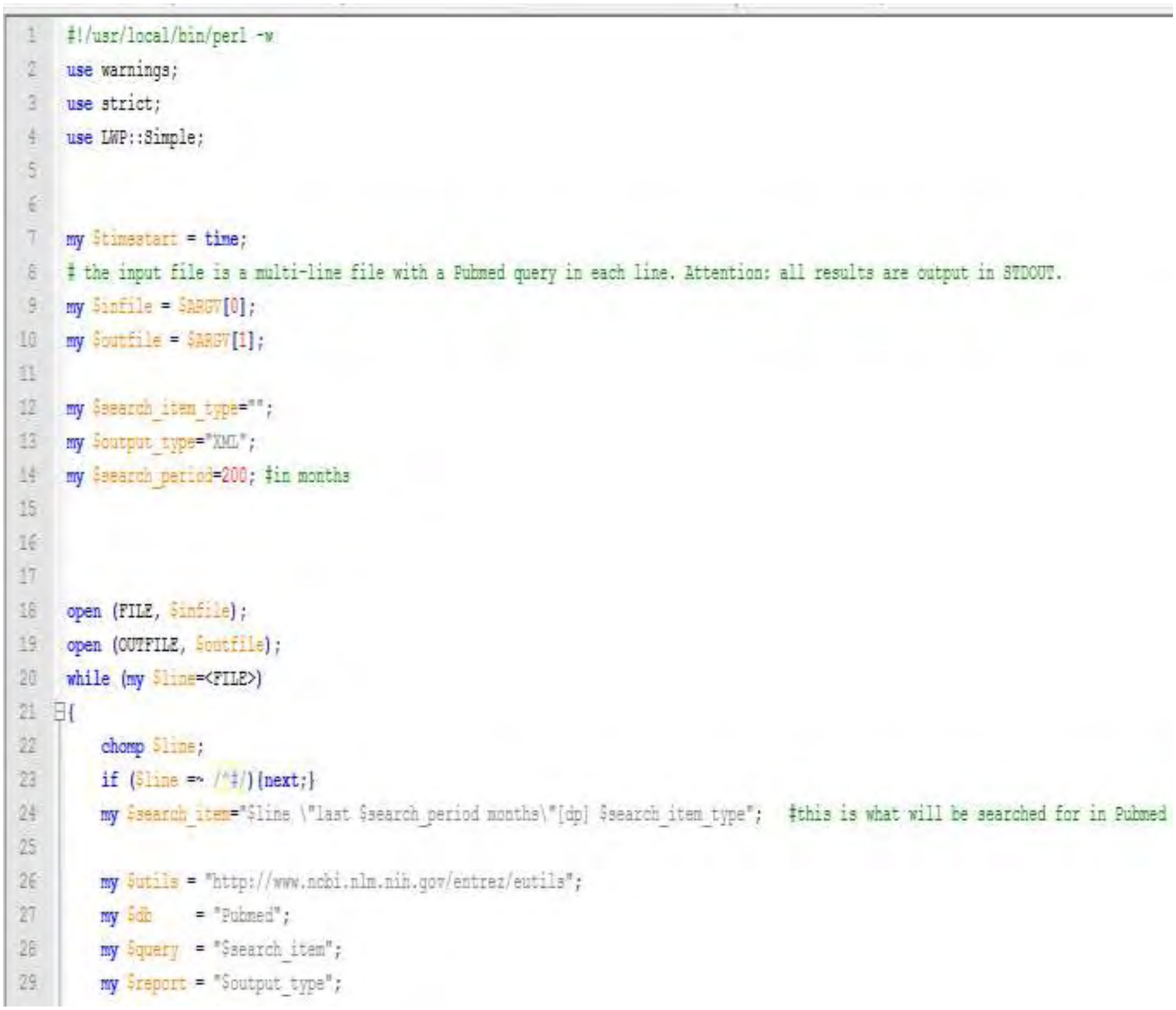
Afterwards, the automatic search and retrieval of all abstracts associated with a microRNA, based on the presence of the name of the miRNA in the title or abstract of the publication, takes place. As mentioned above, the input file with the additional parameters required is the query of the special Esearch call, as described above. Then, the automatic retrieval of corresponding abstracts will follow through Efetch call, which is also described above, using appropriate parameters for retrieving the abstracts in the proper XML format.

The next step is to extract the information needed, that is to say, the corresponding microRNA names, MESH terms, and literature PubMed ID of those abstracts in XML format. The input file given is the output file of the previous step. This process is using pattern matching through Perl's appropriate regular expression which was designed according to MicroRNAs' nomenclature rules.

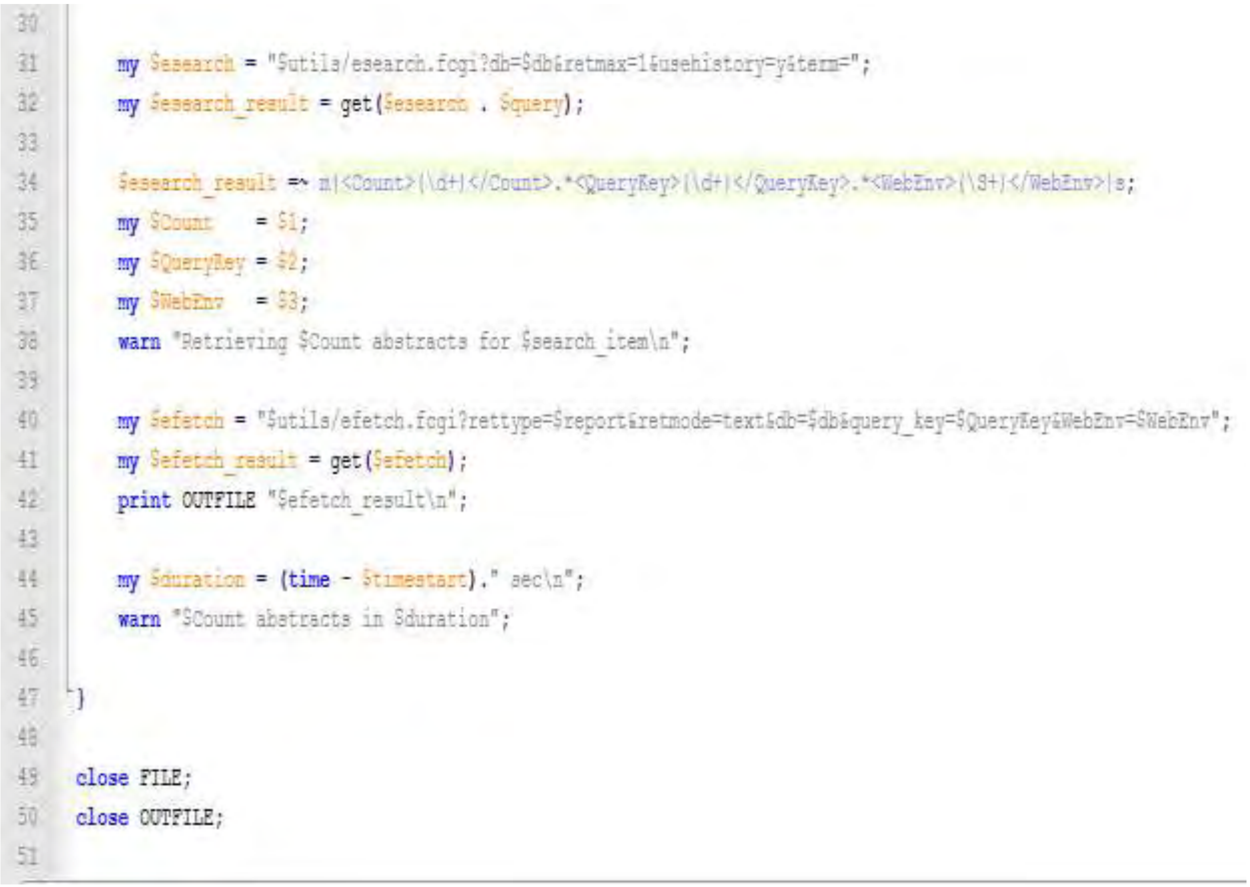
By performing the above steps, we are able to retrieve the information required, all disease associated MeSH terms for a microRNA. This information was properly processed by the team of the DAINA lab so as to be used to the DIANA-microT Web Server update, where some of its features concerning the implementation of this thesis, will be briefly presented in the next chapter.

3.6.1 Automatic Search and Retrieval from PubMed

The Perl script developed for the automatic search and retrieval of all abstracts from PubMed is presented and described below. By executing PubMed_Search_and_Download_xml.pl and through its E-utilities calls we are accessing, searching and retrieving PubMed's data.



```perl
#!/usr/local/bin/perl -w
use warnings;
use strict;
use LWP::Simple;


my $timestart = time;
# the input file is a multi-line file with a Pubmed query in each line. Attention: all results are output in STDOUT.
my $infile = $ARGV[0];
my $outfile = $ARGV[1];

my $search_item_type="";
my $output_type="XML";
my $search_period=200; #in months



open (FILE, $infile);
open (OUTFILE, $outfile);
while (my $line=<FILE>)
{
    chomp $line;
    if ($line =~ /^#/){next;}
    my $search_item="$line \"last $search_period months\"[dp] $search_item_type";   #this is what will be searched for in Pubmed

    my $utils = "http://www.ncbi.nlm.nih.gov/entrez/eutils";
    my $db     = "Pubmed";
    my $query  = "$search_item";
    my $report = "$output_type";
```



```
30
31      my $esearch = "$utils/esearch.fcgi?db=$db&retmax=1&usehistory=y&term=";
32      my $esearch_result = get($esearch . $query);
33
34      $esearch_result =~ m|<Count>(\d+)</Count>.*<QueryKey>(\d+)</QueryKey>.*<WebEnv>(\S+)</WebEnv>|s;
35      my $Count    = $1;
36      my $QueryKey = $2;
37      my $WebEnv   = $3;
38      warn "Retrieving $Count abstracts for $search_item\n";
39
40      my $efetch = "$utils/efetch.fcgi?rettype=$report&retmode=text&db=$db&query_key=$QueryKey&WebEnv=$WebEnv";
41      my $efetch_result = get($efetch);
42      print OUTFILE "$efetch_result\n";
43
44      my $duration = (time - $timestart)." sec\n";
45      warn "$Count abstracts in $duration";
46
47  }
48
49  close FILE;
50  close OUTFILE;
51
```

Line 4: Initially along with the default script header we use the LWP::Simple module in order to fetch the required data as a string from PubMed. It is required, since we are using the ‘get’ function, as a defined library.

Line 7: the $timestart variable is initialized with the value returned from the time function, the number of seconds the system’s epoch, is 00:00:00 UTC, January 1, 1970. We use the $timestart variable as a timestamp so as to calculate the time needed to retrieve the data.

Line 9: the $infile variable refers to the input file, given from the command line, and contains the PubMed query with the Mesh terms. The input file is the infile.txt and contains the query: #"micrornas"[MeSH Terms] AND "neoplasms"[MeSH Terms] "micrornas"[MeSH Terms].

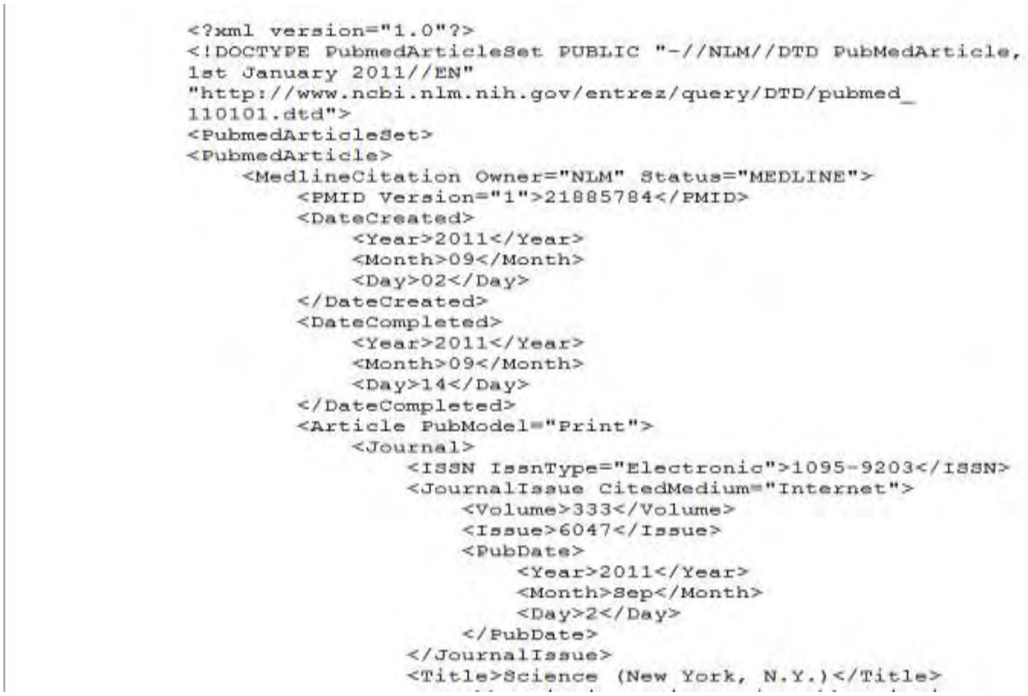
Line 10: the $outfile variable refers to the output file, given also from the command line. This is the file which contains the retrieved abstracts, along with the corresponding MeSH terms and PMID.

Line 12-14: The retrieved data will be the abstracts along with their metadata of the last 200 months in XML format.

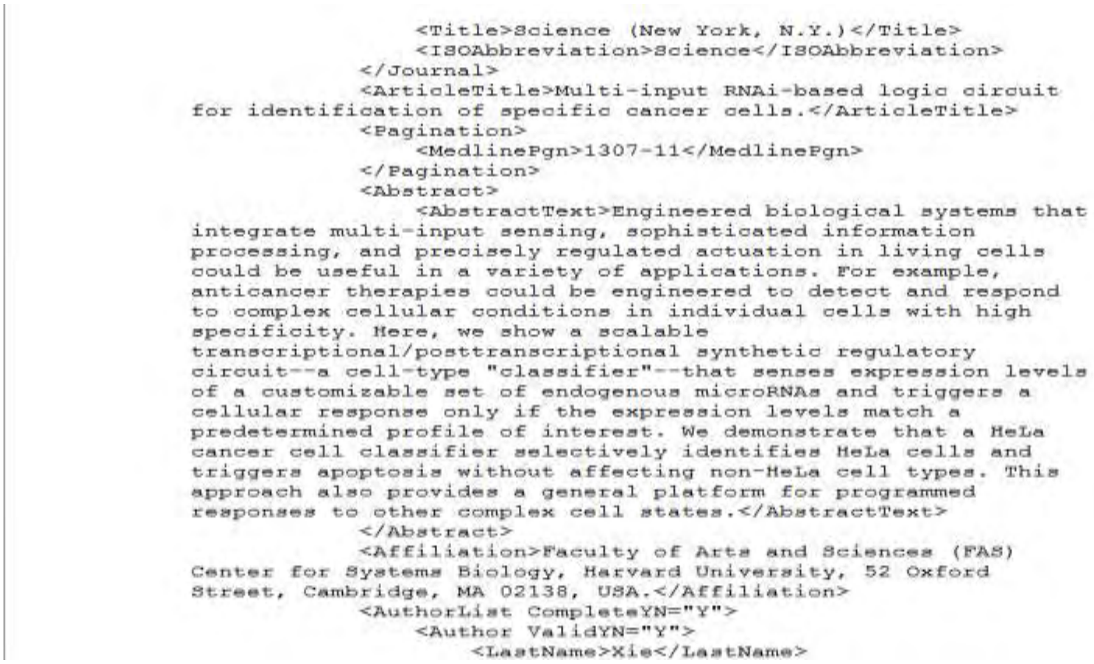
Lines 18-50: we open the input and output file. The input file opens in order for the MeSH terms to be 'read' and to set the $search_item variable with the required parameters. The $search_item variable is the query which will be searched in PubMed. It contains the MeSH terms and the searching period that will be searched. The $utils variable contains the route, the URL, for the E-utilities as we mentioned above. Then searching in PubMed takes place according to the above parameters. The $esearch variable contains the path and parameters for the ESearch call and the $esearch_result variable contains the result of the ESearch call, which are displayed and parsed into the $Count, $QueryKey, and $WebEnv variables for later use. Next, the $efetch variable contains the path and parameters for the EFetch call, among them the desire XML format, and the $efetch_result variable contains the result of the EFetch call, all the abstracts from the last 200 months with a microRNA name in the title or the abstract of PubMed's publications. Afterwards we print the retrieved data to the output file and we display the duration time of the retrieval. Since we are done with the automatic search and retrieval we close both of the files.

To run the script:

perl PubMed_Search_and_Download_xml.pl infile.txt > abstracts_xml.txt. The output file, abstracts_xml.txt, will contain abstracts in the form presented below :



```
<?xml version="1.0"?>
<!DOCTYPE PubmedArticleSet PUBLIC "-//NLM//DTD PubMedArticle,
1st January 2011//EN"
"http://www.ncbi.nlm.nih.gov/entrez/query/DTD/pubmed_
110101.dtd">
<PubmedArticleSet>
<PubmedArticle>
    <MedlineCitation Owner="NLM" Status="MEDLINE">
        <PMID Version="1">21885784</PMID>
        <DateCreated>
            <Year>2011</Year>
            <Month>09</Month>
            <Day>02</Day>
        </DateCreated>
        <DateCompleted>
            <Year>2011</Year>
            <Month>09</Month>
            <Day>14</Day>
        </DateCompleted>
        <Article PubModel="Print">
            <Journal>
                <ISSN IssnType="Electronic">1095-9203</ISSN>
                <JournalIssue CitedMedium="Internet">
                    <Volume>333</Volume>
                    <Issue>6047</Issue>
                    <PubDate>
                        <Year>2011</Year>
                        <Month>Sep</Month>
                        <Day>2</Day>
                    </PubDate>
                </JournalIssue>
                <Title>Science (New York, N.Y.)</Title>
```

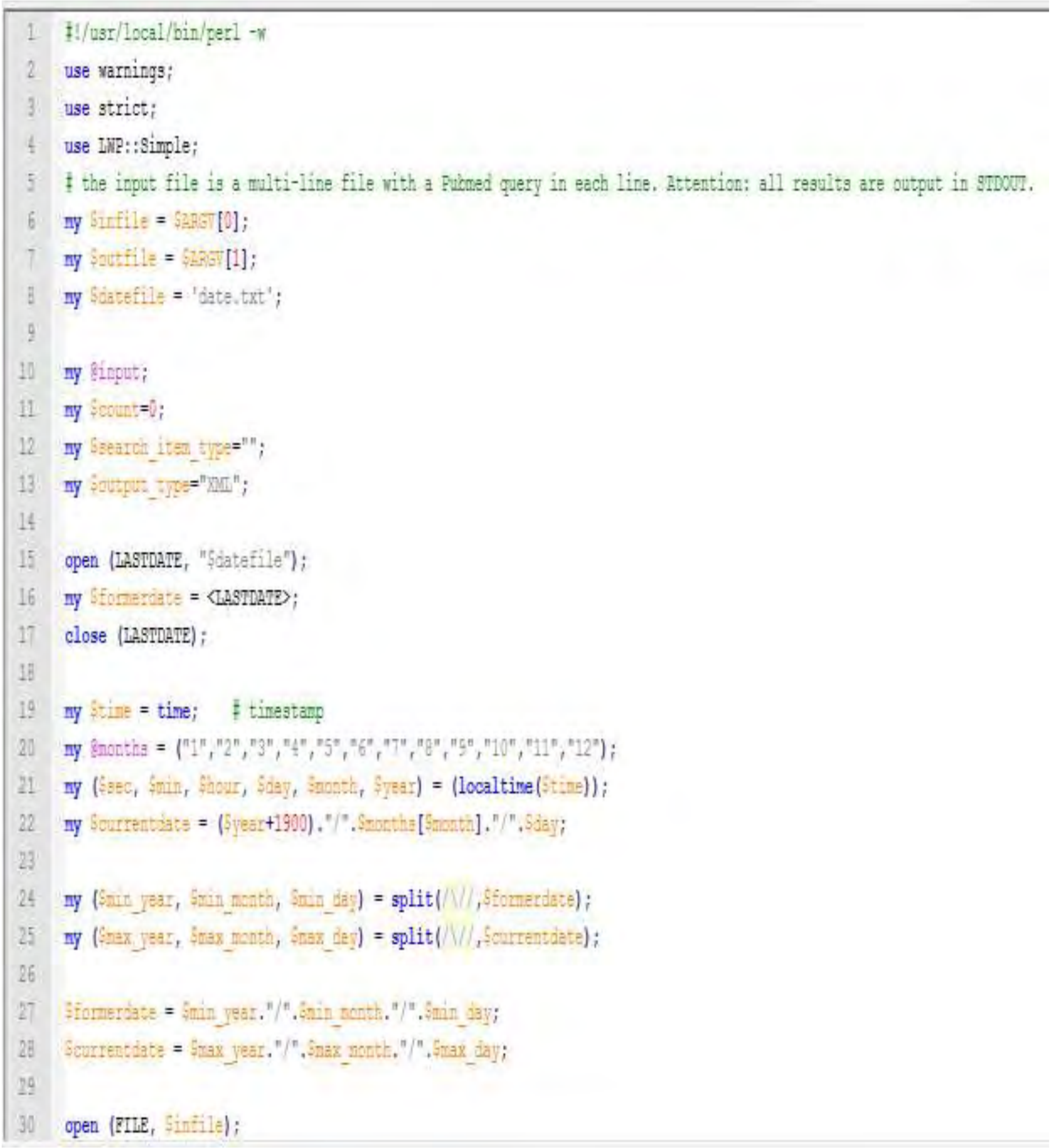


```
                 <Title>Science (New York, N.Y.)</Title>
```

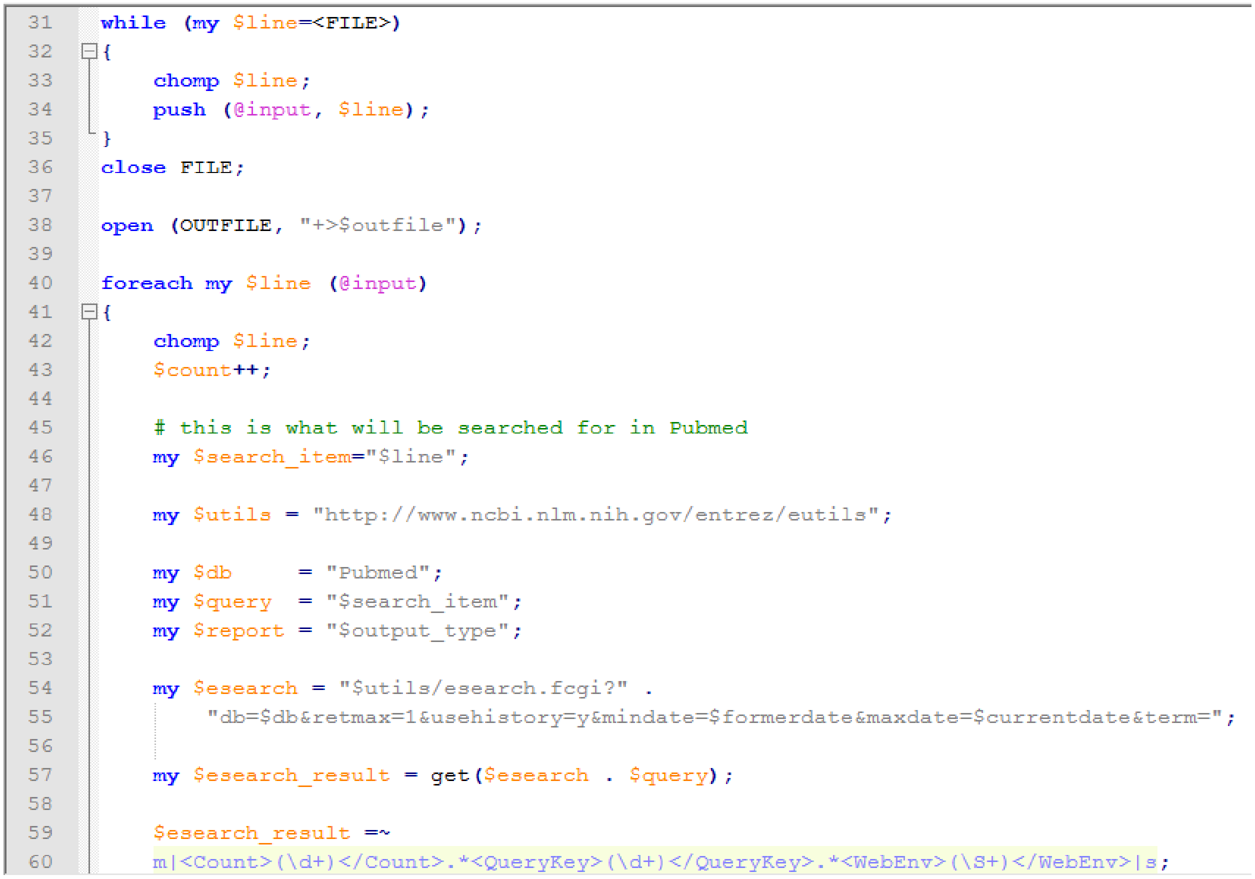
```
                 <ISOAbbreviation>Science</ISOAbbreviation>
             </Journal>
             <ArticleTitle>Multi-input RNAi-based logic circuit
for identification of specific cancer cells.</ArticleTitle>
             <Pagination>
                 <MedlinePgn>1307-11</MedlinePgn>
             </Pagination>
             <Abstract>
                 <AbstractText>Engineered biological systems that
integrate multi-input sensing, sophisticated information
processing, and precisely regulated actuation in living cells
could be useful in a variety of applications. For example,
anticancer therapies could be engineered to detect and respond
to complex cellular conditions in individual cells with high
specificity. Here, we show a scalable
transcriptional/posttranscriptional synthetic regulatory
circuit--a cell-type "classifier"--that senses expression levels
of a customizable set of endogenous microRNAs and triggers a
cellular response only if the expression levels match a
predetermined profile of interest. We demonstrate that a HeLa
cancer cell classifier selectively identifies HeLa cells and
triggers apoptosis without affecting non-HeLa cell types. This
approach also provides a general platform for programmed
responses to other complex cell states.</AbstractText>
             </Abstract>
             <Affiliation>Faculty of Arts and Sciences (FAS)
Center for Systems Biology, Harvard University, 52 Oxford
Street, Cambridge, MA 02138, USA.</Affiliation>
             <AuthorList CompleteYN="Y">
                 <Author ValidYN="Y">
                     <LastName>Xie</LastName>
```

3.6.1.2 Automatic update

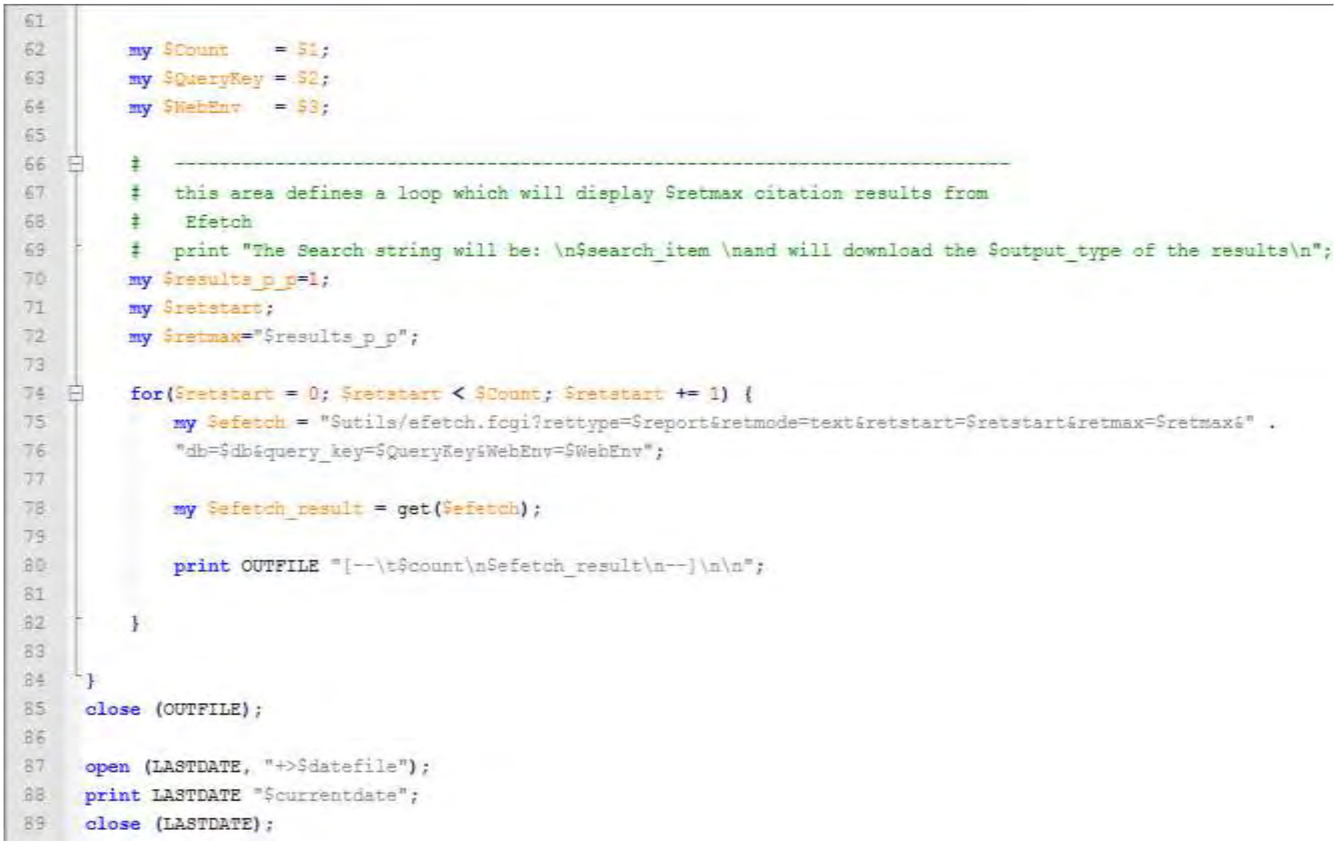
A different version of the perl script described above is shown below. Whenever the following code is executed only the extra abstracts from the previous update are retrieved. We present an automatic update since the program’s last execution. This leads to the retrieval of the most recent abstracts. We use the file “date.txt” to provide the initiation date of the script’s first execution. The code file is the PubMed_Search_and_Download_xml_updated.pl



```
#!/usr/local/bin/perl -w
use warnings;
use strict;
use LWP::Simple;
# the input file is a multi-line file with a Pubmed query in each line. Attention: all results are output in STDOUT.
my $infile = $ARGV[0];
my $outfile = $ARGV[1];
my $datefile = 'date.txt';

my @input;
my $count=0;
my $search_item_type="";
my $output_type="XML";

open (LASTDATE, "$datefile");
my $formerdate = <LASTDATE>;
close (LASTDATE);

my $time = time;    # timestamp
my @months = ("1","2","3","4","5","6","7","8","9","10","11","12");
my ($sec, $min, $hour, $day, $month, $year) = (localtime($time));
my $currentdate = ($year+1900)."/".$months[$month]."/".$day;

my ($min_year, $min_month, $min_day) = split(/\//,$formerdate);
my ($max_year, $max_month, $max_day) = split(/\//,$currentdate);

$formerdate = $min_year."/".$min_month."/".$min_day;
$currentdate = $max_year."/".$max_month."/".$max_day;

open (FILE, $infile);
```



```perl
while (my $line=<FILE>)
{
    chomp $line;
    push (@input, $line);
}
close FILE;

open (OUTFILE, "+>$outfile");

foreach my $line (@input)
{
    chomp $line;
    $count++;

    # this is what will be searched for in Pubmed
    my $search_item="$line";

    my $utils = "http://www.ncbi.nlm.nih.gov/entrez/eutils";

    my $db     = "Pubmed";
    my $query  = "$search_item";
    my $report = "$output_type";

    my $esearch = "$utils/esearch.fcgi?" .
        "db=$db&retmax=1&usehistory=y&mindate=$formerdate&maxdate=$currentdate&term=";

    my $esearch_result = get($esearch . $query);

    $esearch_result =~
    m|<Count>(\d+)</Count>.*<QueryKey>(\d+)</QueryKey>.*<WebEnv>(\S+)</WebEnv>|s;
```



```perl

    my $Count    = $1;
    my $QueryKey = $2;
    my $WebEnv   = $3;

    #  ---------------------------------------------------------------------------
    #  this area defines a loop which will display $retmax citation results from
    #   Efetch
    #  print "The Search string will be: \n$search_item \nand will download the $output_type of the results\n";
    my $results_p_p=1;
    my $retstart;
    my $retmax="$results_p_p";

    for($retstart = 0; $retstart < $Count; $retstart += 1) {
        my $efetch = "$utils/efetch.fcgi?rettype=$report&retmode=text&retstart=$retstart&retmax=$retmax&" .
        "db=$db&query_key=$QueryKey&WebEnv=$WebEnv";

        my $efetch_result = get($efetch);

        print OUTFILE "[--\t$count\n$efetch_result\n--]\n\n";

    }

}
close (OUTFILE);

open (LASTDATE, "+>$datefile");
print LASTDATE "$currentdate";
close (LASTDATE);
```

Line 8: the text file 'date.txt' contains the date, when the user executed the script for the last time and it is assigned to the variable $datefile, so as to be used later.

Lines 10-13: the needed variables are set and the retrieving data is set to be in XML format.

Lines 15-17: the file 'date.txt' opens and the date it contains is assigned to the variable $formerdate. Then we close this file.

Lines 19-22: the $time variable is initialized with the value returned from the 'time' function, we use the $time variable as a timestamp for the current excecution. Then, we define a scalar variable for the months of the year and we assign the arithmetic format of each month. We use the 'localtime' function, which converts the time returned by the ‘time’ function to list with the time of the local time zone. Afterwards, the variable $currentdate is set to be the current date of the computer in the format year/month/day.

Lines 24-28: firstly, the two variables, assigned with the dates, are being split to their components (year, month, day) using the ‘split’ function which splits the string into a list of strings and returns that list. Then are recreated in the format suitable for the next steps.

Lines 30-36: the input file opens and after any trailing string is removed the lines are assigned to the @input array. We close the input file.

Lines 38-85: with a foreach loop for every line of the input file, the search in PubMed takes place. Specifically, the $utils variable contains the route for the E-utilities and the ESearch call, which conducts the search, now is called with two extra parameters, the $mindate and the $maxdate. The variable $mindate sets the minimum publication date from which ESearch will start searching for parers and the $maxdate sets the maximum publication date to which ESearch will stop the searching. As $mindate we provide the value of the variable $formerdate and as $maxdate the value of the variable $currentdate. Then, the results of the ESearch call are stored in the $esearch_result variable and the EFetch call retrieves the abstracts.

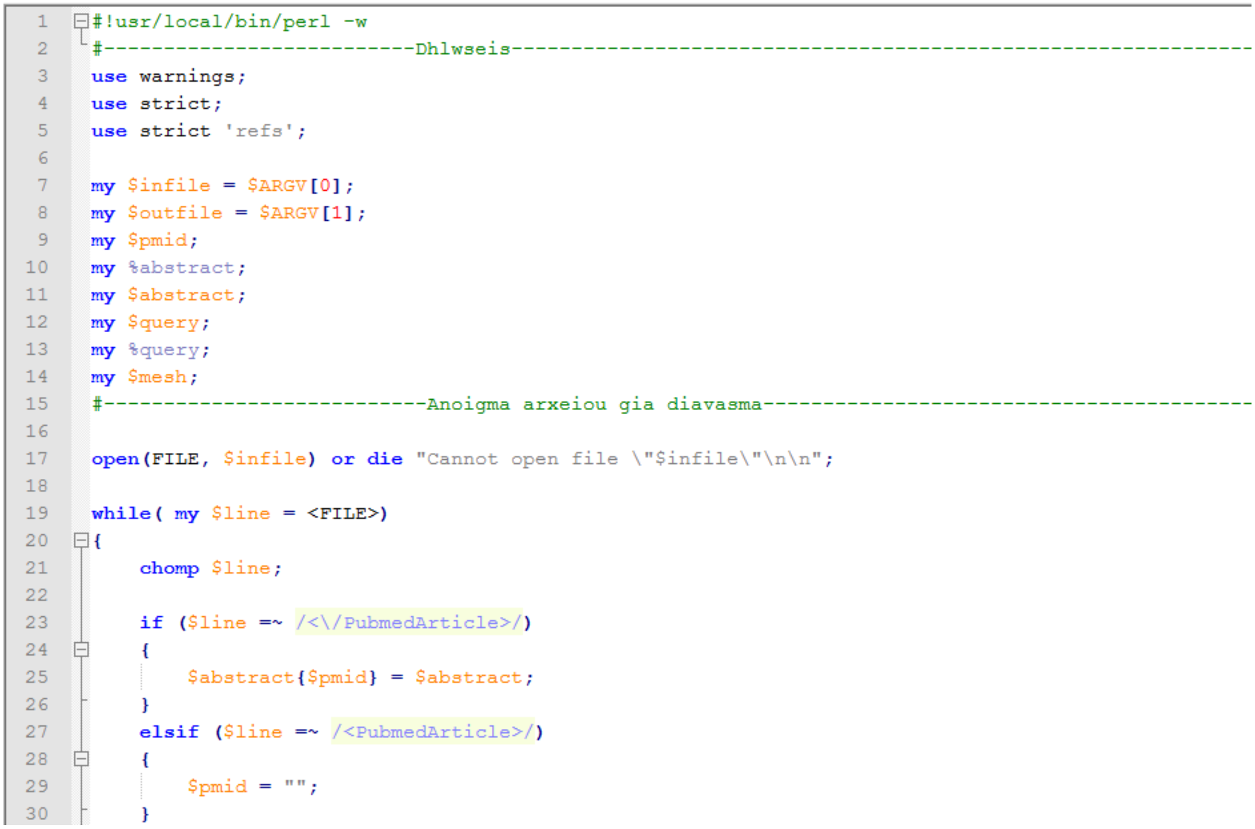
Lines 87-90: we open the file 'date.txt' with the operator +>, which means that its contents are being erased every time it opens, and we rewrite the current date in it. This command will provide the script with the date of last execution. Then, the file is being closed.

To run the script:

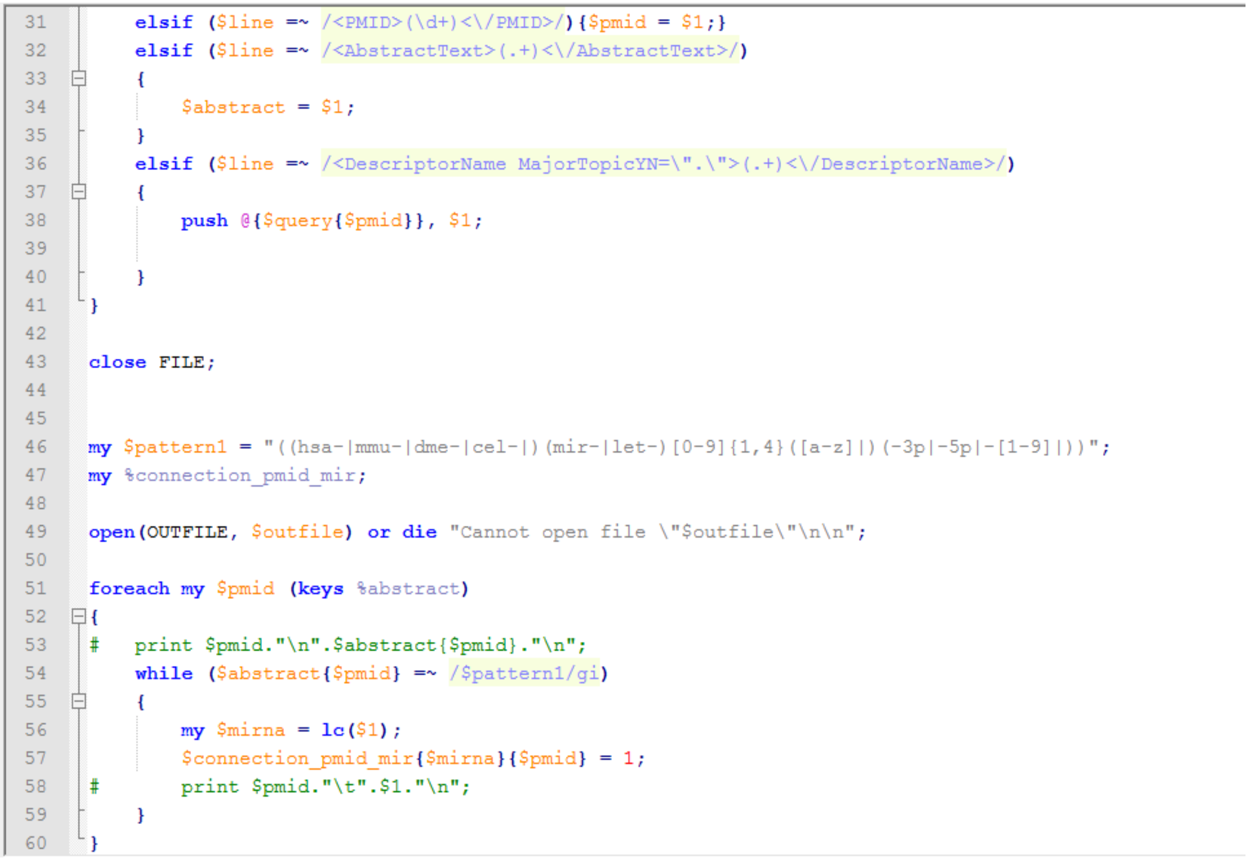
perl PubMed_Search_and_Download_xml_updated.pl infile.txt date.txt > abstracts_xml_updated.txt.

3.6.2 Extraction of MicroRNAs

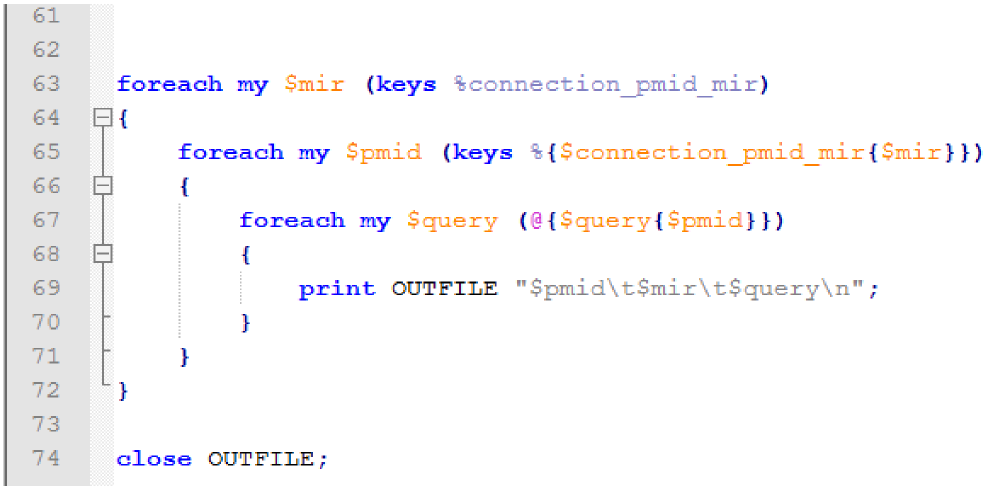
The Perl script presented below, parses the data stored in the output file, retrieved by the previous Perl script, by using a pattern. This pattern is a Perl's regular expression and the information we retrieve contains micoRNA names, MESH terms, and the literature PubMed ID. The code file is the Find_miRNA_in_xml_file.pl:



```perl
#!usr/local/bin/perl -w
#--------------------------Dhlwseis--------------------------------------------------------------
use warnings;
use strict;
use strict 'refs';

my $infile = $ARGV[0];
my $outfile = $ARGV[1];
my $pmid;
my %abstract;
my $abstract;
my $query;
my %query;
my $mesh;
#---------------------------Anoigma arxeiou gia diavasma------------------------------------------

open(FILE, $infile) or die "Cannot open file \"$infile\"\n\n";

while( my $line = <FILE>)
{
    chomp $line;

    if ($line =~ /<\/PubmedArticle>/)
    {
        $abstract{$pmid} = $abstract;
    }
    elsif ($line =~ /<PubmedArticle>/)
    {
        $pmid = "";
    }
```



```perl
    elsif ($line =~ /<PMID>(\d+)<\/PMID>/){$pmid = $1;}
    elsif ($line =~ /<AbstractText>(.+)<\/AbstractText>/)
    {
        $abstract = $1;
    }
    elsif ($line =~ /<DescriptorName MajorTopicYN=\".\">(.+)<\/DescriptorName>/)
    {
        push @{$query{$pmid}}, $1;

    }
}

close FILE;


my $pattern1 = "((hsa-|mmu-|dme-|cel-|)(mir-|let-)[0-9]{1,4}([a-z]|)(-3p|-5p|-[1-9]|))";
my %connection_pmid_mir;

open(OUTFILE, $outfile) or die "Cannot open file \"$outfile\"\n\n";

foreach my $pmid (keys %abstract)
{
#   print $pmid."\n".$abstract{$pmid}."\n";
    while ($abstract{$pmid} =~ /$pattern1/gi)
    {
        my $mirna = lc($1);
        $connection_pmid_mir{$mirna}{$pmid} = 1;
#       print $pmid."\t".$1."\n";
    }
}
```



```
61
62
63  foreach my $mir (keys %connection_pmid_mir)
64  {
65      foreach my $pmid (keys %{$connection_pmid_mir{$mir}})
66      {
67          foreach my $query (@{$query{$pmid}})
68          {
69              print OUTFILE "$pmid\t$mir\t$query\n";
70          }
71      }
72  }
73
74  close OUTFILE;
```

Lines 1-15 : statements of the variables used next. The $infile variable refers to the input file, the output of the previous perl script containing the abstracts retrieved. In line 5 we use the strict pragma ‘refs’ to prevent a possible runtime error.

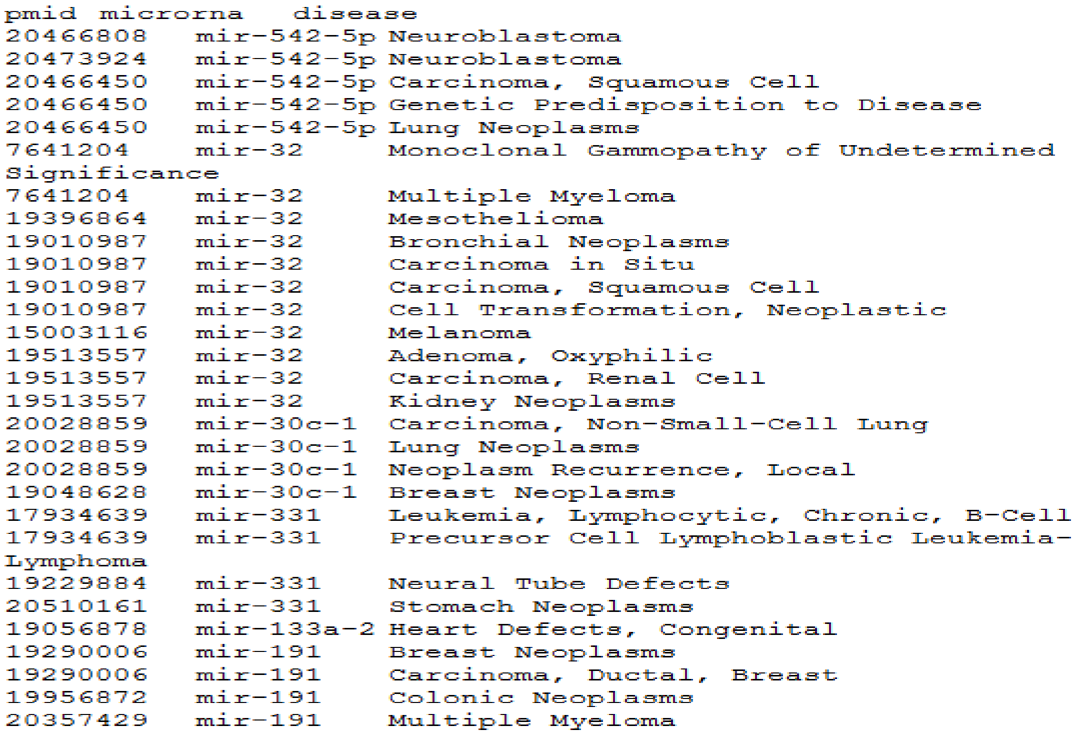
Lines 17-43 : we open the input file and for each abstract we put the corresponding PMID and MeSH terms into an array

Line 46 : we set the pattern, for matching a name of a microRNA in the title or abstract. This search pattern is a Perl’s regular expression based on the nomenclature rules of microRNAs.

Lines 51-60 : next we check,for each abstract based on the PMID, the existence of a microRNA name, using the pattern.

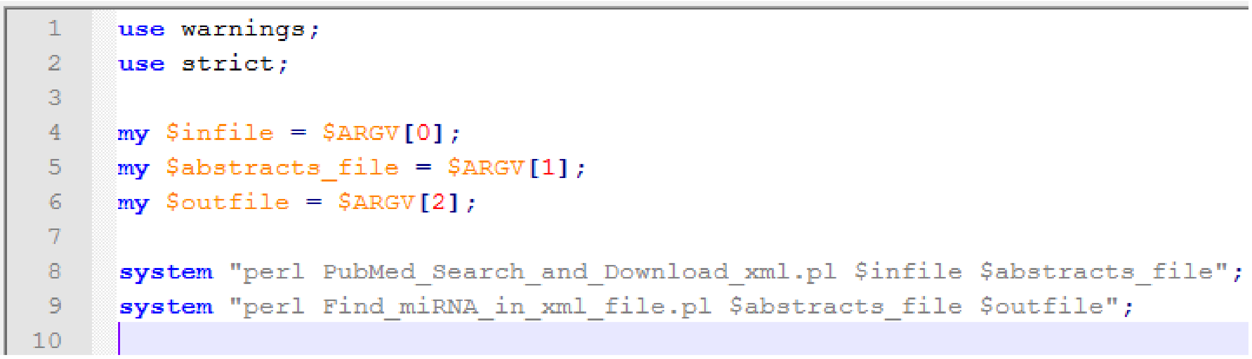
Lines 63-74 : finally, the PumMed ID (PMID), the corresponding microRNA name and the corresponding Mesh term are being retrieved in the ‘extraction_PMID_miRNA_MeSH.txt’ file.

To run the script:
perl Find_miRNA_in_xml_file.pl abstracts_xml.txt > extraction_PMID_miRNA_MeSH.txt. The output file will have the form presented below :



```
pmid  microrna   disease
20466808   mir-542-5p Neuroblastoma
20473924   mir-542-5p Neuroblastoma
20466450   mir-542-5p Carcinoma, Squamous Cell
20466450   mir-542-5p Genetic Predisposition to Disease
20466450   mir-542-5p Lung Neoplasms
7641204    mir-32     Monoclonal Gammopathy of Undetermined
Significance
7641204    mir-32     Multiple Myeloma
19396864   mir-32     Mesothelioma
19010987   mir-32     Bronchial Neoplasms
19010987   mir-32     Carcinoma in Situ
19010987   mir-32     Carcinoma, Squamous Cell
19010987   mir-32     Cell Transformation, Neoplastic
15003116   mir-32     Melanoma
19513557   mir-32     Adenoma, Oxyphilic
19513557   mir-32     Carcinoma, Renal Cell
19513557   mir-32     Kidney Neoplasms
20028859   mir-30c-1  Carcinoma, Non-Small-Cell Lung
20028859   mir-30c-1  Lung Neoplasms
20028859   mir-30c-1  Neoplasm Recurrence, Local
19048628   mir-30c-1  Breast Neoplasms
17934639   mir-331    Leukemia, Lymphocytic, Chronic, B-Cell
17934639   mir-331    Precursor Cell Lymphoblastic Leukemia-
Lymphoma
19229884   mir-331    Neural Tube Defects
20510161   mir-331    Stomach Neoplasms
19056878   mir-133a-2 Heart Defects, Congenital
19290006   mir-191    Breast Neoplasms
19290006   mir-191    Carcinoma, Ductal, Breast
19956872   mir-191    Colonic Neoplasms
20357429   mir-191    Multiple Myeloma
```

Finally, we can run all the Perl scripts described above using one command by executing the script below. We give to the command line:
perl Extraction_PMID_miRNA_MeSH.pl infile.txt abstracts_xml.txt extraction_PMID_miRNA_MeSH.txt,
where the $infile variable refers to the input file containing the MeSH terms given to PubMed, the $abstracts_file variable refers to the abstracts_xml.txt presented above and as a result we have the same as above extraction_PMID_miRNA_MeSH.txt.



```
use warnings;
use strict;

my $infile = $ARGV[0];
my $abstracts_file = $ARGV[1];
my $outfile = $ARGV[2];

system "perl PubMed_Search_and_Download_xml.pl $infile $abstracts_file";
system "perl Find_miRNA_in_xml_file.pl $abstracts_file $outfile";

```

[36]

4

DIANA-microT Web server

In this chapter we will present some of the functions and features of the updated version of DIANA-microT Web server, those in which the implementation of this thesis refers. DIANA-microT Web server is a software application which is used for microRNA target prediction and has been developed by the team of DIANA Lab (DNA Intelligent Analysis) of the Research Center of Biomedical Sciences "Alexander Fleming". The updated version of DIANA-microT Web server, DIANA-microT v4.0, for microRNA target prediction displays the associations between microRNAs and diseases and it is based on Artificial Neural Networks.

4.1 Introduction

One of the key solutions to the problem of decoding and recognizing the role of microRNAs in disease, is the computational microRNA target prediction, since the experimental method of identification of miRNA target genes is time consuming and expensive .In reply to this problem, many programs of microRNA target prediction have been developed, one of which is the DIANA-microT 3.0 developed by the DIANA Lab. This is an algorithm for the individually microRNA target prediction using various parameters. The algorithm provides an indication of the r of the false positive rate prediction by calculating, for each predicted interaction, a signal to noise ratio and score of accuracy. [36]

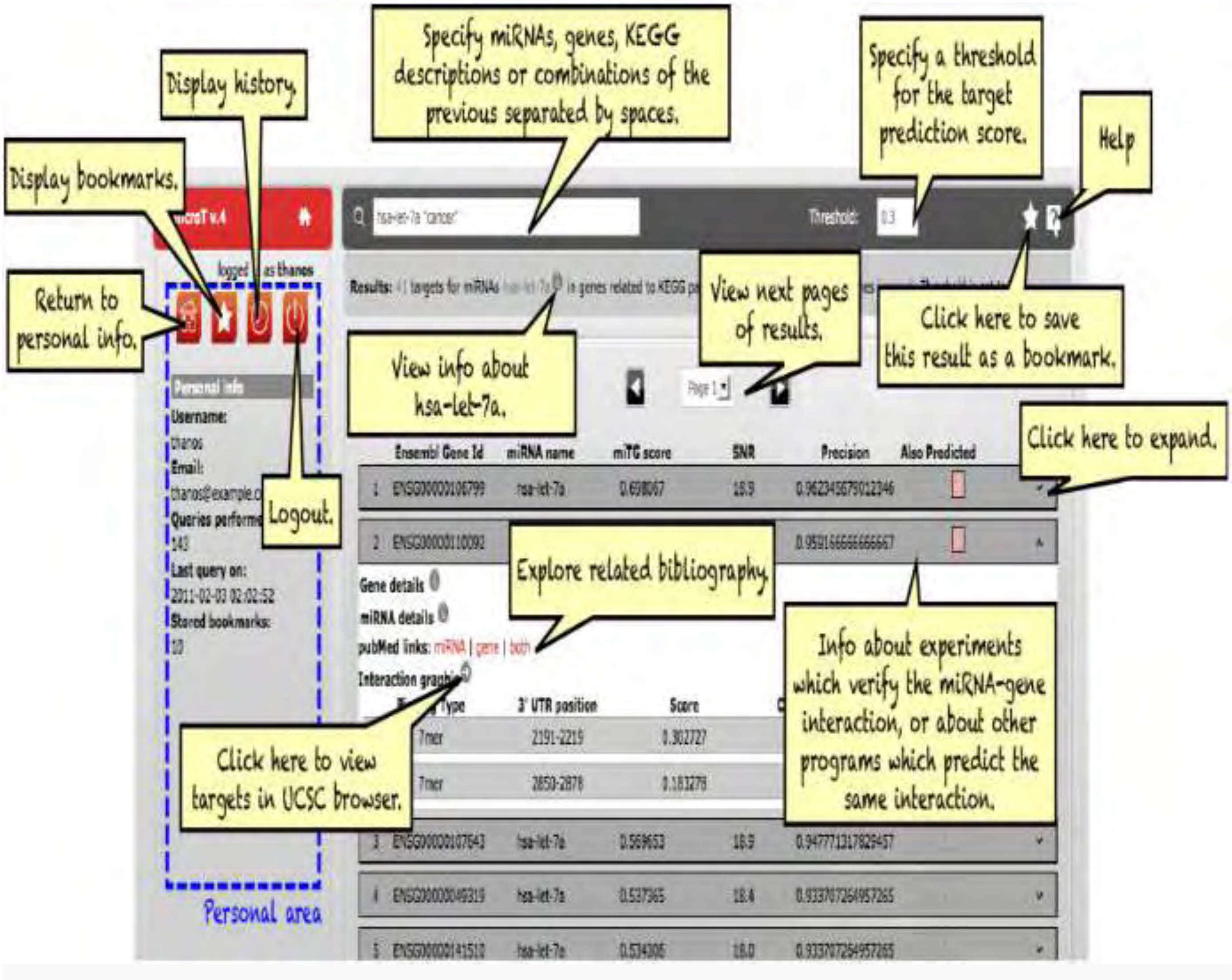
The results of the algorithm can be accessed by a user interface of DIANA-microT Web server, which provides, apart from information of miRNA target prediction, additional information from online biological databases. [37]

The current version of DIANA-microT Web server (DIANA-microT v4.0) was upgraded in such way so that it is able to process predictions for two additional species: Drosophila melanogaster and Caenorhabditis elegans. Moreover, through bibliographic analysis, DIANA-microT v4.0, displays the associations between microRNAs and diseases. With this additional functionality, the scientific research community is provided with useful information about the function of microRNAs in biological processes. DIANA-microT v4.0 home page is publicly available in www.microrna.gr/microT-v4.

The user can search for microRNA targets using the search box presented in the web page. The search can start by inserting various attributes. DIANA-microT v4.0 help page provides some examples :

- microRNA names, for example hsa-let-7a, hsa-miR-155.
- genes (gene name, Ensembl id or Refseq id) for example Fign .
- any combination of the above, for example hsa-let-7a ENSG00000149948. [38]

The help page also provides information presented for each microRNA gene interaction, and information for binding sites. It is also demonstrates some of the most important features of the web server:


Specify miRNAs, genes, KEGG descriptions or combinations of the previous separated by spaces.
Specify a threshold for the target prediction score.
Help
Display history.
Display bookmarks.
Return to personal info.
Logout.
View info about hsa-let-7a.
View next pages of results.
Click here to save this result as a bookmark.
Click here to expand.
Explore related bibliography.
Info about experiments which verify the miRNA-gene interaction, or about other programs which predict the same interaction.
Click here to view targets in UCSC browser.
Personal area
microT v.4
hsa-let-7a "cancer"
Threshold: 0.3
logged as thanos
Page 1

Personal info
Username:
thanos
Email:
thanos@example.c
Queries performe
143
Last query on:
2011-02-03 02:02:52
Stored bookmarks:
10

Results: 41 targets for miRNAs hsa-let-7a in genes related to KEGG p

	Ensembl Gene Id	miRNA name	miTG score	SNR	Precision	Also Predicted
1	ENSG00000106799	hsa-let-7a	0.698067	18.9	0.962345679012346	□
2	ENSG00000110092				0.959166666666667	□
3	ENSG00000107643	hsa-let-7a	0.569653	18.9	0.947771317829457	
4	ENSG00000049319	hsa-let-7a	0.537365	18.4	0.933707264957265	
5	ENSG00000141510	hsa-let-7a	0.534306	18.0	0.933707264957265	

Gene details
miRNA details
pubMed links: miRNA | gene | both
Interaction graphic

Binding Type	3' UTR position	Score
7mer	2191-2219	0.302727
7mer	2850-2878	0.183278

4.1 Features of DIANA-microT Web server

4.2 Relation of MicroRNAs to Diseases and Medical descriptors

Below is presented a part of the artcle *: DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association,* by Manolis Maragkakis, Thanasis Vergoulis, Panagiotis Alexiou, Martin Reczko, Kyriaki Plomaritou, Mixail Gousis, Kornilios Kourtis, Nectarios Koziris, Theodore Dalamagas, and Artemis G. Hatzigeorgiou, which refers relations of microRNAs to some basic functional features, diseases and MeSH terms :

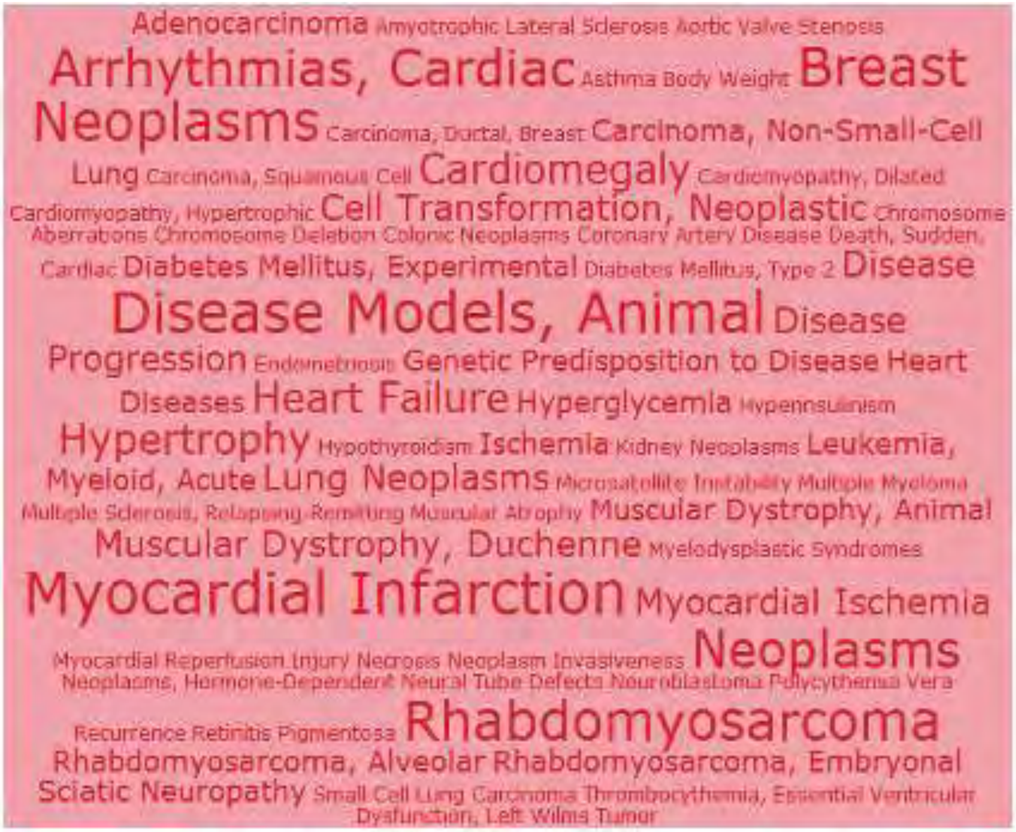
“DIANA microT Web server provides functional analysis of miRNAs that reaches beyond a simple listing of miRNA targets through integration of knowledge extracted from bibliography and known biological pathways.

In the previous version of the Web server, bibliographic integration considered automated searches in PubMed providing publications related to a miRNA, each target gene or combination of the two.

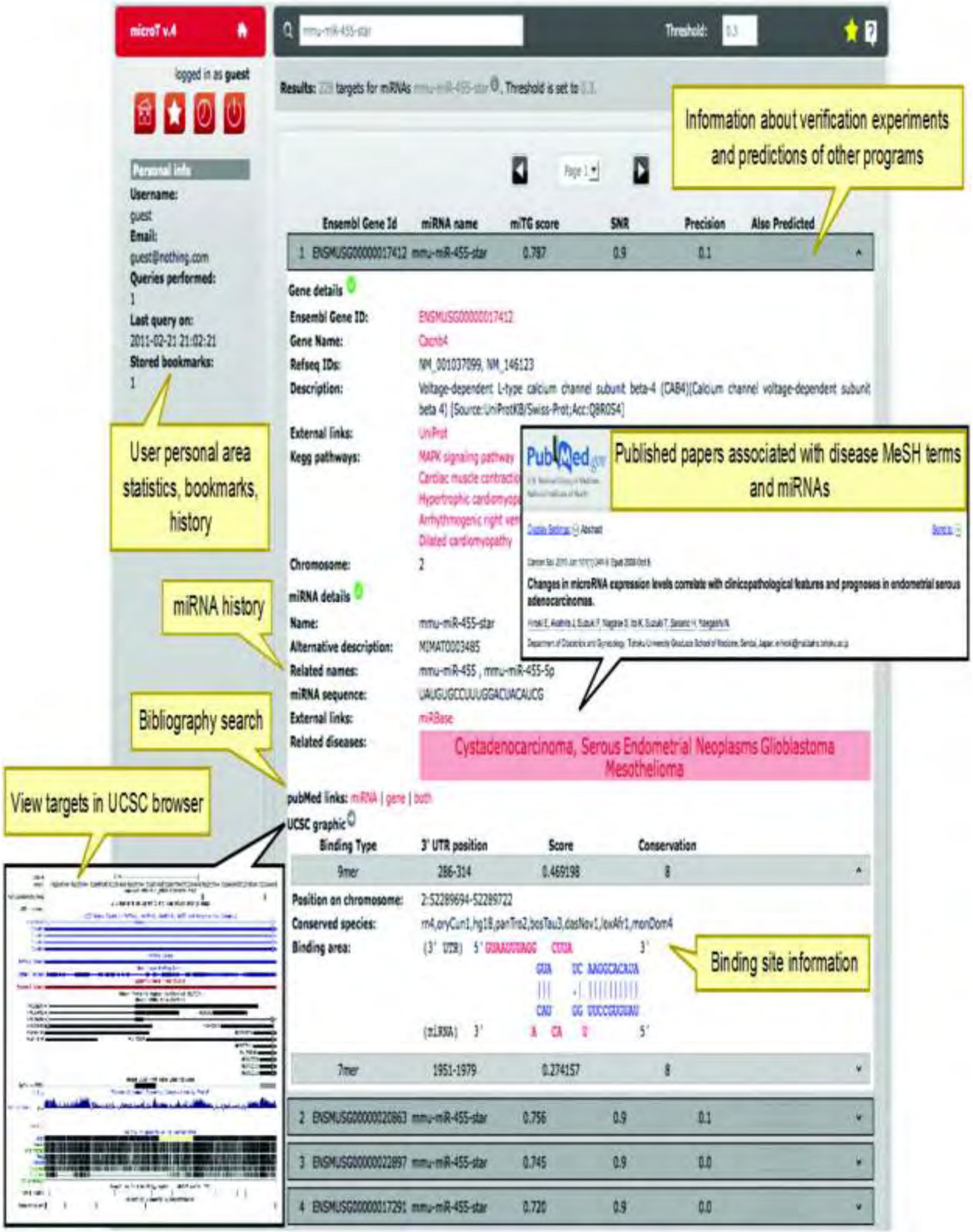
Now, an additional feature noted as ‘Related diseases’ has been added that directly associates a miRNA to publications connected to one or several diseases. This feature is based on information included in the title or the abstract of publications found in PubMed. All abstracts associated with a miRNA are retrieved from PubMed, based on the presence of the name of the miRNA or a member of its family, as defined by miRBase, in the title or abstract of the publication. The retrieved publications are associated with Medical Subject Headings (MeSH), the National Library of Medicine’s controlled vocabulary thesaurus, through their metadata.

All disease associated MeSH terms for a miRNA are counted and visualized through a tag cloud, where MeSH terms appear in a size proportional to the number of publications reporting this miRNA-disease association. The MeSH terms of the tag cloud also serve as hyperlinks to the relevant publications. The Perl script for creating a tag cloud of MeSH terms is presented in Appendix B. For example, in figure 4.3 miR-455-star (miR-455*) has been associated with a publication indicating that lower expression of this miRNA correlates with poor overall survival in endometrial serous adenocarcinoma.

In figure 4.3 we present an example of a DIANA-microT Web server results page. Balloons indicate and explain important features of the results page. ‘Related diseases’ tag cloud contains links to PubMed and specifies all papers which associate the particular disease with the corresponding miRNA. The field ‘PubMed links’ provides automated bibliography searches based only on the name of miRNAs, protein coding genes or the combination of both. The ‘UCSC graphic’ link presents the predicted binding sites in a UCSC genome browser window along with tracks such as SNPs and repeat elements. The left side of the page is devoted to the administration of the user personal space and reports their latest searches and bookmarks.”


Adenocarcinoma Amyotrophic Lateral Sclerosis Aortic Valve Stenosis
Arrhythmias, Cardiac Asthma Body Weight Breast
Neoplasms Carcinoma, Ductal, Breast Carcinoma, Non-Small-Cell
Lung Carcinoma, Squamous Cell Cardiomegaly Cardiomyopathy, Dilated
Cardiomyopathy, Hypertrophic Cell Transformation, Neoplastic Chromosome
Aberrations Chromosome Deletion Colonic Neoplasms Coronary Artery Disease Death, Sudden,
Cardiac Diabetes Mellitus, Experimental Diabetes Mellitus, Type 2 Disease
Disease Models, Animal Disease
Progression Endometriosis Genetic Predisposition to Disease Heart
Diseases Heart Failure Hyperglycemia Hyperinsulinism
Hypertrophy Hypothyroidism Ischemia Kidney Neoplasms Leukemia,
Myeloid, Acute Lung Neoplasms Microsatellite Instability Multiple Myeloma
Multiple Sclerosis, Relapsing-Remitting Muscular Atrophy Muscular Dystrophy, Animal
Muscular Dystrophy, Duchenne Myelodysplastic Syndromes
Myocardial Infarction Myocardial Ischemia
Myocardial Reperfusion Injury Necrosis Neoplasm Invasiveness Neoplasms
Neoplasms, Hormone-Dependent Neural Tube Defects Neuroblastoma Polycythemia Vera
Recurrence Retinitis Pigmentosa Rhabdomyosarcoma
Rhabdomyosarcoma, Alveolar Rhabdomyosarcoma, Embryonal
Sciatic Neuropathy Small Cell Lung Carcinoma Thrombocythemia, Essential Ventricular
Dysfunction, Left Wilms Tumor

4.2 An example of a tag cloud for hsa-miR-1 showing revelant disease assosiated MeSH terms.



logged in as **guest**

Personal info

Username: guest
Email: guest@nothing.com
Queries performed: 1
Last query on: 2011-02-21 21:02:21
Stored bookmarks: 1

Results: 228 targets for miRNAs mmu-miR-455-star, Threshold is set to 0.3.

	Ensembl Gene Id	miRNA name	miTG score	SNR	Precision	Also Predicted
1	ENSMUSG00000017412	mmu-miR-455-star	0.787	0.9	0.1	

Gene details

Ensembl Gene ID: ENSMUSG00000017412
Gene Name: Cacnb4
Refseq IDs: NM_001037099, NM_146123
Description: Voltage-dependent L-type calcium channel subunit beta-4 (CAB4)(Calcium channel voltage-dependent subunit beta 4) [Source:UniProtKB/Swiss-Prot;Acc:Q8R0S4]
External links: UniProt
Kegg pathways: MAPK signaling pathway
Cardiac muscle contractio
Hypertrophic cardiomyop
Arrhythmogenic right ven
Dilated cardiomyopathy
Chromosome: 2

miRNA details

Name: mmu-miR-455-star
Alternative description: MIMAT0003485
Related names: mmu-miR-455 , mmu-miR-455-5p
miRNA sequence: UAUGUGCCUUUGGACUACAUCG
External links: miRBase
Related diseases: Cystadenocarcinoma, Serous Endometrial Neoplasms Glioblastoma Mesothelioma

pubMed links: miRNA | gene | both

UCSC graphic

Binding Type	3' UTR position	Score	Conservation
9mer	286-314	0.469198	8

Position on chromosome: 2:52289694-52289722
Conserved species: rn4,oryCun1,hg18,panTro2,bosTau3,dasNov1,loxAfr1,monDom4
Binding area:

```
(3' UTR)  5' GUAAGUUAGG    CGUA        3'
                      GUA    UC AAGGCACAUA
                      |||    :| ||||||||||
                      CAU    GG UUCCGUGUAU
(miRNA)   3'         A   CA  U            5'
```

Binding Type	3' UTR position	Score	Conservation
7mer	1951-1979	0.274157	8

	Ensembl Gene Id	miRNA name	miTG score	SNR	Precision	Also Predicted
2	ENSMUSG00000020863	mmu-miR-455-star	0.756	0.9	0.1	
3	ENSMUSG00000022897	mmu-miR-455-star	0.745	0.9	0.0	
4	ENSMUSG00000017291	mmu-miR-455-star	0.720	0.9	0.0	

Cancer Sci. 2010 Jan;101(1):241-9. Epub 2009 Oct 8.

Changes in microRNA expression levels correlate with clinicopathological features and prognoses in endometrial serous adenocarcinomas.

Hiroki E, Akahira J, Suzuki F, Nagase S, Ito K, Suzuki T, Sasano H, Yaegashi N.

Department of Obstetrics and Gynecology, Tohoku University Graduate School of Medicine, Sendai, Japan.

Information about verification experiments and predictions of other programs

User personal area statistics, bookmarks, history

Published papers associated with disease MeSH terms and miRNAs

miRNA history

Bibliography search

View targets in UCSC browser

Binding site information

4.3 Example of a DIANA-microT Web server results page.

5 Epilogue

MicroRNAs play a key role as regulators of many cellular functions through post- transcriptional repression of their target genes. The significant increase of available information and interest of the scientific community, in recent years, for the function of microRNAs has led to the need to design and develop computational programs in order to analyze their function.

One of the objectives of this need for developing new computational methods relevant to the microRNAs target prediction, is the association of the function of microRNAs with diseases, which is referred to scientific articles with increasing frequency.

In this thesis, we made an attempt to extract associations of microRNAs with diseases from biological articles which was automatically retrieved from PubMed. Specifically we extracted the information : MicroRNA, MeSH terms, PMID, for each microRNA name that was referred to the title or abstract of the publication in an efficient way

This information was properly processed by the team of the DAINA Lab (DNA intelligent Analysis) of Biomedical Sciences Research Center "Alexander Fleming", so as to be used in the latest version of DIANA-microT Web server, a program for microRNA target prediction, in order to assist biologists and researchers retrieving important data in an efficient way.

6

References

[1] http://en.wikipedia.org/wiki/Bioinformatics

[2] http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html

[3] http://www.ebi.ac.uk/2can/bioinformatics/bioinf_realworld_1.html

[4] http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

[5] http://en.wikipedia.org/wiki/DNA

[6] http://en.wikipedia.org/wiki/Protein_synthesis

[7] http://en.wikipedia.org/wiki/Reverse_transcription#Process_of_reverse_transcription

[8] Du T, Zamore PD : "microPrimer: the biogenesis and function of microRNA", November 1, 2005, doi: 10.1242/dev.02070

[10] http://en.wikipedia.org/wiki/Human_Genome_Project

[11] http://www.genome.gov/25520322

[12] Mattick JS, Makunin IV : "Non-coding RNA, PubMed", PMID: 16651366

[13] http://www.ensembl.org/info/docs/genebuild/ncrna.htm : "Annotation of Non-Coding RNAs"

[14] Stephen F Madden, Susan B Carpenter, Ian B Jeffery, Harry Björkbacka , Katherine A Fitzgerald , Luke A O'Neill and Desmond G Higgins : "Detecting microRNA activity from gene expression data" BMC Bioinformatics 2010, doi:10.1186/1471-2105-11-257

[15] http://en.wikipedia.org/wiki/MicroRNA#cite_note-pmid15685193-13

[16]VICTOR AMBROS, BONNIE BARTEL, DAVID P. BARTEL, CHRISTOPHER B. BURGE, JAMES C. CARRINGTON, XUEMEI CHEN, GIDEON DREYFUSS, SEAN R EDDY, SAM GRIFFITHS-JONES, MHAIRI MARSHALL, MARJORI MATZKE, GARY RUVKUN, and THOMAS TUSCHL : “A uniform system for microRNA annotation”, RNA 2003. 9: 277-279, doi: 10.1261/rna.2183803

[17] Li M, Marin-Muller C, Bharadwaj U, Chow KH, Yao Q, Chen C : "MicroRNAs: control and loss of control in human physiology and disease” , April 2009, PMID: 19030926 [PubMed - indexed for MEDLINE] PMCID: PMC2933043

[18] Cannell IG, Kong YW, Bushell M : How do microRNAs regulate gene expression?, PMID: 19021530 [PubMed - indexed for MEDLINE]

[19]Mark Springer : “Understanding Regulation of Gene Expression by MicroRNAs Using Real-Time PCR Assays”, Reprinted from *American Biotechnology Laboratory* October 2005

[20]Pierre Maziere and Anton J. Enright : “Prediction of microRNAs targets”,Elsevier, June 2007

[21] Harris S Soifer, John J Rossi and Pål Sætrom : “MicroRNAs in Disease and Potential Therapeutic Applications” ,Molecular Therapy (2007) 15 12, 2070–2079, doi:10.1038/sj.mt.6300311

[22 Nicola Meola, Vincenzo Alessandro Gennarino and Sandro Banfi : “microRNAs and genetic diseases” , PathoGenetics (2009), doi: 10.1186/1755-8417-2-7,PubMed: 19889204

[23]Arti Gaur, David A. Jewell, Yu Liang, Dana Ridzon, Jason H. Moore, Caifu Chen, Victor R. Ambros, and Mark A. Israel : “Characterization of MicroRNA Expression Levels and Their Biological Correlates in Human Cancer Cell Lines” , doi: 10.1158/0008-5472.CAN-06-2698

Cancer Res March 15, 2007 67; 2456

[24] Erson AE, Petty EM : “MicroRNAs in development and disease” , PMID: 18713256 [PubMed - indexed for MEDLINE]

[25] Daniele Catalucci, PhD, Paolo Gallo, MD, Gianluigi Condorelli, MD, PhD : “MicroRNAs in Cardiovascular Biology and Heart Disease” Cardiovascular Genetics, 2009; 2: 402-408

doi: 10.1161/CIRCGENETICS.109.857425

[26] Eric M. Small, PhD, Robert J.A. Frost, MD, PhD, Eric N. Olson, PhD : “MicroRNAs Add a New Dimension to Cardiovascular Disease”, 2010; 121: 1022-1032 ,doi: 10.1161/

[27]http://www.regular-expressions.info/tutorial.html

[28]http://www.nlm.nih.gov/pubs/factsheets/mesh.html : Fact Sheet Medical Subject Headings (MeSH®)

[29] http://www.nlm.nih.gov/bsd/disted/mesh/indexprinc.html : Principles of MEDLINE Subject Indexing

[30]-http://www.ncbi.nlm.nih.gov/books/NBK3827/ : PubMed Help

[31]http://www.nlm.nih.gov/pubs/factsheets/pubmed.html : PubMed®: MEDLINE® Retrieval on the World Wide Web

[32]http://www.nlm.nih.gov/pubs/factsheets/medline.html : Fact Sheet MEDLINE®

[33]Monica Romiti, M.L.S. and Peter Cooper, Ph.D. : Entrez Help, http://www.ncbi.nlm.nih.gov/books/NBK3837/

[34 http://www.ncbi.nlm.nih.gov/books/NBK25501/ : Entrez Programming Utilities Help

[35 http://en.wikipedia.org/wiki/Perl

[36 http://perldoc.perl.org/ : Perl manual

[37]Manolis Maragkakis , Panagiotis Alexiou , Giorgio L Papadopoulos, Martin Reczko , Theodore Dalamagas , George Giannopoulos , George Goumas , Evangelos Koukis , Kornilios Kourtis, Victor A Simossis , Praveen Sethupathy , Thanasis Vergoulis , Nectarios Koziris , Timos Sellis , Panagiotis Tsanakas and Artemis G Hatzigeorgiou : “Accurate microRNA target prediction correlates with protein repression levels”, BMC Bioinformatics 2009, doi:10.1186/1471-2105-10-295

[38]M. Maragkakis, M. Reczko, V. A. Simossis, P. Alexiou, G. L. Papadopoulos, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas and A. G. Hatzigeorgiou, : “DIANA-microT web server: elucidating microRNA functions through target prediction”, Nucl. Acids Res. (2009) 37 (suppl 2): W273-W276.,doi: 10.1093/nar/gkp292

[39]http://diana.cslab.ece.ntua.gr/DianaTools/index.php?r=site/help : DIANA-microT v4.0 help page