Institut für Informatik
der Naturwissenschaftlichen Fakultät III
der
Martin-Luther-Universität Halle-Wittenberg

BIOINFORMATICS APPROACH FOR microRNA TARGET PREDICTION AND FUNCTIONAL ANALYSIS

Dissertation

zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

vorgelegt von

Emmanouil Maragkakis
geb. am 31.05.1983 in Athen

Gutachter:

1. Prof. Dr. Ivo Grosse
2. Dr. Artemis Hatzigeorgiou
3. Prof. Dr. Wojciech Makalowski

Datum der Verteidigung: 8 Juli 2011

Halle/Saale 2011

Bioinformatics approach for microRNA target prediction and functional analysis

Cumulative thesis submitted to the Faculty of Natural Sciences III in partial fulfillment of the requirements for the PhD Degree of the Martin-Luther University Halle-Wittenberg

by

Emmanouil Maragkakis

Halle/Saale 2011

“This dissertation is submitted as a cumulative thesis according to the guidelines provided by the PhD-program of Martin-Luther University Halle-Wittenberg. The thesis includes ten original papers addressing one topic, four of which comprise the majority of my research work during the course of PhD. The other papers are a collaborative effort.”

Emmanouil Maragkakis

CONTENTS

1. ABBREVIATIONS

RNA	ribonucleic acid
miRNA	microRNA
mRNA	messenger RNA
RISC	RNA-Induced Silencing Complex
CoProHMM	Conditional Profile Hidden Markov Model
3’UTR	3’ untranslated region
CDS	coding sequence
rRNA	ribosomal RNA
tRNA	transfer RNA
snRNA	small nuclear RNA
snoRNA	small nucleolar RNA
Ago	Argonaute
nt	nucleotide
endo-siRNA	endogenous small interfering RNA
esiRNA	endogenous small interfering RNA
piRNA	Piwi interacting RNA
pol II	polymerase II
H. sapiens	Homo sapiens
M. musculus	Mus musculus
D. melanogaster	Drosophila melanogaster
C. elegans	Caenorhabditis elegans

2. SUMMARY

2.1. English version

From September 2007 to May 2011 I have been working on the field of computational microRNA (miRNA) biology with emphasis on the development of accurate miRNA target prediction algorithms and the bioinformatics analysis of miRNA function. Here, I present a brief summary of the thesis resulted from this work.

miRNAs are small endogenous RNA molecules that play a key role in development and diseases through post-transcriptional regulation of gene expression. They are part of the RNA-Induced Silencing Complex (RISC) which they guide on the mRNA of target genes and induce translational repression and/or mRNA degradation (Ambros 2001; Bartel 2009). They have been found to confer a novel layer of genetic regulation in a wide range of biological processes and are involved in many stages of cancer progression by both promoting and/or suppressing oncogenesis (Joels, Matthews et al. 2005; Lee and Dutta 2007; Tagawa, Karube et al. 2007; Ivanovska, Ball et al. 2008). They are also involved in several biological functions and developmental stages and have been linked to several human pathologies such as cardiovascular and neurodegenerative diseases as well as human malignancies (Croce and Calin 2005; Esquela-Kerscher and Slack 2006; Garzon, Fabbri et al. 2006; Slack and Weidhaas 2006; Fabbri, Ivan et al. 2007; Gartel and Kandel 2008).

miRNA biology is a newly evolved field and there is great need for programs to address the scientific questions raised. Arguably, the most important role of miRNAs in the living cell is their targeting of mRNA molecules. Even though experimental validation of miRNA targets has been progressing in bounds in the past few years, the larger part of targeted genes still remains unverified. Therefore, the understanding of miRNA function goes hand in hand with the development of computational target prediction programs. Since the initial discovery of miRNAs several algorithms for miRNA target prediction have been developed but all of them still lack in terms of specificity and sensitivity (Kiriakidou, Nelson et al. 2004; Selbach, Schwanhausser et al. 2008). Additionally, there is great need for tools which can assess miRNA function in biological experiments and allow users to extract relevant information from biological data and resources.

In my work I have addressed miRNA target prediction by implementing two major releases of the microT program (papers 3, 10) and by contributing in the development of a novel alignment algorithm based on probabilistic models (paper 7). Also, I have contributed in a combined computational and experimental approach for assessing viral miRNA targeting against host genes (paper 8). I have participated in the development of two programs, one for the identification of miRNAs involved in the differential expression of genes (paper 5) and a second one for the assessment of miRNA involvement in biological pathways (paper 2). I have also developed a Web server which serves as an interface between bioinformatics tools and researchers and offers unique information regarding miRNA function (papers 1, 10). In addition, I have co-authored two reviews, one for evaluating available miRNA target prediction programs (paper 4) and the second for describing available online miRNA resources (paper 9). Overall, my work has resulted in 11 publications in international peer-reviewed journals. In the following I give an overview, in chronological order, of the scientific contributions of each of the papers.

In the miRNA field, information has been expanding in an increasing way in the last years. For this, the development of tools such as a target prediction program which provide primary data is not on its own sufficient and there is need for applications, primarily web based, which will serve as an interface between bioinformatics tools and researchers. These applications need to be able to organize the available information and present it in an intuitive and integrated way. To address this, I have developed a Web server which provides extensive information and wide connectivity to online biological resources in a user friendly interface. Target gene and miRNA functions are elucidated through automated bibliographic searches and functional information is extracted through KEGG (Kanehisa, Goto et al. 2004) pathways. Also, the server offers links to nomenclature, sequence and protein databases and users are facilitated by being able to search for targeted genes using different nomenclatures or functional features. Importantly, since miRNA target prediction is a computationally intensive task, I have developed an infrastructure in the computer cluster of the National Technical University of Athens to allow users to perform prediction for custom miRNA sequences, as part of the Web server. The work has been published in Maragkakis *et al* (Maragkakis, Reczko et al. 2009).

In order to identify molecular pathways potentially affected by the expression of single or multiple miRNAs I have contributed in the development of DIANA-mirPath, a functional analysis tool incorporating miRNA targets in biological pathways. It is a Web based application whose algorithm consists of an enrichment analysis of miRNA target genes within manually designed biological pathways. The combinatorial effect of co-expressed miRNAs in the modulation of a given pathway is taken into account through the analysis of multiple miRNAs simultaneously. This work has been published in Papadopoulos *et al* (Papadopoulos, Alexiou et al. 2009).

My work regarding miRNA target prediction resulted in a major release of the microT program, denoted as microT-v3.0. The program uses parameters which are calculated individually for each miRNA and computes a total score for predicted miRNA:target gene interactions as the weighted sum of scores for evolutionarily conserved and non-conserved binding sites. It is based on assessing whether in terms of evolutionary conservation a predicted miRNA binding site can be distinguished from a random background or not. The prediction performance of the program has been evaluated independently in a work published in Nature by Selbach *et al* (Selbach, Schwanhausser et al. 2008) and has been shown to be the most precise program available. The work has been published in Maragkakis *et al* (Maragkakis, Alexiou et al. 2009).

High-throughput gene expression experiments are widely used to identify the role of genes involved in biological conditions of interest. Similarly, the identification of miRNAs and the genes they regulate may provide potential ways for diagnosis and therapy in human diseases. Although miRNA expression levels may not be routinely measured in high-throughput experiments, a possible involvement of miRNAs in the deregulation of gene expression can be computationally predicted and quantified through analysis of overrepresented motifs in the 3′UTR sequences of deregulated genes. For this, I have participated in the development of DIANA-mirExTra to allow the comparison of frequencies of miRNA associated motifs between sets of genes that can lead to the identification of miRNAs responsible for the deregulation of large numbers of genes. I have also customized this program to be able to run in the computer cluster mentioned earlier allowing users to run the program through a Web interface. This work has been published in Alexiou *et al* (Alexiou, Maragkakis et al. 2010).

One of the major steps in a miRNA target prediction program is the alignment of the miRNA sequence against the target mRNA sequence for the identification of putative binding sites. For this, several approaches have been suggested but most of them are based on heuristic assumptions driven only by a few experimental data. To address this issue, I have contributed in the development of a novel data driven method based on the notion of Profile Hidden Markov Models. This method has been denoted as Conditional Profile Hidden Markov Model (CoProHMM) and has been shown to outperform existing alignment methods. This work has been published in Grau *et al* (Grau, Arend et al. 2010).

Usually, top performing target prediction programs exploit information regarding evolutionary conservation of predicted miRNA binding sites. However, this seemingly informative feature might as well decrease prediction performance in specific cases. This happens when miRNA targeting has a negative effect on the organism's survival and therefore the organism tends to avoid it. This is the case regarding viral miRNA targeting against a host organism. Taking this into account, I have participated in a combined computational and experimental approach for viral miRNAs of Epstein-Barr virus and found that ebv-miR-BART6-5p silence Dicer through multiple target sites located in the 3'UTR of Dicer mRNA and that mutation and A-to-I editing appear to be adaptive mechanisms that antagonize ebv-miR-BART6 activities that consequently affect viral latency. This work has been published in Iizasa *et al* (Iizasa, Wulff et al. 2010).

To enhance the scientific significance of the DIANA Web server I updated it to support predictions for two widely studied species: *Drosophila melanogaster* and *Caenorhabditis elegans*. Most importantly, in the updated version, I have associated miRNAs to diseases through bibliographic analysis and therefore provide insights for the potential involvement of miRNAs in biological processes. Also, I have contributed in the analysis of the nomenclature used to describe mature miRNAs along different miRBase (Griffiths-Jones 2006) versions, and have extracted the naming history of each miRNA. This enables the identification of miRNA publications regardless of possible nomenclature changes. The work has been published in Maragkakis, Vergoulis *et al* (Maragkakis, Vergoulis et al. 2011).

Chi *et al* (Chi, Zang et al. 2009) released a set of biological data which allowed the development of miRNA target prediction programs based on machine learning techniques. These data and the data of Hafner *et al* (Hafner, Landthaler et al. 2010) served as the base for the development of another release of microT denoted as microT-CDS which is a miRNA target prediction program that stands out not only because it is a purely data driven approach but also because it succeeds in assessing miRNA targeting both in the 3'UTR and the coding sequence of genes. Importantly, it is shown that targeting in the coding sequence is not only functional but also confers an important biological meaning. This is evident, since the inclusion of targets in the coding sequence increases prediction sensitivity by more than 10% and also increases prediction precision. The manuscript describing the work involved in the development of microT-CDS is currently under submission (Reczko, Maragkakis *et al,* under submission).

Finally, I have co-authored two reviews for miRNAs. The first one discusses and evaluates available miRNA target prediction methods (Alexiou, Maragkakis et al. 2009). This is of great importance particularly because miRNAs is a rather new scientific topic and in the last years more than a dozen miRNA target prediction programs have been developed. Therefore the evaluation of the prediction performance is a very critical step is choosing which of the programs are best to be used for experimental designs. The second review discusses available online resources for miRNA analysis in an attempt to assist biologist in acquainting themselves with available tools designed for miRNA analysis.

My publications have been cited 92 times since my first publication in 2009, resulting in an h-index of 4. The DIANA Web server which may be accessed at www.microrna.gr has received more than 697,931 page views by more than 110,016 users from more than 60 countries and it currently receives more than 30,000 page views by more than 5500 users per month.

2.2. German version

Seit September 2007 arbeite ich auf dem Gebiet der Bioinformatik an der Entwicklung neuer Algorithmen zur Vorhersage von Genen, die durch microRNAs (miRNAs) reguliert werden, sowie an der Funktionsaufklärung von miRNAs. Im Folgenden gebe ich eine kurze Zusammenfassung der daraus entstandenen Doktorarbeit.

MiRNAs sind kleine endogene RNA-Moleküle, die eine Schlüsselrolle in der Entwicklung der Zelle sowie in verschiedenen Krankheiten durch post-transkriptionelle Regulation der Genexpression spielen. Sie sind Teil des "RNA-Induced Silencing Complex" (RISC), den sie zur mRNA des Zielgens führen. Sie sind damit verantwortlich für die Induktion von translationaler Repression und/oder mRNA-Abbau (Ambros 2001; Bartel 2009). Diese Induktion wurde als neuer Mechanismus der Genregulation in einer Vielzahl von biologischen Prozessen identifiziert, und miRNAs sind an vielen Stadien der Tumorprogression durch Unterstützung und/oder Unterdrückung der Onkogenese beteiligt (Joels, Matthews et al. 2005; Lee and Dutta 2007; Tagawa, Karube et al. 2007; Ivanovska, Ball et al. 2008). MiRNAs sind ebenfalls an einer Vielzahl von biologischen Funktionen und zellulären Entwicklungsstadien beteiligt und wurden mit verschiedenen Erkrankungen wie beispielsweise Herz-Kreislauf- und neurodegenerativen Erkrankungen sowie Tumoren in Verbindung gebracht (Croce and Calin 2005; Esquela-Kerscher and Slack 2006; Garzon, Fabbri et al. 2006; Slack and Weidhaas 2006; Fabbri, Ivan et al. 2007; Gartel and Kandel 2008).

Die molekulare Funktionsweise von miRNAs ist noch weitestgehend unverstanden, und das Forschungsgebiet der MiRNA-Biologie ist noch sehr jung. Daher besteht ein großer Bedarf an neuen Algorithmen zur Bewältigung der experimentellen Daten mit dem Ziel, wichtige wissenschaftliche Fragen auf diesem sich rasant entwickelnden Forschungsgebiet zu beantworten. Eine wichtige Funktion von miRNAs besteht in ihrer gezielten Interaktion mit mRNA-Molekülen in der lebenden Zelle. Obwohl die experimentelle Verifikation dieser MiRNA-Ziel-Interaktionen in den letzten Jahren Fortschritte machte, ist der größte Teil der Zielgene bislang nicht verifiziert. Dies wiederum limitiert unser Verständnis der Funktionen von miRNAs. Mit Hilfe von Computermodellen zur Vorhersage von MiRNA-Ziel-Interaktionen lässt sich diese Kluft jedoch teilweise schließen. Seit der Entdeckung von miRNAs wurden verschiedene Algorithmen zur Vorhersage von MiRNA-Ziel-Interaktionen entwickelt, die jedoch in Bezug auf ihre Sensitivität und Spezifität starke Defizite aufweisen (Kiriakidou, Nelson et al. 2004; Selbach, Schwanhausser et al. 2008). Darüber hinaus besteht ein großer Bedarf an Softwaresystemen, die die Funktion von miRNAs in biologischen Experimenten beurteilen können und den Nutzer bei der Zusammenstellung relevanter Informationen aus verschiedenen biologischen Datenquellen unterstützen.

In meiner Arbeit habe ich mich mit der Vorhersage von MiRNA-Ziel-Interaktionen befasst. Konkret habe ich zwei verschiedenen Varianten des DIANA-microT-Programms entwickelt (3, 11) und zur Entwicklung eines neuartigen, auf probabilistischen Modellen basierenden, Alignment-Algorithmus beigetragen (7). Weiterhin habe ich eine Kombination aus experimentellen und computergestützten Methoden zur Analyse viraler MiRNA-Ziel-Interaktionen mit Genen des Wirtes des Virus entwickelt (8). Ich habe bei der Entwicklung eines Programms zur Identifikation von miRNAs, die an der differentiellen Expression von Genen beteiligt sind (5), sowie eines Programms zur Beurteilung der Wirkung von miRNAs in Signalwegen (2) mitgewirkt. Darüber hinaus habe ich einen Web-Server zur Funktionsaufklärung von miRNAs entwickelt (1, 10). Ich bin Koautor zweier Übersichtsartikel zur vergleichenden Bewertung von verfügbaren Programmen zur Vorhersage von MiRNA-Ziel-Interaktionen (4) sowie zur Beschreibung der auf dem Gebiet der MiRNA-Biologie verfügbaren Ressourcen (9). Zusammenfassend wurden meine Untersuchungen in 11 Publikationen in internationalen und begutachteten Zeitschriften veröffentlicht. Im Folgenden gebe ich einen Überblick der wissenschaftlichen Beiträge jeder Veröffentlichung in chronologischer Reihenfolge.

Die Menge verfügbarer Informationen auf dem Gebiet der MiRNA-Biologie ist in den letzten Jahren stark angewachsen. Unter diesem Gesichtspunkt ist die Entwicklung von einzelnen Werkzeugen wie beispielsweise eines Programms zur Vorhersage von MiRNA-Ziel-Interaktionen nicht ausreichend. Vielmehr werden vor allem Web-basierte Anwendungen benötigt, die experimentellen Forschern eine Schnittstelle zu Bioinformatik-Werkzeugen bieten. Diese Anwendungen müssen die verfügbare Information organisiert, intuitiv und als Integration verschiedener Quellen darstellen. Zu diesem Zweck habe ich einen benutzerfreundlichen Webserver entwickelt, der umfangreiche Informationen und eine Online-Anbindung an verschiedene biologische Ressourcen bietet. Die Funktion von Zielgenen und miRNAs wird durch automatische Literaturrecherchen aufgeklärt, und funktionelle Informationen werden aus der KEGG-Datenbank zu genetischen Signalwegen (Kanehisa, Goto et al. 2004) extrahiert. Der Webserver bietet ebenfalls Verweise zu Nomenklatur, Sequenz- und Protein-Datenbanken, und die Benutzer werden bei ihren Suchen durch die Verwendung unterschiedlicher Nomenklatur oder funktioneller Merkmale unterstützt. Da die Vorhersage der MiRNA-Ziel-Interaktionen sehr rechenintensiv ist, habe ich eine parallele Implementierung für den Rechen-Cluster der Nationalen Technischen Universität von Athen realisiert, die auch die Vorhersage benutzerdefinierter Sequenzen durch den Webserver erlaubt (Maragkakis, Reczko et al. 2009).

Zur Identifizierung molekularer Signalwege, die durch die Expression einzelner miRNAs oder von Gruppen von miRNAs potentiell beeinflusst sind, habe ich an der Entwicklung von DIANA-mirPath, einem funktionellen Analyse-Werkzeug, das MiRNA-Ziel-Interaktionen in biologische Signalwege einbezieht, mitgewirkt. DIANA-mirPath ist eine Web-basierte Anwendung, deren Algorithmus auf einer Anreicherungs-Analyse von MiRNA-Zielgenen innerhalb der durch Biologen annotierten Signalwege beruht. Die kombinatorische Wirkung von ko-exprimierten miRNAs bei der Modulation eines gegebenen Signalweges wird durch die parallele Analyse mehrerer miRNAs berücksichtigt (Papadopoulos, Alexiou et al. 2009).

Meine Arbeiten an der Vorhersage von MiRNA-Ziel-Interaktionen führten zur dritten Haupt-Version des DIANA-microT-Programms. Dieses Programm verwendet Parameter, die für jede miRNA individuell angepasst werden, und berechnet einen Gesamtscore für die vorhergesagten MiRNA-Zielgen-Interaktionen aus einer gewichteten Summe der Scores von evolutionär konservierten und nicht-konservierten Bindungsstellen. Die Bewertung basiert auf der Wahrscheinlichkeit, mit der eine vorhergesagte MiRNA-Bindungsstelle im Hinblick auf die evolutionäre Konservierung von einer zufälligen Hintergrundverteitung unterschieden werden kann. Die Vorhersagegenauigkeit des Programms wurde in einer unabhängigen in Nature veröffentlichten Vergleichsstudie (Selbach, Schwanhausser et al. 2008) auf der Basis von experimentellen Daten ausgewertet und erwies sich unter den derzeit verfügbaren Programmen als das mit der höchsten Vorhersagegenauigkeit. Diese Arbeiten wurden in (Maragkakis, Alexiou et al. 2009) veröffentlicht.

Genexpressions-Experimente mit hohem Durchsatz werden häufig zur Detektion von Genen verwendet, die unter wichtigen biologischen Bedingungen exprimiert werden. In ähnlicher Weise können auch miRNAs und die durch sie regulierten Gene identifiziert werden, was wiederum potentiell neue Wege zur Diagnose und Therapie verschiedener Krankheiten eröffnet. Obwohl die Expressionswerte von miRNAs nicht routinemäßig in Genexpressions-Experimenten gemessen werden, kann dennoch eine mögliche Beteiligung von miRNAs bei der Deregulierung der Genexpression computergestützt vorhergesagt und durch die Analyse von überrepräsentierten Sequenz-Motiven in den 3-'UTR-Sequenzen von deregulierten Genen quantifiziert werden. Zu diesem Zweck wurde das DIANA-mirExTra-Programm zum Vergleich der Häufigkeiten der mit miRNAs assoziierten Sequenzmotive in verschiedenen Gruppen von Genen entwickelt, um die für einen größeren Teil der deregulierten Gene verantwortlichen miRNAs zu identifizieren. Ich habe auch dieses Programm für den oben genannten Rechen-Cluster parallelisiert, um interaktive Analysen mit einer Web-Schnittstelle zu ermöglichen. Diese Arbeit wurde in (Alexiou, Maragkakis et al. 2010) veröffentlicht.

Eine der wichtigsten Teilaufgaben bei der Vorhersage von MiRNA-Ziel-Interaktionen ist die Ausrichtung der miRNA-Sequenz entlang der mRNA-Sequenz zur Identifizierung von putativen Bindungsstellen. Zu diesem Zweck wurden verschiedene Ansätze vorgeschlagen, die jedoch meistens auf heuristischen Annahmen basieren, die nur durch wenige experimentelle Daten gestützt werden. In diesem Zusammenhang habe ich ein evidenzbasiertes Modell mitentwickelt, das auf Profile Hidden Markov Modellen aufbaut und Conditional Profile Hidden Markov Modell (CoProHMM) genannt wird. Wir konnten zeigen, dass CoProHMMs anderen Sequenz-Ausrichtingsmethoden überlegen sind und haben diese Arbeit in (Grau, Arend et al. 2010) veröffentlicht.

In den meisten Fällen verwenden die leistungsfähigsten Programme zur Vorhersage von MiRNA-Ziel-Interaktionen Informationen über evolutionäre Konservierung der vorhergesagten MiRNA-Bindungsstellen. Interessanterweise kann dieses scheinbar informative Merkmal in bestimmten Fällen die Vorhersagegenauigkeit verringern. Einer dieser Fälle ergibt sich, wenn die Regulation durch eine miRNA eine negative Wirkung auf die Überlebenchancen des Organismus hat und der Organismus daher versucht, die Bindung der miRNA zu vermeiden. Dies ist beispielsweise bei auf den Wirtsorganismus gerichteten viralen miRNAs der Fall. Unter Berücksichtigung dieser Einflüsse führte ich eine kombinierte computergestützte und experimentelle Untersuchung für die viralen miRNAs des Epstein-Barr-Virus durch. Es wurde festgestellt, dass die virale miRNA EBV-miR-BART6-5p das Dicer-Protein durch mehrere Bindungsstellen in der 3'-UTR der Dicer-mRNA unterdrückt und dass sowohl Mutationen als auch aktive Adenosine-zu-Inosine-RNA-Bearbeitung (A-to-I editing) als adaptive Mechanismen wirksam zu sein scheinen, die dem Einfluss von EBV-miR-BART6 entgegenwirken und dadurch die virale Latenz beeinflussen. Diese Arbeit wurde in (Iizasa, Wulff et al. 2010) veröffentlicht.

Um die Funktionalität des DIANA-Webservers zu erhöhen, habe ich zwei weitere Modellorganismen, *Drosophila melanogaster* und *Caenorhabditis elegans*, aufgenommen. Im Rahmen dieser Aktualisierung wurde auch eine automatische Literaturanalyse zur Assoziation von miRNAs mit verschiedenen Krankheiten implementiert, die neue Einblicke in die mögliche Beteiligung von miRNAs an verschiedenen biologischen Prozessen ermöglicht. Weiterhin habe ich die Entwicklung der in verschiedenen Versionen der MiRBase-Datenbank (Griffiths-Jones 2006) verwendeten Nomenklatur analysiert und so eine Identifikator-Historie für jede miRNA extrahiert. Dies ermöglicht eine von Namensänderungen unabhängige Identifizierung von miRNAs in Publikationen. Diese Arbeit ist in (Maragkakis, Vergoulis et al. 2011) erschienen.

(Chi, Zang et al. 2009) haben eine Reihe von biologischen Daten veröffentlicht, die die Anwendung von Methoden des Maschinellen Lernens zur Entwicklung von Programmen zur Vorhersage von MiRNA-Ziel-Interaktionen ermöglicht. Diese und die Daten von (Hafner, Landthaler et al. 2010) dienten als Grundlage für die Entwicklung einer weiteren Version von DIANA-microT. Das mit microT-CDS bezeichnete Programm zur Vorhersage von MiRNA-Ziel-Interaktionen zeichnet sich nicht nur durch seinen ausschließlich datengetriebenen Ansatz aus, sondern auch durch seine erfolgreiche Vorhersage von MiRNA-Bindungsstellen, die sowohl in der 3'-UTR als auch in der protein-kodierenden Sequenz von Genen auftauchen können. Wir konnten nachweisen, dass die MiRNA-Bindungen in den kodierenden Sequenzen funktional sind und dass die Berücksichtigung der Bindungsstellen in den kodierenden Sequenzen zu einer signifikanten Erhöhung der Vorhersagegenauigkeit führt. Das Manuskript zur Beschreibung dieser Untersuchung ist derzeit in Vorbereitung.

Ich bin Mitautor zweier Übersichtsartikel im Bereich der computergestützten Analyse von miRNAs. Der erste präsentiert und vergleicht die derzeit verfügbaren Programme zur Vorhersage von MiRNA-Ziel-Interaktionen (Alexiou, Maragkakis et al. 2009). Ein solcher Vergleich ist wichtig, da das Forschungsgebiet der MiRNA-Biologie relativ jung ist und in den letzten Jahren mehr als ein Dutzend Programme zur Vorhersage von MiRNA-Ziel-Interaktionen entwickelt wurden, ein objektiver Leistungsvergleich dieser Programme jedoch bislang fehlte. Ein solcher Vergleich aber ist ein notwendiger Schritt bei der Entscheidung, welches Programm am besten zum Entwurf von Experimenten zu verwenden ist. Der zweite Übersichtsartikel beschreibt verfügbare Online-Ressourcen zur computergestützten Analyse von miRNAs und versucht, Biologen mit diesen Ressourcen vertraut zu machen.

Meine Publikationen sind seit meiner ersten Veröffentlichung im Jahr 2009 92 Mal zitiert worden, was einem h-Index von 4 entspricht. Der DIANA-Webserver, der unter www.microrna.gr aufgerufen werden kann, erhielt bislang mehr als 697.931 Seitenaufrufe von mehr als 110.016 Benutzern aus über 60 Ländern. Er erhält derzeit monatlich mehr als 30.000 Seitenaufrufe von mehr als 5500 Nutzern.

3. GENERAL INTRODUCTION

The traditional notion until the end of the last century has been that the primary and almost exclusive role of RNA in cells is to carry genetic information for the translation of DNA into protein. However, the discovery of small non-coding RNAs and other non-coding RNAs has forced a paradigm shift on the perceived roles of RNA in the cell and gene regulation in general.

Historically, the term “small RNA” has been used for a wide variety of RNAs which have short length. It has been used for ribosomal RNA (rRNA), transfer RNAs (tRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs) and has also been largely associated with short regulatory RNAs. The later are largely implicated in eukaryotic cell silencing pathway and are distinguished primarily by their small size (~20–30 nucleotides) and their association with the Argonaute (Ago) family proteins. In humans there are at least three classes of small regulatory RNAs that are encoded. Based on their biogenesis mechanism and the type of the Ago protein that they are associated with, they are clustered into three distinct types: microRNAs (miRNAs), endogenous small interfering RNAs (endo-siRNAs or esiRNAs) and Piwi interacting RNAs (piRNAs).

miRNAs are small; approximately 21 to 22 nucleotide long, single stranded non-coding regulatory RNAs which were first discovered in *Caenorhabditis elegans* for their role in regulating the expression of protein coding genes (Lee, Feinbaum et al. 1993). However, although they were identified as early as 1993 it was not until 2001 that they were suggested to be widespread and abundant in cells (Lagos-Quintana, Rauhut et al. 2001; Lau, Lim et al. 2001; Lee and Ambros 2001). Since then, hundreds of more miRNA molecules have been identified in an increasing number of species including viruses, plants, nematodes, mice, and humans, suggesting a deep and important role in gene regulation and cell biology in general (Bartel and Bartel 2003; Berezikov and Plasterk 2005).

miRNAs have been found to confer a novel layer of genetic regulation in a wide range of biological processes. Their involvement in cellular commitment and cell cycle regulation gives them an important role in animal development and human diseases. Specifically, miRNAs have been linked to several human pathologies such as cardiovascular and neurodegenerative diseases (Hebert and De Strooper 2007; Hebert, Horre et al. 2008; Zhang 2008) as well as in human malignancies (Croce and Calin 2005; Esquela-Kerscher and Slack 2006; Garzon, Fabbri et al. 2006; Slack and Weidhaas 2006; Fabbri, Ivan et al. 2007; Gartel and Kandel 2008). Also they have been found to regulate various developmental stages in animals such as *Caenorhabditis elegans* (Lee, Feinbaum et al. 1993; Reinhart, Slack et al. 2000; Lau, Lim et al. 2001; Lee and Ambros 2001), *Danio rerio* (Wienholds, Kloosterman et al. 2005), *Drosophila melanogaster* (Aravin, Lagos-Quintana et al. 2003), *Mus musculus* (Baroukh, Ravier et al. 2007), *Homo sapiens* (Chen, Li et al. 2004; Yi, O'Carroll et al. 2006; Lu, Thomson et al. 2007) and in plants (Kidner and Martienssen 2005). In particular, miRNAs are believed to be involved in many stages of cancer progression by both promoting and/or suppressing oncogenesis (He, Thomson et al. 2005; Lee and Dutta 2007; Tagawa, Karube et al. 2007; Ivanovska, Ball et al. 2008), tumor growth (Johnson, Esquela-Kerscher et al. 2007; Si, Zhu et al. 2007), invasion and metastasis (Ma, Teruya-Feldstein et al. 2007; Asangani, Rasheed et al. 2008; Huang, Gumireddy et al. 2008; Tavazoie, Alarcon et al. 2008; Zhu, Wu et al. 2008).

For many years, researchers have been analyzing microarray expression data of protein coding genes in different cancer types in order to identify specific expression signatures. The limited number of miRNAs makes them an ideal candidate for this type of analysis. Currently, there are approximately 700 human miRNAs registered in miRBase (Griffiths-Jones, Saini et al. 2008), and according to estimates their number may reach 1000. Analyzing their expression, several miRNA signatures have already been successfully associated with human cancers (Calin and Croce 2006) such as leukemia (Calin and Croce 2007; Landais, Landry et al. 2007), thyroid carcinomas (He, Jazdzewski et al. 2005), breast (Iorio, Ferracin et al. 2005), lung (Yanaihara, Caplen et al. 2006) and pancreatic cancer (Lee, Gusev et al. 2007).

Most mammalian miRNAs are transcribed by RNA polymerase II (pol II) (Lee, Kim et al. 2004), the same polymerase that directs the transcription of protein coding genes. However there are some Alu repeat associated miRNAs which are known to be transcribed by RNA polymerase III (pol III) (Borchert, Lanier et al. 2006). miRNAs are encoded in sense or anti-sense orientation within introns of protein coding genes and in non-coding transcripts. Most mammalian miRNA genes have multiple isoforms (paralogues) that are probably the result of gene duplications. Approximately 50% of mammalian miRNA loci are found in close proximity to other miRNAs. These clustered miRNAs are considered to be transcribed from a single polycistronic transcription unit (Lee, Jeon et al. 2002), although there may be exceptional cases in which individual nearby miRNAs are derived from separate gene promoters (Figure 1).

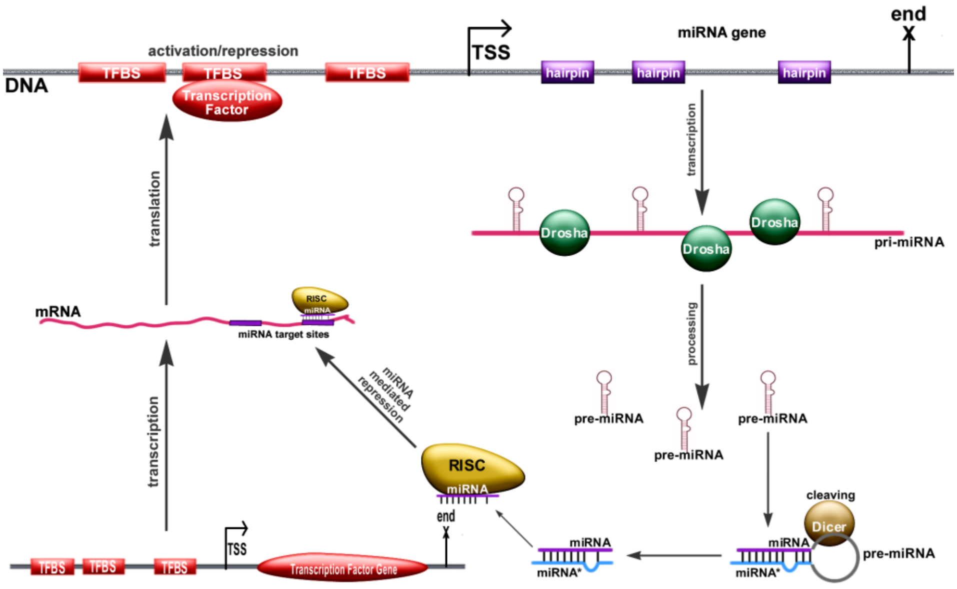


Figure 1: A miRNA gene is controlled by several TFs whose binding sites (TFBSs) are located near the Transcription Start Site of this gene. When transcribed, the miRNA gene produces a long pri-miRNA molecule. The pri-miRNA molecule is cleaved by Drosha and yields the pre-miRNA stem-loop (hairpin) structure. The enzyme Dicer cleaves the loop of the hairpin and produces the miRNA-miRNA* duplex. One chain of the miRNA duplex is incorporated into the RISC complex and can regulate mRNA translation. In this example, the miRNA regulates the translation of the promoter in a typical negative feedback control loop. Image taken from Alexiou *et al* (Alexiou, Vergoulis et al. 2010)

The primary transcripts (pri-miRNAs) that are generated by Pol II are usually several kilobases long and contain local stem-loop structures. The first step of miRNA maturation is cleavage at the stem of the hairpin structure, which releases a small hairpin that is termed a pre-miRNA. This reaction takes place in the nucleus by the nuclear RNase III type protein Drosha (Lee, Ahn et al. 2003). Drosha requires a cofactor, the DiGeorge syndrome critical region gene 8 (DGCR8) protein in humans (Pasha in *Drosophila melanogaster* and *Caenorhabditis elegans*) (Denli, Tops et al. 2004; Gregory, Yan et al. 2004; Han, Lee et al. 2004; Landthaler, Yalcin et al. 2004). Together with DGCR8 (or Pasha), Drosha forms a large complex known as the Microprocessor complex.

Following nuclear processing, pre-miRNAs are exported to the cytoplasm where they are cleaved near the terminal loop by Dicer, releasing approximately 22 nucleotide long miRNA duplexes (Hutvagner, McLachlan et al. 2001). Thus the first end of the pre-miRNA is determined by Drosha through hairpin processing while the other end of the mature miRNAs is determined by Dicer.

Following Dicer cleavage, the resulting approximately 22nt long RNA duplex is loaded onto an Ago protein and generates the effector complex, RISC. One strand of the ~22nt RNA duplex remains in Ago as a mature miRNA called the guide strand, whereas the other strand called the passenger strand is usually degraded. Studies on miRNA precursors suggest that it is common for the strand with relatively unstable base pairs at the 5′ end to typically survive and get incorporated to the Ago protein (Khvorova, Reynolds et al. 2003). Because strand selection is often not a stringent process, some hairpins produce miRNAs from both strands at similar frequencies.

Gene regulation by miRNAs is mediated through the association of the miRNA-loaded RISC to complementary sequences in target mRNAs (Figure 2). It has been postulated that miRNAs can cause the premature removal of translating ribosomes from target mRNA (Petersen, Bordeleau et al. 2006). However, the absence of truncated protein products from targeted mRNAs has resulted in speculation that miRNAs primarily direct degradation of target mRNAs (Eulalio, Huntzinger et al. 2008; Guo, Ingolia et al. 2010). Other studies in mammalian systems have co-sedimented miRNAs with target mRNAs not associated with ribosomes (Pillai 2005), indicating that Ago2 may mediate miRNA translation repression by blocking access of translation initiating factors to the 5’ cap of target mRNA.

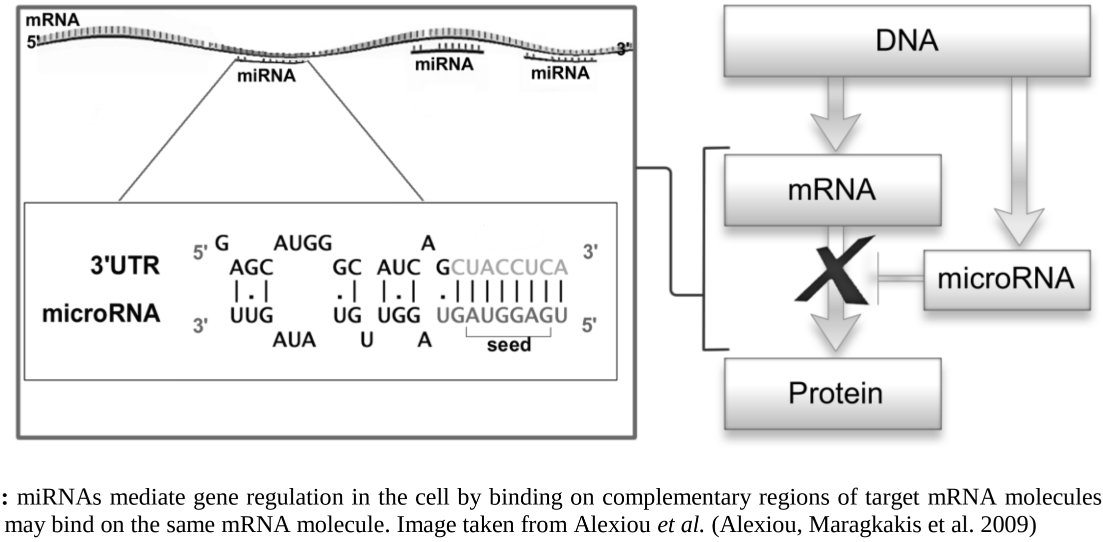


Figure 2: miRNAs mediate gene regulation in the cell by binding on complementary regions of target mRNA molecules. Multiple miRNAs may bind on the same mRNA molecule. Image taken from Alexiou *et al.* (Alexiou, Maragkakis et al. 2009)

Arguably, the most important role of miRNAs in the living cell is targeting of mRNA molecules and several factors have been postulated to have an effect on this mechanism. Even though experimental validation of miRNA targets has been progressing in bounds in the past few years, a large portion of targeted genes still remains unverified. For this, deciphering miRNA function goes hand in hand with the development of computational target prediction programs. Since the initial discovery of miRNAs several algorithms for miRNA target prediction have been developed. However all of them lack both in terms of specificity and sensitivity (Kiriakidou, Nelson et al. 2004; Selbach, Schwanhausser et al. 2008). Most target prediction algorithms to date are biased towards heuristic assumptions regarding miRNA targeting such as the existence of a seed (nucleotides 2-7 of the miRNA) match between the miRNA and the target gene. However, recently it has been shown that such heuristics although informative are not always required and other types of binding are also possible (Shin, Nam et al. 2010).

In order to increase specificity, most target prediction programs use several features such as evolutionary conservation, structural accessibility (Kertesz, Iovino et al. 2007), nucleotide composition (Grimson, Farh et al. 2007) and localization within the gene (Gaidatzis, van Nimwegen et al. 2007; Grimson, Farh et al. 2007). The limitation regarding most of these features is that they have been heuristically defined based on only a few data and therefore the contained information is not always fully assessed.

In addition, traditionally, predictions for miRNA binding sites have been limited to the 3'UTR of mRNAs with only few exceptions which have nevertheless been shown (Kiriakidou, Nelson et al. 2004; Selbach, Schwanhausser et al. 2008) to perform poorly on experimental data. Interestingly, the advent of high throughput sequencing data (Chi, Zang et al. 2009; Hafner, Landthaler et al. 2010) has revealed that miRNAs tend to bind in approximately equal proportion on the 3'UTR and the coding sequence (CDS) of target mRNAs. Hafner et al. using microarrays suggested that miRNA targeting in the CDS usually infers little but measurable effect on miRNA mediated mRNA degradation. Also, analysis on the proteomics data of Selbach et al (Selbach, Schwanhausser et al. 2008) reveals that approximately half of the targeted genes, following miRNA transfection, carry not a single corresponding miRNA seed match on their 3'UTR sequence indicating that alternative targeting mechanisms may be at play.

Since miRNA biology constitutes a newly evolved scientific field, the related information has been expanding in an increasing way in the last years. For this, the development of tools that provide primary data is not on its own sufficient and there is need for applications, primarily web based, which will organize the available information and serve as an interface between bioinformatics tools and researchers. These applications need to be able to optimally organize the available data and present it in an intuitive and integrated way.

4. RESEARCH OBJECTIVES

In this thesis there are two primary objectives which have been addressed. The first objective is the development of advanced miRNA target prediction programs capable of identifying the underlying nature of miRNA targeting and the second is the development of integrated tools which will reveal aspects regarding miRNA function. Both of these objectives aim in understanding the role of miRNAs in the cell and deciphering their function. The overall perspective is to provide researchers with integrated tools which they may use to design and analyze biological experiments.

The former objective has several important aspects that need to be addressed starting from the design of the algorithm and ending to the evaluation of the prediction performance. The first step is to search and collect available data which confer information for miRNA targeting. Such a dataset might be the one of experimentally verified miRNA targets stored in TarBase (Sethupathy, Corda et al. 2006; Papadopoulos, Reczko et al. 2009). The next step would be the analysis of the available datasets and the decision on which of these may be used for developing the algorithm and which may be used for evaluating its performance. Next, the design of the algorithm needs to be addressed and a decision needs to be made on whether it would be a pure data driven approach or it would implement heuristic parameters introduced through the analysis of the biological data. Finally, the most important step is the evaluation of the prediction performance on several different datasets to verify that the algorithm can generalize on different kinds of datasets.

The later objective also includes several steps that need to be addressed. One of the most important is the need to integrate available information from several primary resources such as for example a target prediction program, a gene expression experiment or a database for gene annotation. Specifically, the tools which have been developed as part of this thesis and which are described in the following have been designed to offer high interconnection with each other and with external databases and resources. Also one of the goals that has been set and which is usually set aside by many bioinformatics applications is the ease in using the tools by an average researcher. For this, all of the tools have been implemented as Web applications and are supported by an intuitive graphical interface.

5. microRNA TARGET PREDICTION

Arguably, the most important role of miRNAs in the cell is the regulation of gene expression through post transcriptional targeting of mRNA molecules. Several factors have been postulated to confer information regarding this mechanism such as evolutionary conservation, structural accessibility (Kertesz, Iovino et al. 2007), nucleotide composition (Grimson, Farh et al. 2007) and localization within the gene (Gaidatzis, van Nimwegen et al. 2007; Grimson, Farh et al. 2007). In this chapter I describe all my related published work regarding miRNA target prediction. In chronological order, the work consists of a major release of the microT program called microT-v3.0, a novel probabilistic approach for miRNA binding site identification called CoProHMM, a combined computational and experimental approach for the identification of viral miRNA targets and finally another major release of the microT program denoted as microT-CDS which addresses the identification of predicted miRNA targets within the coding sequence of genes.

5.1. Accurate miRNA target prediction correlates with protein repression levels.

In the following publication we describe the program denoted as microT-v3.0. In short it is a program that uses parameters which are calculated individually for each miRNA and that computes a total score for predicted miRNA:target gene interactions as the weighted sum of scores for evolutionarily conserved and non-conserved binding sites. Primarily, the program assesses whether, in terms of evolutionary conservation, a predicted miRNA binding site can be distinguished from a random background or not. Also it is shown that the prediction score of the program correlates with protein repression levels. The prediction performance of the program has been evaluated independently in a work published in Nature by Selbach *et al* (Selbach, Schwanhausser et al. 2008) and has been shown to be the most precise program available. The work was published in Maragkakis *et al* (Maragkakis, Alexiou et al. 2009).

BMC Bioinformatics

BioMed Central

Research article

Open Access

Accurate microRNA target prediction correlates with protein repression levels

Manolis Maragkakis*†[1,2], Panagiotis Alexiou†[1,3], Giorgio L Papadopoulos[1], Martin Reczko[1,4], Theodore Dalamagas[5], George Giannopoulos[5,6], George Goumas[7], Evangelos Koukis[7], Kornilios Kourtis[7], Victor A Simossis[1], Praveen Sethupathy[8], Thanasis Vergoulis[5,6], Nectarios Koziris[7], Timos Sellis[5,6], Panagiotis Tsanakas[7] and Artemis G Hatzigeorgiou*[1,9]

Address: [1]Institute of Molecular Oncology, Biomedical Sciences Research Center 'Alexander Fleming', Vari, Greece, [2]Institute of Computer Science, Martin Luther University Halle-Wittenberg, 06120 Halle, Germany, [3]School of Biology, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece, [4]Synaptic Ltd., Heraklion, Greece, [5]Institute for the Management of Information Systems, "Athena" Research Center, Athens, Greece, [6]Knowledge and Database Systems Lab, Department of Computer Science, School of Electrical and Computer Engineering, National Technical University of Athens, Greece, [7]Computing Systems Laboratory, Department of Computer Science, School of Electrical and Computer Engineering, National Technical University of Athens, Greece, [8]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, 20876, USA and [9]Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA, USA

Email: Manolis Maragkakis* - maragkakis@fleming.gr; Panagiotis Alexiou - pan.alexiou@fleming.gr; Giorgio L Papadopoulos - papadopoulos@fleming.gr; Martin Reczko - reczko@fleming.gr; Theodore Dalamagas - dalamag@imis.athena-innovation.gr; George Giannopoulos - giann@dblab.ece.ntua.gr; George Goumas - goumas@cslab.ece.ntua.gr; Evangelos Koukis - vkoukis@cslab.ece.ntua.gr; Kornilios Kourtis - kkourt@cslab.ece.ntua.gr; Victor A Simossis - simossis@fleming.gr; Praveen Sethupathy - sethupathyp@mail.nih.gov; Thanasis Vergoulis - bergoulis@dblab.ece.ntua.gr; Nectarios Koziris - nkoziris@cslab.ece.ntua.gr; Timos Sellis - timos@imis.athena-innovation.gr; Panagiotis Tsanakas - tsanakas@admin.grnet.gr; Artemis G Hatzigeorgiou* - artemis@fleming.gr

* Corresponding authors †Equal contributors

Published: 18 September 2009

BMC Bioinformatics 2009, **10**:295 doi:10.1186/1471-2105-10-295

Received: 15 April 2009
Accepted: 18 September 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/295

Abstract

Background: MicroRNAs are small endogenously expressed non-coding RNA molecules that regulate target gene expression through translation repression or messenger RNA degradation. MicroRNA regulation is performed through pairing of the microRNA to sites in the messenger RNA of protein coding genes. Since experimental identification of miRNA target genes poses difficulties, computational microRNA target prediction is one of the key means in deciphering the role of microRNAs in development and disease.

Results: DIANA-microT 3.0 is an algorithm for microRNA target prediction which is based on several parameters calculated individually for each microRNA and combines conserved and non-conserved microRNA recognition elements into a final prediction score, which correlates with protein production fold change. Specifically, for each predicted interaction the program reports a signal to noise ratio and a precision score which can be used as an indication of the false positive rate of the prediction.

Conclusion: Recently, several computational target prediction programs were benchmarked based on a set of microRNA target genes identified by the pSILAC method. In this assessment DIANA-microT 3.0 was found to achieve the highest precision among the most widely used microRNA target prediction programs reaching approximately 66%. The DIANA-microT 3.0 prediction results are available online in a user friendly web server at http://www.microrna.gr/microT

Background

MicroRNAs (miRNAs) are short, endogenously expressed RNA molecules that regulate gene expression by binding directly and preferably to the 3' untranslated region (3'UTR) of protein coding genes [1]. Each miRNA is 19-24 nucleotides in length and is processed from a longer transcript which is referred to as the primary transcript (pri-miRNA). These transcripts are processed in the cell nucleus to short, 70-nucleotide stem-loop structures known as pre-miRNAs. Pre-miRNAs are processed to mature miRNAs in the cytoplasm by interaction with the endonuclease Dicer which cleaves the pre-miRNA stem-loop into two complementary short RNA molecules. One of these molecules is integrated into the RISC (RNA induced silencing complex) complex and guides the whole complex to the mRNA, thus inhibiting translation or inducing mRNA degradation [2]. Since their initial identification, miRNAs have been found to confer a novel layer of genetic regulation in a wide range of biological processes. miRNAs were first identified in 1993 [3] via classical genetic techniques in C. elegans, but it was not until 2001 that they were found to be widespread and abundant in cells [4-6]. This finding served as the primary impetus for the development of the first computational miRNA target prediction programs. DIANA-microT [7] and TargetScan [8] were the first algorithms to predict miRNA targets in humans, and led to the identification of an initial set of experimentally supported mammalian targets. Such targets are now collected and reported in TarBase [9] which contains more than one thousand entries for human and mouse miRNAs.

In the last years several groups suggested that the first nucleotides of a miRNA sequence are crucial for recognizing and binding to the messenger of a protein. Kiriakidou *et al.* [7] showed the need for a nearly consecutive binding of the first 9 miRNA nucleotides (*driver* sequence) (figure 1b) to the 3'UTR of protein coding genes in order to repress translation. A statistical approach by Lewis et al. [10] revealed that complementary motifs to nucleotides 2-7 of the miRNA driver sequence (miRNA *seed* region) remain preferentially conserved in several species. Typically, it is believed that a binding of at least 7 consecutive Watson-Crick (WC) base pairing nucleotides between the miRNA driver sequence and the miRNA Recognition Element (MRE) is required for sufficient repression of protein production. However, experimental evidence [11] show that weaker bindings, involving only six consecutively paired nucleotides or including imperfect bindings (e.g. G:U wobble, bulge) may also confer protein repression although they might generally be less effective [12]. For this reason, miRNA target prediction programs mostly rely on sequence alignment of the miRNA seed region to the 3'UTR sequences of candidate target genes in order to identify putative miRNA binding sites. Their specificity is usually increased by additionally assessing the commonly observed binding site evolutionary conservation or by using additional features such as binding site structural accessibility [13,14], nucleotide composition flanking the binding sites [15] or proximity of one binding site to another within the same 3' UTR [12,15,16].

DIANA-microT 3.0, the algorithm described here, utilizes the above mentioned features and categorizes as putative MREs those sites that have seven, eight or nine nucleotide long consecutive WC base pairing with the miRNA driver sequence, starting from position 1 or 2 of the 5'end of the miRNA. For sites with additional base pairing involving the 3'end of the miRNA, a single G:U wobble pair or binding of only 6 consecutive nucleotides to the driver sequence are allowed. Briefly, the DIANA-microT 3.0 algorithm consists of (figure 1a): a) alignment of the miRNA driver sequence on the 3'UTR of a protein coding gene, b) identification of putative MREs based on specific binding rules, c) scoring of individual MREs according to their binding type and conservation profile, d) calculation of an overall miRNA target gene (miTG) score through the weighted sum of all MRE scores lying on the 3'UTR. The program is designed to use up to 27 different species to estimate MRE conservation scores and combines both conserved and non-conserved MREs in a final miTG score (figure 1c). The miTG score correlates with fold changes in protein expression. Additionally, since the algorithm calculates all weights and scores independently for each miRNA it allows for the calculation of signal to noise ratio (SNR) at different miTG score cut-offs providing precision scores which serve as an indication of the false positive rate of the predicted interactions.

Generally, miRNAs can repress the expression of proteins in two ways: via mRNA degradation or via repression of mRNA translation. Until recently, high throughput experiments were only able to measure miRNA-mediated changes at the mRNA level (degradation), allowing the characterization of only a subset of direct miRNA targets [17,18]. However, recently two groups [12,19] have independently developed methods to characterize miRNA-mediated gene expression changes at both the mRNA and the protein level. Selbach *et al.* [19] used microarrays and pulsed stable isotope labeling with amino acids in cell culture (pSILAC) assays to determine the genes targeted by each of five over-expressed miRNAs in HeLa cells. Using this set of experimentally supported targets the authors performed a comparative assessment of several target prediction programs. The benchmark revealed that the simplest prediction method involving the search for complementary sequences of the miRNA seed region on the 3'UTR of genes achieved a precision (the fraction of

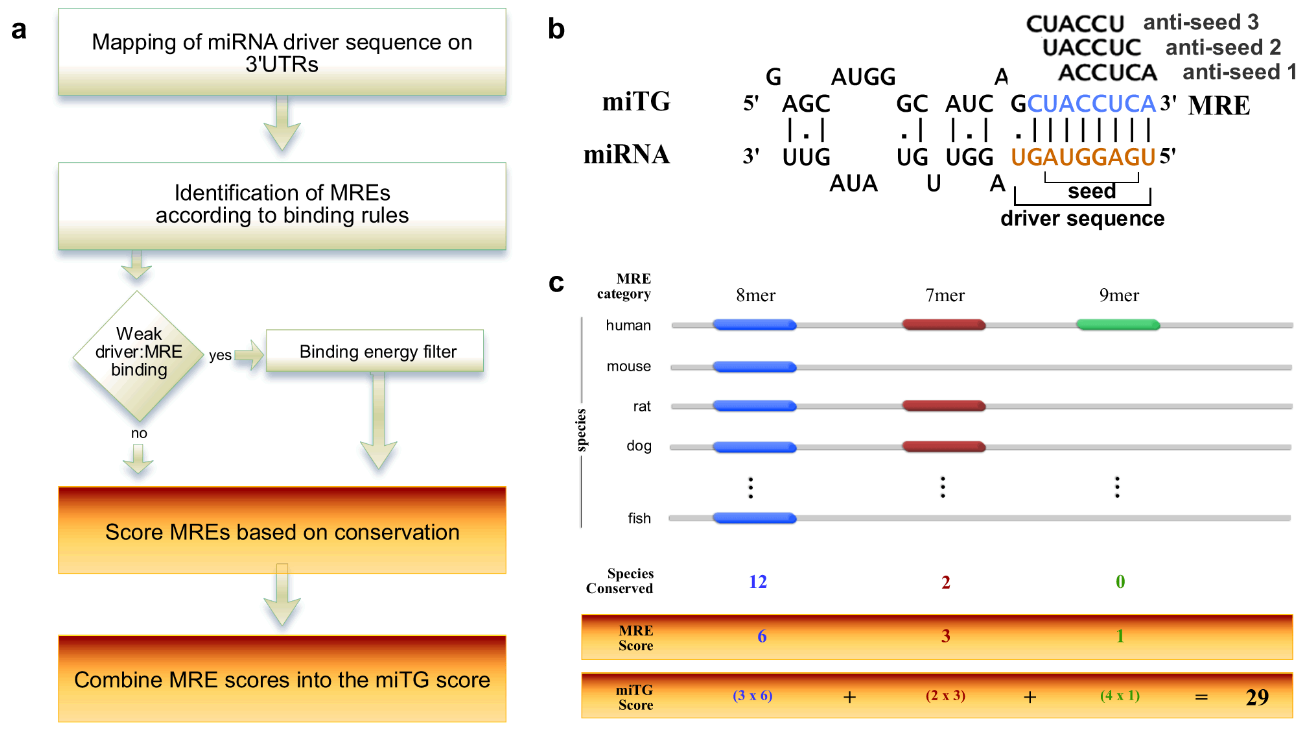

a
Mapping of miRNA driver sequence on 3'UTRs
Identification of MREs according to binding rules
Weak driver:MRE binding
yes
no
Binding energy filter
Score MREs based on conservation
Combine MRE scores into the miTG score
b
CUACCU anti-seed 3
UACCUC anti-seed 2
ACCUCA anti-seed 1
miTG
miRNA
5' AGC GC AUC GCUACCUCA 3' MRE
3' UUG UG UGG UGAUGGAGU 5'
G AUGG A
AUA U A
seed
driver sequence
c
MRE category
8mer
7mer
9mer
species
human
mouse
rat
dog
fish
Species Conserved
12
2
0
MRE Score
6
3
1
miTG Score
(3 x 6) + (2 x 3) + (4 x 1) = 29

Figure 1

The DIANA-microT 3.0 algorithm. (a) A schematic overview of the algorithm. The miRNA driver sequence is mapped onto a 9 nt length window that slides along the 3'UTR sequence. The binding category of the driver:MRE interaction is defined by the number of binding nucleotides between the two sequences. G:U wobble pairs or less than 7 consecutive WC matches are only allowed if the free binding energy of the miRNA:MRE heteroduplex is under a binding category specific threshold (lower free binding energy corresponds to stronger binding). MREs are scored according to their binding category and degree of conservation in other species. The final miTG score is the weighted sum of all MREs on the miTG. (b) The top sequence (MRE) is part of the 3'UTR of a gene. The nine nucleotide region near the 5'end of the miRNA is called the driver sequence of the miRNA (shown in red). Sequences on the MRE, corresponding to positions 1-6, 2-7 and 3-8 from the miRNA 5'end are called anti-seed 1, anti-seed 2 and anti-seed 3 respectively. (c) An example of the miTG score calculation. The top line represents the 3'UTR sequence of a human gene containing three MREs with different conservation levels. Individual MRE scores are calculated depending on the degree of conservation of the MRE, and multiplied by a weight depending on the MRE binding category. The sum of all weighted MRE scores defines the final miTG score.

the predicted targets that were actually downregulated) of 44% while only three of the prediction programs (including an initial version of DIANA-microT 3.0) achieved significantly higher precision. PicTar [20] and TargetScanS [10] achieved approximately 62% precision compared to DIANA-microT 3.0 with approximately 66%.

Methods

Identification of putative miRNA binding sites through sequence alignment

The program identifies the highest scoring alignment between every nine nucleotide long window of the 3'UTR with the miRNA driver sequence using a dynamic programming algorithm. The alignment is based on the following binding rules. Firstly, a minimum of six consecutive matches (Watson-Crick (W-C) or G:U) is required. If the six matches are W-C and the binding starts at position 1 or 2 of the miRNA driver sequence, then the MRE is considered a 6mer. A 7mer (8mer, 9mer) has seven (eight, nine) consecutive W-C matches starting at position 1 or 2 of the miRNA driver. A single G:U wobble pair is allowed as long as there are at least six W-C pairs, yielding 7mers, 8mers and 9mers, each with a wobble base pair.

Filter of putative miRNA binding sites depending on binding energy

For sites with less than 7 consecutive W-C matches (6mer, 7mer with wobble, 8mer with wobble, 9mer with wobble) an additional energy filter is applied. Using RNAhybrid [21] the algorithm estimates the free binding energy between the miRNA sequence and the 3'UTR sequence flanking the identified putative binding site and compares it to the perfect complement energy of the miRNA. As "perfect complement energy" we denote the hypothetical energy of the perfect binding between the miRNA sequence and its reverse complement sequence. Therefore an imperfect site, in terms of alignment, is considered as MRE only if the ratio of the free binding energy to the perfect complement energy is higher than a binding-category specific threshold. A threshold of 0.6 is used for 9mers and 8mers containing a G:U wobble pair, and a threshold of 0.74 is used for 7mers with a G:U wobble pair and 6mers. The energy thresholds have been calculated by comparing the predicted binding sites of the real miRNA sequence versus the predicted binding sites of several shuffled miRNA sequences. The shuffled miRNA sequences are designed to have the same driver as the real miRNA but a shuffled 3' end with the same nucleotide composition as the real miRNA. The free binding energy ratio e_i is defined as the ratio of the free binding energy between the miRNA sequence and the 3'UTR sequence flanking the identified putative binding site over the miRNA perfect complement energy. Additionally, $N_r(e_i)$ is defined as the number of binding sites of the real miRNAs that have energy ratios greater than e_i and as $N_S(e_i)$ the number of binding sites of the shuffled miRNAs that have energy ratios greater than e_i. The ratio $R(e_i) = N_r(e_i)/N_S(e_i)$ indicates how much more prevalent the free binding energy $e_i$for real binding sites compared to the shuffled ones is. An example of the way this ratio $R(e_i)$ fluctuates is provided in figure 2. For each binding category the energy thresholds have been chosen at the point where the ratio $R(e_i)$ becomes greater than 2 indicating that at this energy value one can generally find two times more real binding sites than random binding sites.

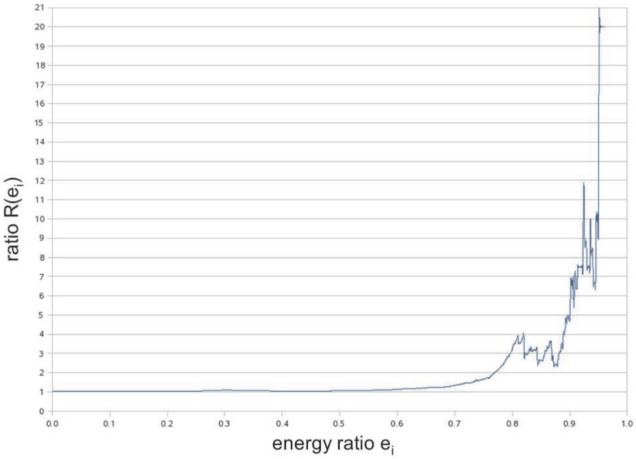


Figure 2
Hybridization energy ratio. Ratio $R(e_i)$ (vertical axis) is plotted against the energy ratio e_i (horizontal axis). The curve corresponds to the binding category which consists of seven WC pairs and a single G:U wobble pair.

Mock miRNAs

Mock miRNAs are artificially produced miRNA sequences which are independently created for each real miRNA. These artificial miRNA sequences are designed to have approximately the same number of predicted MREs as the corresponding real miRNA and are generated through the following procedure. Initially, all 3'UTR sequences are scanned for sites perfectly complementary to each possible 6 nucleotide long motif (hexamer) excluding those motifs corresponding to positions 1-6, 2-7 and 3-8 of real miRNAs. The 60 hexamers having the closest number of complementary sites to those of the seed of the real miRNA are chosen. These hexamers are then used as the seed of each artificially created mock miRNA. The remaining sequence of the mock miRNAs is then produced by randomly shuffling the remaining nucleotides of the real miRNA.

miRNA Recognition Elements score (MRE score)

The identified MREs are checked for sequence conservation in several species based on the sequence alignment of ortholog UTRs. An MRE X is considered conserved in species A if X can also be identified at the exact same position on the ortholog 3'UTR sequence of species A. The conservation score $c$of an MRE is defined as the number of species in which the MRE is conserved. The MRE score is calculated individually for each real miRNA r, each binding category b and each conservation score c. Analytically, for each binding category the number of MREs $N_{r,\,b}(c)$ of the real miRNA and the number of MREs $M_{r,\,m,\,b}(c)$ of the corresponding mock miRNAs with conservation score equal or greater than c are counted and the ratio of the two defines the MRE score (of binding category b at conservation score c). The equation defining this procedure is $R_{r,b}(c) = 60 \cdot N_{r,b}(c) / \sum_{m=1}^{60} M_{r,m,b}(c)$ in which r is the index of the real miRNA, b corresponds to the binding category, c defines the conservation score and m defines the index of the mock miRNA from the set of mock miRNAs corresponding to the real miRNA r. In the described procedure the ratio is kept constant if $N_{rb}(c)$ or $M_{r,\,m,\,b}(c)/60$ become less than 20. Figure 3 shows an example of R_{rb} for 2 binding categories at different MRE conservation scores.

miRNA target gene score

The scores of the MREs identified on the same 3'UTR are combined through a weighted sum to produce the final

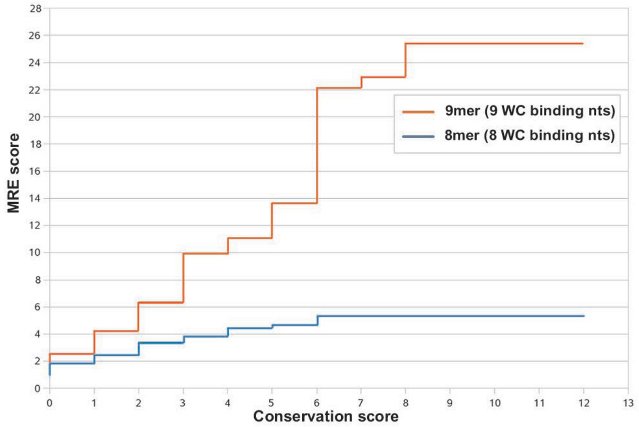


Figure 3
miRNA recognition element score. The MRE score (vertical axis) is plotted against the MRE conservation score (horizontal axis) for two different binding categories.

miTG score. The weights w_b for each binding category b are calculated using 75 miRNAs conserved in human, chimpanzee, mouse, rat, dog and chicken, by comparing them to 375 mock sequences (5 mock miRNAs for each miRNA). The analysis is similar to the calculation of the MRE score explained previously but in this case the 75 miRNAs are not treated independently but as a total. The ratio $R'_b(c)$ for binding category b and conservation score c is calculated as $R'_b(c) = 5 \cdot \sum_{r=0}^{r=75} N_{r,b}(c) / \sum_{r=1}^{75} \sum_{m=1}^{5} M_{r,m,b}(c)$ where $N_{rb}(c)$ is the number of MREs of the r real miRNA categorized to binding category b and having a conservation score greater than c, $M_{r,m,b}(c)$ represents the number of MREs of the m mock miRNA categorized to binding category b succeeding a conservation score greater than c and corresponding to real miRNA r. As shown in figure 4 the weights for each binding category are estimated based on the slope of a fitted line. Fitting is performed based on linear least squares approximation. For each binding category the weight is defined as $w(bindingcategory) = slope(bindingcategory)/slope(9\text{mer})$. For example, the weight for category "8mer" would be $w_{8\text{mer}} = 0.31/0.39 = 0.79$. Except for "9mer", "8mer" and "7mer" the remaining categories do not differ significantly from the mock background and consequently in this analysis no specific weights are calculated for these categories. In order to approximate the estimated weights Dw_b based on the above analysis, each MRE score is multiplied by a specific weight mw_b which depends on the binding category of the MRE (table 1).

miTG score threshold assessment

A common challenge among miRNA target prediction programs is the decision on a score threshold that will reduce the number of misclassifications. Here a set of 100 experimentally supported targets for 43 different human miRNAs, provided by TarBase 5.0 [9], has been used in order to determine a biologically meaningful score threshold. Based on this dataset, an analysis was performed to test the capability of the algorithm to identify supported targets when increasing the miTG score threshold. As expected, the algorithm's capability reduces as the miTG score increases (figure 5). However, there are two distinct miTG scores (7.3 and 19.0) with significantly higher performance reduction. For this reason, these miTG score values have been chosen as a loose and strict miTG score threshold respectively. However, users are still allowed to adjust the threshold at will to exchange between specificity and sensitivity levels.

Precision

The precision of a prediction is defined as the ratio of correct positive predictions over all positive predictions [$precision = truepositive / (truepositive + falsepositive)$]. In the case of DIANA-microT 3.0, the average number of miTGs for mock miRNAs provides an estimation of the number of false positive targets predicted. Therefore, the number of

Table 1: Binding category weights

Category	Estimated Weights (w_b)	Multiplication weights (mw_b)	Overall Diana weights $Dw_b = mw_b/mw_{9mer}$
9mer	1	4	1.00 = 4/4
8mer	0.79	3	0.75 = 3/4
7mer	0.41	2	0.50 = 2/4
other	-	1	0.25 = 1/4

The binding weights estimated for each binding category and the weights used in DIANA-microT 3.0.

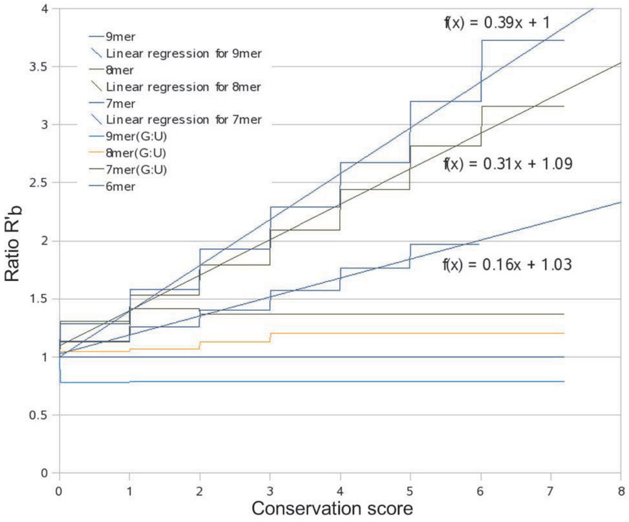


Figure 4

Binding categories differ from the mock background. Ratio R'_b (vertical axis) versus the conservation score (horizontal axis) for the set consisting of 75 miRNAs conserved in human, chimp, mouse, rat, dog, chicken. This diagram indicates how each binding category may be differentiated as the conservation score increases (more conserved MREs). It may be seen that 9mers tend to differentiate more than 8mers and 8mers more than 7mers. Except for categories "9mer", "8mer" and "7mer" the remaining categories do not seem to differ significantly from the background.

true positive predicted miTGs can be calculated by subtracting the average number of predicted miTGs for the mock miRNAs from the total number of predicted miTGs for the real miRNA. In detail, the precision for miRNA *r* at miTG score *s* is calculated by $precision_r(s) = \left[W_r(s) - \bar{W}_{r,m}(s)\right] / W_r(s)$ where W_r is the number of miTGs of the *r* real miRNA having miTG scores from *s* to *s* + Δ*s*, $\bar{W}_{r,m}$ is the average number of miTGs of the mock miRNAs corresponding to miRNA *r* having miTG scores from *s* to *s* + Δ*s* and Δ*s* is a specified miTG score window (Δ*s* = 3).

miRNA sequences

The human and mouse miRNA sequences used by DIANA-microT 3.0 have been downloaded from miRBase Build 10.0 [22].

3'UTR sequences

The gene 3'UTR sequences have been downloaded from Ensembl, release 48 [23]. Those 3'UTR sequences that correspond to the same gene but to different gene transcripts have been filtered to keep only the longest 3'UTR sequence.

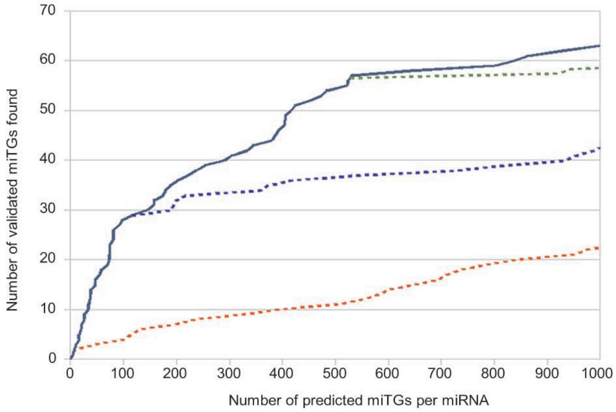


Figure 5

Define biologically meaningful score threshold. Experimentally validated targets correctly predicted by DIANA-microT 3.0 versus the average number of predicted miTGs per miRNA. The slope of this curve corresponds to the rate in which correct validated targets are discovered as more miTGs are predicted. There are two distinct points in which the slope changes. These points correspond to miTG score values of 19 and 7.3 which are proposed as the strict and loose miTG score thresholds respectively. As a control, the order of miTGs with scores lower than each threshold was shuffled. The discovery rate of these controls is shown with dotted lines. The red line shows all miTGs in random order, the blue line those with miTG score under 19 and the green line those with miTG score under 7.3. The difference in slope between the solid line and each dotted line shows the improvement on the discovery rate achieved by the DIANA-microT scoring scheme. Two other target prediction programs (Pictar and TargetScan 4.2) have been compared to DIANA-microT 3.0 on the same dataset achieving similar precision levels (figure 9).

Multiple Alignment Files (MAFs)

The multiple genome alignment files have been downloaded from the UCSC Genome Browser [24]. The file used for human (hg18) is the alignment to 16 vertebrate genomes while for mouse (mm9) 29 vertebrate genomes are used.

Results

Signal to Noise Ratio (SNR) assessment

The signal to noise ratio for a prediction algorithm is typically used for the evaluation of its specificity. For DIANA-microT 3.0 the overall SNR is defined as the average signal to noise ratio calculated individually for each miRNA. The individual miRNA signal to noise ratio calculation is performed by dividing the number of predicted miTGs of a real miRNA by the number of predicted miTGs for the set of corresponding mock miRNAs. It is assumed that the predicted miTGs for the mock miRNA sequences provide an unbiased estimate of the number of miTGs predicted by chance alone. Analytically, the SNR value of miRNA *r* at miTG score *s* is calculated as $SNR_r(s) = 60 \cdot NG_r(s) / \sum_{1}^{60} MG_{r,m}(s)$. In this formula $NG_r(s)$ refers to the number of miTGs of the real miRNA *r* having miTG scores greater than *s* while $MG_{r,m}(s)$ refers to the number of miTGs of the mock miRNA *m* corresponding to the real miRNA *r* having miTG score greater than *s*. Figure 6 presents a graph of the *SNR* for seven different miRNAs. The overall SNR calculation for DIANA-microT 3.0 is performed on two different sets of miRNAs. The first set consists of 75 miRNAs conserved in 6 vertebrate species while the second set consists of 227 unique miRNAs each one representing a miRNA family with varying conservation levels. Figure 7 shows the diagram for the number of predicted miTGs versus the miTG score. For an miTG score threshold that yields an average of approximately 100 predicted target genes per miRNA, DIANA-microT 3.0 achieves an overall SNR of 3.9 for the first dataset and an overall SNR of 2.2 for the second dataset which indicates that conserved miRNAs tend to achieve higher SNR values.

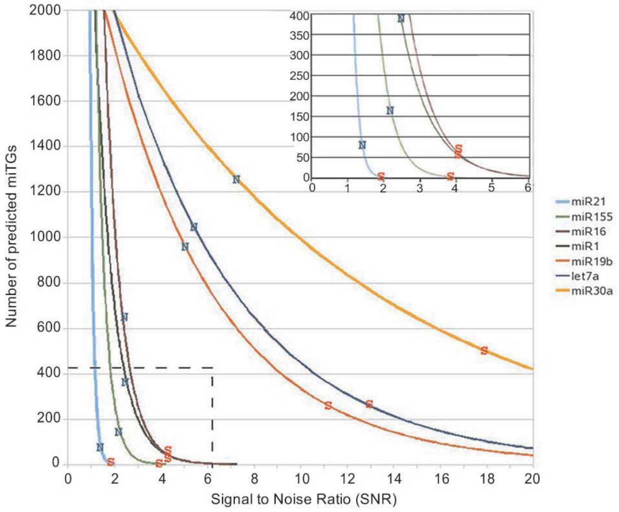


Figure 6
Signal to noise ratio for 7 miRNAs. Curves showing the number of predicted miTGs versus the SNR for 7 miRNAs. The loose and strict thresholds have been marked in the figure with the symbols "N" and "S" respectively.

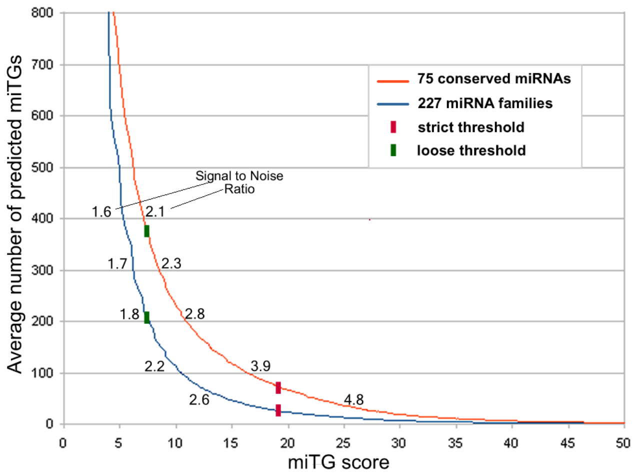


Figure 7
Overall signal to noise ratio. The mean number of predicted miTGs per miRNA for different miTG score cutoffs. The red curve corresponds to a set of 75 miRNAs conserved in at least six species (human, chimp, mouse, rat, dog, chicken), whereas the blue curve corresponds to a set of 227 miRNAs which represent the miRNA families (with varying conservation levels). The values next to the curves indicate the overall SNR. Higher miTG score leads to fewer predicted miTGs with higher overall SNR, which suggests a lower number of false positive predicted miTGs. The suggested strict (red bars) and loose (green bar) miTG score thresholds are marked on the curves. For the strict miTG score threshold (miTG score = 19), the estimated overall SNR for the set of 227 miRNAs (blue line) is 3, meaning that approximately one in three predicted miTGs might be a false positive. In comparison, at the loose suggested threshold (miTG score = 7.3), approximately one in two predicted miTGs might be a false positive.

Receiver Operating Characteristics (ROC) analysis on proteomics data

Until recently a common difficulty in assessing the performance of a prediction algorithm was that the available experimental data could not easily distinguish between true and false targets. However, the recent study of Selbach *et al.* provides both classes of targets allowing for the estimation of both the true positive rate as well as the false positive rate of a prediction. Using a $\log_2$ fold change cutoff of -0.2 to distinguish between targeted and non-targeted genes, the performance of DIANA-microT 3.0 is assessed and presented as a ROC curve (figure 8).

Correlation of miTG score to the repression of protein production

In the study by Selbach *et al*[19], it was observed that there is a correlation between the $\log_2$-fold change of protein production with the number of occurrences of the hexamer corresponding to the seed of a miRNA in the 3'UTR

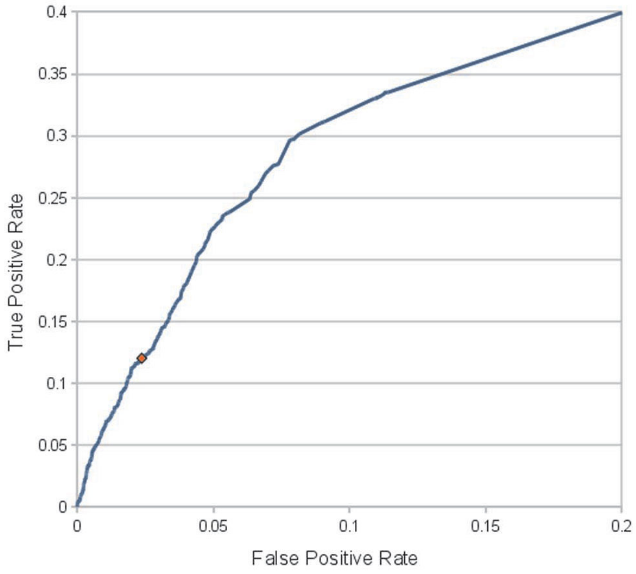


Figure 8
DIANA-microT 3.0 ROC curve. The ROC curve for DIANA-microT 3.0 calculated on the pSILAC data [19]. The suggested loose threshold of DIANA-microT 3.0 has been marked on the diagram with a red dot.

of downregulated genes. When investigating the same data using DIANA-microT 3.0, a similar correlation between the level of protein down-regulation and the predicted miTG scores, SNR, and precision values is observed (figure 9a). Interestingly, a linear regression analysis shows that the combination of miTG score, precision, SNR, and the number of anti-seeds (regions on the gene 3'UTR complementary to the motifs 1-6, 2-7, 3-8 of the miRNA) as regressors provides the best accuracy in attempting to predict such fold changes in protein expression. Figure 9b demonstrates the relationship between the protein expression fold change versus the number of occurrences of the miRNA anti-seed 2 (adjusted $R^2 = 0.12$) as well as the protein expression fold change versus the combined regressor (adjusted $R^2 = 0.15$).

Discussion and conclusion

In the last five years more than two dozen miRNA target prediction programs for mammalian genomes have been published [25]. Using data from a high throughput experiment on five miRNAs [19] as a true-positive set of targets, it has been shown that DIANA-microT 3.0 achieves comparable precision to two other leading target prediction programs, TargetScanS [8] and PicTar [20]. Additionally, DIANA-microT 3.0 provides prediction scores which cor-

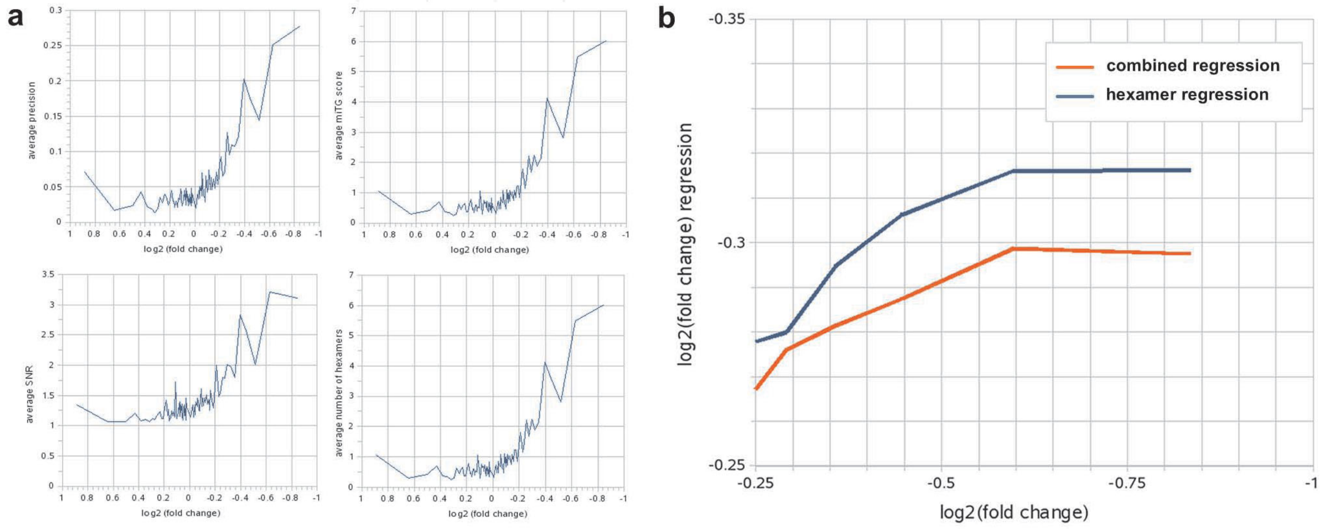


Figure 9
Correlation of DIANA-microT 3.0 prediction measures to protein repression. Fold changes are calculated for approximately 5,000 proteins after overexpression of a miRNA. The results for five miRNAs, as provided by Selbach et al., are used. The fold change and the miTG score is averaged in groups of 150 proteins sorted by fold change. (a) The correlation of several miRNA target prediction measures with protein production fold change induced by the same miRNAs. It may be observed that there is a trend for values of all the measures to increase as the level of downregulation increases. (b) The red line indicates the correlation between the anti-seed 2 occurrences on the 3'UTRs of downregulated genes with the protein production fold change of the corresponding genes using a linear regression. The blue line shows the corresponding correlation for a linear regressor based on a combination of the miTG score, the precision, the SNR and the anti-seed 2 frequency. The combined linear regressor correlates better with the protein production fold change than the regressor based solely on the anti-seed 2 frequency.

Table 2: Number of miTGs predicted in common by programs

	Diana-microT	PicTar	TargetScan 4.2
Diana-microT	**22391**	8882	10651
PicTar		**17135**	12902
TargetScan 4.2			**19299**

The table diagonal corresponds to the total number of miTGs predicted by each program for all the miRNAs which are included in the set of experimentally verified targets. The number of miTGs predicted in common by each two target prediction programs is shown in the table. For example, TargetScan and PicTar have 12902 predicted targets in common while DIANA-microT and PicTar have 8882.

relate with protein production fold change and may be used as an indication of the expected fold change in protein production. The performance of the algorithm has been analyzed further by using a different set of supported miRNA targets which has been extracted by the database of experimentally supported targets [9]. The results also indicate that the three programs (DIANA-microT 3.0, PicTar and TargetScan 4.2) achieve similar precision levels (figure 10). However, as shown in table 2 and 3 there are significant differences among the miTGs predicted by DIANA-microT 3.0 and those predicted by each of the other programs. Table 3 indicates that only 40% of the miTGs predicted by DIANA-microT 3.0 are also predicted by PicTar, and only 48% are predicted by TargetScan 4.2. This leaves in either case approximately 50% of the targets predicted only by DIANA-microT 3.0.

Recently, the rapid growth in the discovery rate of novel miRNA sequences due to extensive usage of deep sequencing technology [14], and the fact that miRNAs have been shown to undergo A-to-I RNA editing [15] have underlined the need for a web based program which would allow for miRNA target predictions based on user defined miRNA sequences. DIANA-microT 3.0 is one of the few programs offering such a service, supporting the scientific community with a tool which in total can be extensively used for the analysis of miRNA dependent processes. This tool can be accessed thought the DIANA-microT [26] web server at http://www.microrna.gr/microT which includes an optimized prediction algorithm that provides several features, combined with a user friendly interface which assists in the identification of interactions of interest.

Table 3: Percentage of common predictions among programs

	Diana-microT	PicTar	TargetScan 4.2
Diana-microT	**100%**	39.67%	47.57%
PicTar	51.84%	**100%**	75.30%
TargetScan 4.2	55.19%	66.85%	**100%**

The percentage of each program's predicted targets (rows) which are also predicted by another program (columns) for all the miRNAs which are included in the set of experimentally verified targets. For example, from the miTGs predicted by DIANA-microT 3.0, 39.67% are also predicted by PicTar and 47.57% by TargetScan 4.2.

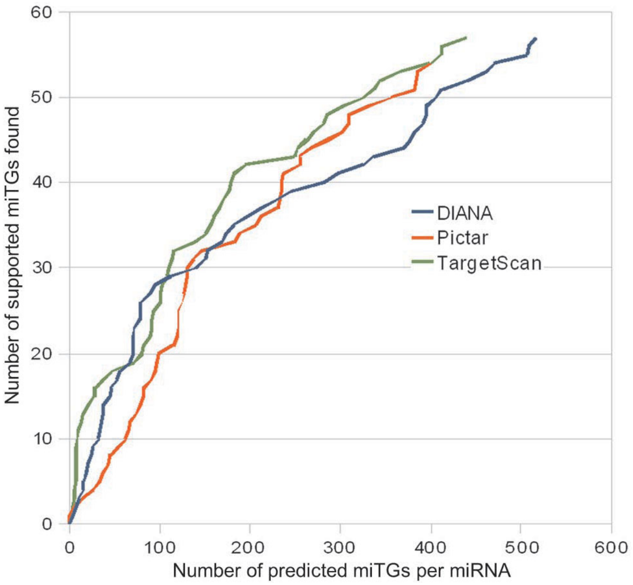


Figure 10
Comparison on experimentally supported targets. Comparison of three target prediction programs (DIANA-microT 3.0, Pictar and TargetScan 4.2) on the experimentally supported dataset. The average number of predicted miTGs per miRNA is presented on the horizontal axis. The total number of correctly predicted experimentally validated targets is shown on the vertical axis. All three programs tested perform similarly.

As already mentioned, DIANA-microT 3.0 takes into account both conserved and not conserved MREs. This attribute provides the algorithm with a highly important capability to predict targets of viral miRNA sequences. Generally, targets of viral miRNAs are not expected to be conserved and this limits the ability of algorithms dependent on conservation to identify them. However, since DIANA-microT 3.0 algorithm accepts non conserved MREs it can successfully cope with viral miRNA sequences.

Authors' contributions

MM and PA designed and developed the algorithm, performed the statistical analysis and drafted the paper. GLP contributed in the algorithm's implementation. MR participated in the algorithm's design and drafted the paper. TD, GG (Giannopoulos G.), TV and TS participated in the design and implementation of the web server database. GG (Goumas G.), EK, KK, NK, PT participated in the implementation of the algorithm's parallelization and contributed in the development of the online execution of the algorithm. VAS contributed in the web server design and development. PS helped to draft the paper and participated in the early development of the algorithm. AGH conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Funding: Aristeia Award from General Secretary Research and Technology, Greece

References

1. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell* 2004, **116(2):**281-297.
2. Liu J, Carmell MA, Rivas FV, Marsden CG, Thomson JM, Song JJ, Hammond SM, Joshua-Tor L, Hannon GJ: **Argonaute2 is the catalytic engine of mammalian RNAi.** *Science* 2004, **305(5689):**1437-1441.
3. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell* 1993, **75(5):**843-854.
4. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T: **Identification of novel genes coding for small expressed RNAs.** *Science* 2001, **294(5543):**853-858.
5. Lau NC, Lim LP, Weinstein EG, Bartel DP: **An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans.** *Science* 2001, **294(5543):**858-862.
6. Lee RC, Ambros V: **An extensive class of small RNAs in Caenorhabditis elegans.** *Science* 2001, **294(5543):**862-864.
7. Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, Hatzigeorgiou A: **A combined computational-experimental approach predicts human microRNA targets.** *Genes Dev* 2004, **18(10):**1165-1178.
8. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB: **Prediction of mammalian microRNA targets.** *Cell* 2003, **115(7):**787-798.
9. Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, Hatzigeorgiou AG: **The database of experimentally supported targets: a functional update of TarBase.** *Nucleic Acids Res* 2009:D155-158.
10. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.** *Cell* 2005, **120(1):**15-20.
11. Brennecke J, Stark A, Russell RB, Cohen SM: **Principles of microRNA-target recognition.** *PLoS Biol* 2005, **3(3):**e85.
12. Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP: **The impact of microRNAs on protein output.** *Nature* 2008, **455(7209):**64-71.
13. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E: **The role of site accessibility in microRNA target recognition.** *Nat Genet* 2007, **39(10):**1278-1284.
14. Long D, Lee R, Williams P, Chan CY, Ambros V, Ding Y: **Potent effect of target structure on microRNA function.** *Nat Struct Mol Biol* 2007, **14(4):**287-294.
15. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP: **MicroRNA targeting specificity in mammals: determinants beyond seed pairing.** *Mol Cell* 2007, **27(1):**91-105.
16. Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M: **Inference of miRNA targets using evolutionary conservation and pathway analysis.** *BMC Bioinformatics* 2007, **8:**69.
17. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM: **Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs.** *Nature* 2005, **433(7027):**769-773.
18. Sood P, Krek A, Zavolan M, Macino G, Rajewsky N: **Cell-type-specific signatures of microRNAs on target mRNA expression.** *Proc Natl Acad Sci USA* 2006, **103(8):**2746-2751.
19. Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N: **Widespread changes in protein synthesis induced by microRNAs.** *Nature* 2008, **455(7209):**58-63.
20. Lall S, Grun D, Krek A, Chen K, Wang YL, Dewey CN, Sood P, Colombo T, Bray N, Macmenamin P, *et al.*: **A genome-wide map of conserved microRNA targets in C. elegans.** *Curr Biol* 2006, **16(5):**460-471.
21. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R: **Fast and effective prediction of microRNA/target duplexes.** *Rna* 2004, **10(10):**1507-1517.
22. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics.** *Nucleic Acids Res* 2008:D154-158.
23. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, *et al.*: **Ensembl 2008.** *Nucleic Acids Res* 2008:D707-714.
24. Karolchik D, Hinrichs AS, Kent WJ: **The UCSC Genome Browser.** *Curr Protoc Bioinformatics* 2007, **Chapter 1(Unit 1):**4.
25. Sethupathy P, Megraw M, Hatzigeorgiou AG: **A guide through present computational approaches for the identification of mammalian microRNA targets.** *Nat Methods* 2006, **3(11):**881-886.
26. Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, *et al.*: **DIANA-microT web server: elucidating microRNA functions through target prediction.** *Nucleic Acids Res* 2009:W273-276.

5.2. Predicting miRNA targets utilizing an extended profile HMM

In the following publication we describe a data driven alignment method for miRNA targeting. One of the most critical steps in a miRNA target prediction program is the alignment of the miRNA sequence against the target mRNA sequence for the identification of putative miRNA binding sites. For this, several approaches have been suggested but most of them are based on heuristic assumptions guided by a few experimental data. To address this issue, we have developed a novel data driven method based on Profile Hidden Markov Models. This method has been denoted as Conditional Profile Hidden Markov Model (CoProHMM) and is shown to outperform existing alignment methods. This work was published in Grau *et al* (Grau, Arend et al. 2010).

Predicting miRNA targets utilizing an extended profile HMM

Jan Grau[1,*], Daniel Arend[1], Ivo Grosse[1], Artemis G. Hatzigeorgiou[2], Jens Keilwagen[3], Manolis Maragkakis[1,2], Claus Weinholdt[1], and Stefan Posch[1]

[1] Institute of Computer Science, Martin Luther University Halle–Wittenberg, Germany

[2] Institute of Molecular Oncology, Biomedical Sciences Research Center 'Alexander Fleming', Vari, Greece

[3] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

Abstract: The regulation of many cellular processes is influenced by miRNAs, and bioinformatics approaches for predicting miRNA targets evolve rapidly. Here, we propose conditional profile HMMs that learn rules of miRNA-target site interaction automatically from data. We demonstrate that conditional profile HMMs detect the rules implemented into existing approaches from their predictions. And we show that a simple UTR model utilizing conditional profile HMMs predicts target genes of miRNAs with a precision that is competitive compared to leading approaches, although it does not exploit cross-species conservation.

1 Introduction

miRNAs are short ($\sim$ 22 nt) endogeneous RNAs that bind to partially complementary sites on mRNA target sequences. They induce cleavage of the miRNA-mRNA duplex or repress translation of the bound mRNA [BSRC05]. Hence, miRNAs influence gene expression and introduce a novel level of gene regulation. For instance, several miRNA signatures have already been successfully associated with human cancers. In animals, miRNAs preferentially bind to the 3' untranslated region (UTR) of the mRNA, and for binding a high complementarity between miRNA and target is required only at the 5' end of the miRNA. Computational miRNA target prediction plays a key role in deciphering the functional role of miRNAs. Several dozen programs have been therefore developed in the last years, and in the following, we describe the main idea behind some of the most widely used programs.

[LSJR$^+$03] propose an algorithm for the prediction of targets of vertebrate miRNAs called TargetScan. TargetScan requires perfect complementarity between positions 2 and 8 at the 5'-end of the miRNA and a potential target, and the free energy of binding between miRNA and target is computed. Predictions are verified using orthologous UTR sequences from other organisms. [LBB05] propose a refined version called TargetScanS, which demands a shorter region of the target to be complementary to nucleotides $2-7$ of the miRNA. TargetScan 5.0 [FFBB09] additionally considers the distance from the 3' UTR and `AU` content.

In contrast to TargetScan, miRanda [EJG+03] does not require perfect complementarity at the seed region, but uses an algorithm similar to Smith-Waterman sequence alignment with similarity scores of $+5$ for `G:C` and A:U basepairs, $+2$ for `G:U` basepairs, and -3 for mismatches, and the scores for the first 11 positions of the alignment are weighted by a factor of 2. Potential target sites (TSs) are filtered for a minimum similarity score and a minimum free energy.

PicTar [KGP+05] searches for perfectly complementary seed regions of 7 nt starting from position 1 or 2 of the miRNA. Mismatches in the seed region are allowed if these do not increase the free energy. Additionally, a filter with respect to the free energy of the complete miRNA-mRNA duplex is applied.

DIANA-microT [MRS+09] prefers perfect complementarity of 7 to 9 nt starting from position 1 or 2 of the miRNA. However, if the considered TS shows good complementarity to the 3' end of the miRNA, the length of this seed region may be reduced to 6 nt, and single `G:U` basepairs are allowed. DIANA-microT uses orthologous UTRs from up to 27 organisms for assessing the conservation of TSs. Finally, the score of a potential UTR target is computed as a weighted average of all predicted TSs.

In contrast to previous approaches, we propose a fully statistical approach for predicting TSs of given miRNAs that is capable of learning rules of miRNA-TS binding from data sets comprising pairs of miRNAs and associated TSs. This approach employs an extension of profile hidden Markov models (HMMs) [KBM+94], which we call *conditional profile HMM* (CoProHMM), and learns parameters by the discriminative maximum supervised posterior (MSP) principle [CdM05, GKK+07]. Since all parameters of CoProHMMs are learned from training data, this approach is not biased towards heuristic assumptions about miRNA-TS interaction like the existence or length of a seed region, unless the training data are.

2 Methods

In the following, we introduce CoProHMMs for modeling the binding between miRNA and TS. We describe how we learn CoProHMMs from data, and we explain how we combine several predictions of a learned CoProHMM to finally predict target genes of a given miRNA.

2.1 Conditional profile HMMs

At the basis of the CoProHMM modeling miRNA TSs, we use a standard profile HMM architecture [KBM+94], which is illustrated in Fig. 1. This architecture is also referred to as "plan9" due to its 9 transitions at each layer of the model. We define a total of K match states M_k, which emit a nucleotide of the TS with a probability that is conditional on the nucleotide at position k of the miRNA. Here, we use $K = 22$, since this is the length of a typical miRNA and, hence, the model covers all positions of the miRNA that are potentially interacting with the TS. If a TS and the associated miRNA are perfectly complementary, we anticipate that only match states are visited for emitting the complete sequence of the TS. Otherwise, silent delete states D_k allow for the insertion of gaps into the TS, insert states I_k allow for including gaps in the miRNA, and match states also allow to replace nucleotides. In Fig. 1, edges represent transition probabilities not fixed to 0. From each node of column k, we can reach node I_k in the same column, and nodes M_{k+1} and D_{k+1} in the next column. Each admissible path starts at D_0 and ends at D_{K+1}. Hence, the states M_0, I_{K+1}, and M_{K+1} are never visited in admissible paths, and are only included to simplify recursive definitions in the following.

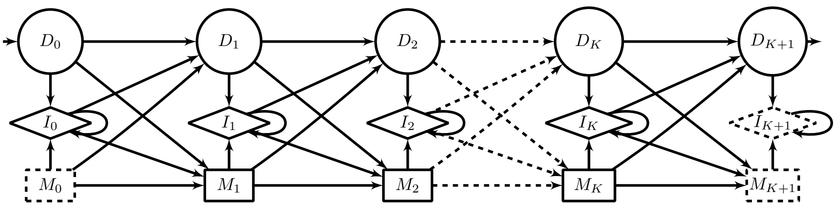


Figure 1: Plan9 architecture of the proposed CoProHMMs. Circles represent silent delete states that do not emit nucleotides of the TS, diamonds represent insert states that emit nucleotides of the TS without considering the nucleotides of the miRNA, and rectangles represent match states that emit nucleotides of the TS with probabilities conditional on the nucleotides of the miRNA. Admissible paths start at D_0 and end at D_{K+1}. States with dashed borders are not visited in admissible paths.

We parameterize the transition probabilities and the emission probabilities by normalized exponentials [Mac98, BB01] using real-valued parameters, since this allows for an unconstrained numerical optimization of the parameters with respect to the discriminative MSP principle.

According to the plan9 architecture, we define the transition probability $P_T(V|S_k, \boldsymbol{\beta}_{T,S_k})$ of going from node $S_k \in \{I_k, M_k, D_k\}$ to node V given parameters $\boldsymbol{\beta}_{T,S_k}$ as

$$P_T(V|S_k, \boldsymbol{\beta}_{T,S_k}) = \begin{cases} \frac{\exp(\beta_{V|S_k})}{\sum_{\tilde{V} \in \{I_k, M_{k+1}, D_{k+1}\}} \exp(\beta_{\tilde{V}|S_k})} & \text{if } V \in \{I_k, M_{k+1}, D_{k+1}\} \\ 0 & \text{otherwise} \end{cases},$$

where $\boldsymbol{\beta}_{T,S_k} = (\beta_{I_k|S_k}, \beta_{M_{k+1}|S_k}, \beta_{D_{k+1}|S_k}), \beta_{V|S_k} \in \mathbb{R}$.

In contrast to standard profile HMMs, we use conditional probabilities depending on the nucleotides of the miRNA for the emissions of the match states. For match state M_k, we define the conditional emission probability $P_{M_k}(a|r_k, \boldsymbol{\beta}_{M_k})$ of symbol a in the TS given the k-th symbol r_k of the miRNA and parameters $\boldsymbol{\beta}_{M_k}$ as

$$P_{M_k}(a|r_k, \boldsymbol{\beta}_{M_k}) = \frac{\exp(\beta_{a|r_k, M_k})}{\sum_{\tilde{a} \in \Sigma} \exp(\beta_{\tilde{a}|r_k, M_k})}, \tag{1}$$

where $\boldsymbol{\beta}_{M_k} = (\beta_{A|A,M_k}, \beta_{C|A,M_k}, \ldots, \beta_{U|U,M_k}), \beta_{a|b,M_k} \in \mathbb{R}$.

Finally, we parameterize the emission probability $P_{I_k}(a|\boldsymbol{\beta}_{I_k})$ of symbol a at insert state I_k given parameters $\boldsymbol{\beta}_{I_k}$ in analogy to equation (1).

We define *forward* variables $\mathcal{F}_{S_k}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta})$ as the probability of observing the first ℓ symbols of the TS sequence $\boldsymbol{x}$ and visiting node S_k in state interval $s(\ell, \boldsymbol{x}|\boldsymbol{r})$ given parameters $\boldsymbol{\beta}$ and the sequence $\boldsymbol{r}$ of the miRNA, i.e.,

$$\mathcal{F}_{S_k}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}) = P\left(x_1, \ldots, x_\ell, S_k \in s(\ell, \boldsymbol{x}|\boldsymbol{r})|\boldsymbol{r}, \boldsymbol{\beta}\right). \tag{2}$$

A node S_k is visited in state interval $s(\ell, \boldsymbol{x}|\boldsymbol{r})$ if it is contained in a path from D_0 to D_{K+1}, and the symbols x_1 to x_ℓ have been emitted either by predecessors of S_k in the path or by S_k itself, whereas $x_{\ell+1}$ is emitted by a successor of S_k in this path.

We use these forward variables for defining the likelihood $P(\boldsymbol{x}|ts, \boldsymbol{r}, \boldsymbol{\beta}_{ts})$ of TS $\boldsymbol{x}$ given the class ts of TS, the sequence of the miRNA $\boldsymbol{r}$, and parameters $\boldsymbol{\beta}_{ts}$, i.e.

$$P(\boldsymbol{x}|ts, \boldsymbol{r}, \boldsymbol{\beta}_{ts}) = \mathcal{F}_{D_{K+1}}(L, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}_{ts}). \tag{3}$$

Using this definition, the likelihood $P(\boldsymbol{x}|ts, \boldsymbol{r}, \boldsymbol{\beta}_{ts})$ is not necessarily normalized over all possible sequences $\boldsymbol{x} \in \Sigma^L$ of given length L.

Similar to original profile HMMs, we recursively derive the forward variables of match state M_k using its predecessors $S_{k-1} \in \{I_{k-1}, D_{k-1}, M_{k-1}\}$ from the previous column of the plan9 architecture (cf. Fig. 1) as

$$\begin{aligned}\mathcal{F}_{M_k}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}) = {} & P_{M_k}(x_\ell|r_k, \boldsymbol{\beta}_{M_k}) \\ & \sum_{S_{k-1}} \mathcal{F}_{S_{k-1}}(\ell-1, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta})\, P_T(M_k|S_{k-1}, \boldsymbol{\beta}_{T,S_{k-1}}).\end{aligned} \tag{4}$$

In analogy, we derive the forward variables of insert states and delete states.

We initialize the forward variables as follows: We can observe D_0 only before the emission of the first symbol. Hence, we set $\mathcal{F}_{D_0}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta})$ to 1 if $\ell = 0$ and to 0 otherwise. We cannot reach M_0 in any admissible path and, thus, $\mathcal{F}_{M_0}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}) = 0$. Finally, we set $\mathcal{F}_{S_k}(0, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}) = 0$ for all emitting states S_k.

2.2 Discriminative training

For learning the parameters of the CoProHMM discriminatively, we need an additional background model. Here, we use a homogeneous Markov model of order 1 with parameters $\boldsymbol{\beta}_{bg}$ that do not depend on the miRNA $\boldsymbol{r}$, i.e.,

$$P(\boldsymbol{x}|bg, \boldsymbol{r}, \boldsymbol{\beta}_{bg}) = P_{hMM(1)}(\boldsymbol{x}|\boldsymbol{\beta}_{bg}). \tag{5}$$

We derive the class posterior of class $c \in \{ts, bg\}$ using the likelihoods $P(\boldsymbol{x}|c, \boldsymbol{r}, \boldsymbol{\beta}_c)$ of equations (3) and (5) as

$$P\left(c\,|\,\boldsymbol{x}, \boldsymbol{r}, \boldsymbol{\beta}\right) = \frac{P(c|\boldsymbol{\beta})P(\boldsymbol{x}|c, \boldsymbol{r}, \boldsymbol{\beta}_c)}{\sum_{\tilde{c}} P(\tilde{c}|\boldsymbol{\beta})P(\boldsymbol{x}|\tilde{c}, \boldsymbol{r}, \boldsymbol{\beta}_{\tilde{c}})}, \tag{6}$$

where $P(c|\boldsymbol{\beta})$ denotes the a-priori probability of class c, which we parameterize in analogy to equation (1).

For Bayesian inference, we define a prior on the parameters $\boldsymbol{\beta}$. For the homogeneous Markov model of class bg, we use a transformed product-Dirichlet prior [Mac98] with equivalent sample size (ESS) [HGC95] $\alpha_{bg} \cdot K$. We define another transformed product-Dirichlet prior with ESS α_{ts} for the parameters of the CoProHMM, which is the product of independent transformed Dirichlet priors for each set of transition parameters and each set of emission parameters. We use Dirichlet priors, since these are conjugate to the likelihood of the homogeneous Markov model and to the distribution of transitions and (conditional) emissions. Hence, their hyper-parameters can be intuitively interpreted as pseudo counts. In the following studies, we use $\alpha_{bg} = \alpha_{ts} = 4$.

We learn all parameters $\boldsymbol{\beta}$ on a set of labelled training data $(\boldsymbol{x}_1, \boldsymbol{r}_1, c_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{r}_n, c_N)$. These training data comprise a sufficient number of TSs, i.e. $c_n = ts$, and non-TSs of several miRNAs. Learning the parameters on the TSs of multiple miRNAs conjointly is motivated by the expectation that by this means, CoProHMM may detect general rules of miRNA-TS binding, that could not be detected if we, for instance, learned a standard profile HMM on the TSs of a single miRNA.

We optimize the parameters with respect to the discriminative MSP principle [CdM05, GKK$^+$07], i.e.,

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left[\prod_{n=1}^{N} P\left(c_n \,|\, \boldsymbol{x}_n, \boldsymbol{r}_n, \boldsymbol{\beta}\right) \right] q\left(\boldsymbol{\beta} \,|\, \alpha_{bg}, \alpha_{ts}\right), \tag{7}$$

where $q\left(\boldsymbol{\beta} \,|\, \alpha_{bg}, \alpha_{ts}\right)$ denotes the product-Dirichlet priors on the parameters $\boldsymbol{\beta}$. This optimization must be carried out numerically, which we accomplish by a quasi-Newton second order method.

2.3 Predicting target genes

In the following, we describe how we utilize a CoProHMM for predicting target genes of a miRNA $\boldsymbol{r}$. We assume that the CoProHMM has already been trained on a set of miRNAs – not necessarily including $\boldsymbol{r}$ – and associated TSs and non-TSs. To this end, we extract the UTR $\boldsymbol{y}_n$ of each gene n. Using a sliding window of width $|\boldsymbol{r}|$, we apply the CoProHMM to each sub-sequence of $\boldsymbol{y}_n$ and compute the log-likelihood according to equation (3) given miRNA $\boldsymbol{r}$. For each UTR, we consider the I sub-sequences yielding the largest log-likelihoods $s_{n,i}$, which end at positions $q_{n,i}$. Let $d_n = q_{n,1}$ and $d'_n = |\boldsymbol{y}_n| - q_{n,1}$ be the distance of the sub-sequence with the largest log-likelihood to the 3' and 5' end of the UTR, respectively. Let $(p_{n,1}, \ldots, p_{n,I})$ denote the positions $(q_{n,1}, \ldots, q_{n,I})$ sorted ascendingly. Let $\boldsymbol{z}_n = (s_{n,1}, \ldots, s_{n,I}, d_n, d'_n, p_{n,1}, \ldots, p_{n,I})$ denote the vector of these features representing UTR $\boldsymbol{y}_n$.

By inspecting histograms of the scores $s_{n,i}$, we find that these may be modeled by a mixture of two Gaussian densities, i.e.,

$$P(s_{n,i}|\boldsymbol{\beta}^s_{c,i}) = P(u^s = 1|\boldsymbol{\beta}^{s,m}_{c,i})\, \mathcal{N}(s_i|\mu_{1,i,c}, \kappa_{1,i,c}) + P(u^s = 2|\boldsymbol{\beta}^{s,m}_{c,i})\, \mathcal{N}(s_i|\mu_{2,i,c}, \kappa_{2,i,c}),$$

where $\boldsymbol{\beta}^s_{c,i} = (\boldsymbol{\beta}^{s,m}_{c,i}, \mu_{1,i,c}, \kappa_{1,i,c}, \mu_{2,i,c}, \kappa_{2,i,c})$, $\mu_{k,i,c}$ and $\kappa_{k,i,c}$ denote the mean and

the log-precision of Gaussian density k, respectively, and the component probabilities $P(u^s = u|\boldsymbol{\beta}_{c,i}^{s,m})$ are parameterized in analogy to equation (1).
To allow for variability in TS positioning, we model d_n and d'_n each by a mixture of two gamma densities, i.e.,

$$P(d_n|\boldsymbol{\beta}_c^d) = P(u^d = 1|\boldsymbol{\beta}_c^{d,m})\, \mathcal{G}(d_n|\alpha_{1,c}^d, \beta_{1,c}^d) + P(u^d = 2|\boldsymbol{\beta}_{c,i}^{d,m})\, \mathcal{G}(d_n|\alpha_{2,c}^d, \beta_{2,c}^d),$$

where $\boldsymbol{\beta}_c^d = (\boldsymbol{\beta}_c^{d,m}, \alpha_{1,c}^d, \beta_{1,c}^d, \alpha_{2,c}^d, \beta_{2,c}^d)$, and $\alpha_{k,c}^d$ and $\beta_{k,c}^d$ denote the log-shape and log-rate of gamma density k, respectively. We define the density $P(d'_n|\boldsymbol{\beta}_c^{d'})$ in analogy.

We model the distances $p_{n,i+1} - p_{n,i}$ by another gamma density, i.e.,

$$P(p_{n,i+1} - p_{n,i}|\boldsymbol{\beta}_c^p) = \mathcal{G}(p_{n,i+1} - p_{n,i}|\alpha_c^p, \beta_c^p),$$

where $\boldsymbol{\beta}_c^p = (\alpha_c^p, \beta_c^p)$.

The complete likelihood of $\boldsymbol{z}_n$ representing UTR $\boldsymbol{y}_n$ of gene n employing convenient independence assumptions amounts to

$$P(\boldsymbol{z}_n|c, \boldsymbol{\beta}_c) \propto \prod_{i=1}^{I} P(s_{n,i}|\boldsymbol{\beta}_{c,i}^s)\; P(d_n|\boldsymbol{\beta}_c^d)\; P(d'_n|\boldsymbol{\beta}_c^{d'}) \prod_{i=1}^{I-1} P(p_{n,i+1} - p_{n,i}|\boldsymbol{\beta}_c^p). \tag{8}$$

In the following studies, we use $I = 5$.

In analogy to equation (6), we define the class posterior in terms of likelihoods $P(\boldsymbol{z}_n|c, \boldsymbol{\beta}_c)$ and a-priori class probabilities $P(c|\boldsymbol{\beta})$. As for the training of the TS model, we optimize the parameters with respect to the discriminative MSP principle (cf. equation (7)) using a training data set of target and non-target genes. In this case, we use beta priors on the parameters of the component probabilities, normal-gamma priors on the parameters of the Gaussian densities, and the conjugate prior according to the definition of the exponential family for the gamma densities. Again, we use an ESS of 4 for both classes. We finally predict target genes based on the class posterior.

3 Results & Discussion

In the following, we first investigate if CoProHMMs can learn characteristics of TSs from data. To this end, we use TSs predicted by existing approaches. Second, we evaluate the utility of CoProHMMs for the prediction of target genes of miRNAs on a benchmark data set.

3.1 Pilot study: Learning CoProHMMs from predictions

We learn CoProHMMs on the predictions of miRanda and TargetScan to investigate if CoProHMMs can learn the rules implemented into these approaches from their predictions. We choose miRanda and TargetScan, because their approaches differ notably. If

CoProHMMs can detect such characteristics from predictions, we might expect that they are also capable of learning novel or refined rules of miRNA-TS binding from experimentally verified TS.

We extract all human TSs and associated miRNAs predicted by TargetScan and miRanda from miRNAMap[1] [HCT$^+$08]. For TargetScan, we use all 244,389 TSs, while we randomly sample 500,000 TSs from the predictions of miRanda. We generate a non-target data set by randomly selecting miRNAs from the mature human miRNAs listed at miRBase[2] [GJSvDE08]. As non-TSs of these miRNAs, we randomly draw 500,000 subsequences of length $|\boldsymbol{r}| \pm 3$ from 3'-UTRs of human genes according to NCBI Genbank[3] human genome build 37.1.

We present a graphical representation of the CoProHMMs learned on the miRanda data set and the TargetScan data set in Fig. 2. Here, we depict only the most interesting region around the seed, while the complete CoProHMMs for miRanda and TargetScan as well as other approaches are available online[4]. For the states, we use the same shapes as in Fig. 1. The thickness of outgoing edges represents the transition probabilities to the successors of a node. We illustrate the emission probabilities of insert states by a row of grayscale boxes, where the first box corresponds to A, the second box corresponds to C, the third box corresponds to G, and the fourth box corresponds to U. The darker a box, the higher is the corresponding emission probability. In analogy, the conditional emission probabilities of match states are represented by a matrix comprising such rows, where each row corresponds to the conditional probability distribution given one nucleotide of the miRNA. The probabilities of visiting a state are visualized by the darkness of the background of each node. The darker the background of a node the higher the probability of visiting this node.

Considering the CoProHMM learned on the miRanda data set, we recover many rules built into miRanda. From the conditional emission probabilities of the match states, we observe a general tendency to complementary base pairings between the TS and the miRNA. This tendency is especially pronounced for the match states in the seed region, but can also be observed for the match states at position 1 and positions 9 to 11. We also detect a slight preference for `G:U` wobble basepairs. These observations are most likely a result of the Smith-Waterman like alignment employed by miRanda. Additionally, miRanda assigns a weight of 2 to the first 11 positions of the alignment, which is reflected by the increased probabilities of visiting match states in the seed region, although this preference already begins to decline at position 8 of the learned CoProHMM.

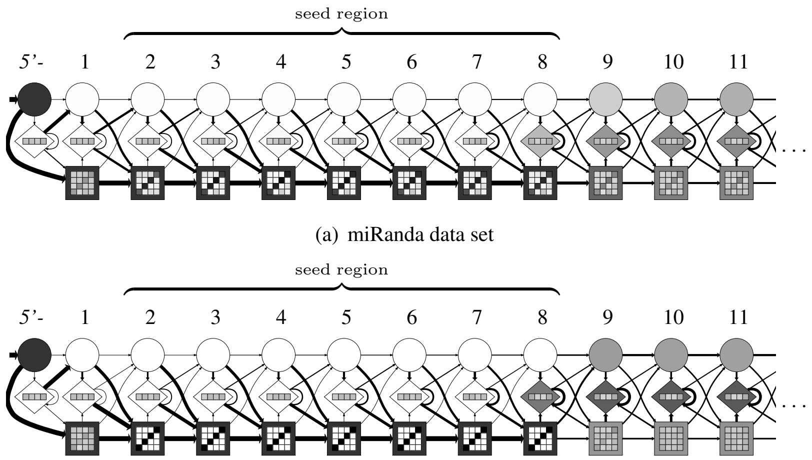
As a second example, we consider the CoProHMM learned on the TargetScan data set in Fig. 2(b). Notable differences between the CoProHMM for the TargetScan data set and the miRanda data set can be observed for the conditional emission probabilities at the match states. At positions 2 to 8 of Fig. 2(b), we find complementary basepairs almost exclusively, while a slight preference for complementary basepairs is present at the bordering

[1] ftp://mirnamap.mbc.nctu.edu.tw/miRNAMap2/miRNA_Targets/Homo_sapiens/miRNA_targets_hsa.txt.tar.gz

[2] http://www.mirbase.org

[3] http://www.ncbi.nlm.nih.gov

[4] http://www.jstacs.de/index.php/MiRNAs



(a) miRanda data set

(b) TargetScan data set

Figure 2: CoProHMMs learned on the miRanda data set (a) and TargetScan data set (b).

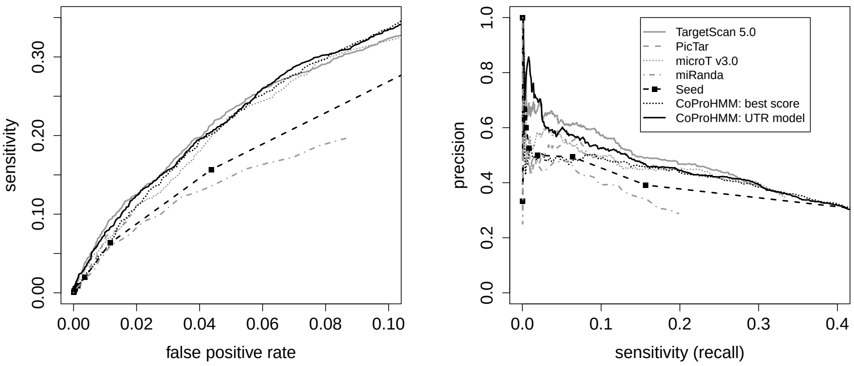
positions 1 and 9. In contrast, the remaining positions exhibit only very slight preferences for specific basepairs. Again, these findings are closely related to the main characteristics built into TargetScan. The perfect complementarity at positions 2 to 8 of the CoProHMM reflects the requirements of TargetScan. We also observe a preference for complementary basepairs at positions 1 and 9, which most likely can be attributed to the fact that initial perfect matches in the seed region may be elongated to either side in TargetScan.

These findings suggest that CoProHMMs are indeed capable of recovering the rules built into miRanda and TargetScan from prediction and, hence, may also be capable of inferring the rules underlying miRNA-TS binding from experimentally verified TSs, once these become available in sufficient quantity.

3.2 Benchmark study: Predicting miRNA target genes

We investigate the utility of CoProHMMs for the prediction of miRNA target genes using the pSILAC data of Selbach *et al.*, which have also been used in recent benchmark studies [SST$^+$08, AMP$^+$09]. To this end, we learn a CoProHMM using a foreground data set that comprises 12 verified TSs and 667 predicted TSs within UTRs of verified target genes extracted from mirecords[5] v. 1 [XZC$^+$09]. As these TSs are too few to reliably learn the models, we also include the TargetScan data set and 405,569 TSs predicted by DIANA-microT. We use predictions of these two approaches, since they yield reasonable precisions in the benchmark studies. We use the same background data set as in the pilot study. We assign a weight of 500 to all verified TSs and a weight of 50 to all predicted TSs in verified target genes to reflect our increased confidence in these data, while we assign a weight of

[5] `http://mirecords.biolead.org/download_data.php?v=1`



(a) ROC curve

(b) Precision-recall curve

Figure 3: ROC curve (a) and precision-recall curve (b) of the classifier using the UTR model (solid black line) and the classifier using the best score of the CoProHMM within each UTR sequence (dotted black line) compared to other approaches.

1 to all other TSs. All TSs of miRNAs contained in the Selbach benchmark data set are excluded when training the CoProHMM to allow for unbiased evaluation.

We extract the UTRs of all genes considered in [SST$^+$08] according to [AMP$^+$09]. For these genes, Selbach *et al.* measured the influence of overexpression or underexpression of a miRNA on the abundance of the corresponding proteins for 5 different miRNAs. For each of these miRNAs, we partition the UTRs into target and non-target UTRs using a threshold of -0.2 on the protein log-fold changes. We assess the performance of the UTR model using the predictions of the CoProHMM in a 5-fold cross validation. In each iteration of the cross validation, we train the parameters of the UTR model on the numeric vectors $\boldsymbol{z}_n$ obtained for 4 of the 5 miRNAs, and we compute the log-likelihood ratios using this trained UTR model for the numeric vectors obtained for the remaining miRNA. In analogy to [AMP$^+$09], we finally use all log-likelihood ratios to compute sensitivity, precision, and false positive rate for different thresholds.

In Fig. 3, we compare the performance of the classifier using the UTR model (solid black line) to other approaches by means of the precision-recall curve and the ROC curve. As a reference, we also include the performance of a classifier that only uses the best score of the CoProHMM over each UTR sequence, i.e., $s_{n,1}$, (dotted black line). Considering Fig. 3(a), we find that even this classifier using only the best score yields a substantially higher sensitivity than miRanda and Seed for a broad range of false positive rates. Surprisingly, the classifier using the simple UTR model, which does not exploit conservation across species, achieves comparable or slightly improved sensitivities compared to miRanda, Seed, PicTar, and microT, while it performs only slightly worse than TargetScan 5.0 for false positive rates below 0.06.

Turning to the precision-recall curve in Fig. 3(b), we find a similar picture. Notably, the classifier using the UTR model again achieves comparable or even higher precisions than miRanda, Seed, PicTar, and microT. However, it can outperform TargetScan 5.0 only for very low sensitivities and yields lower precisions for sensitivities between 0.03 and 0.28.

The performance of both classifiers using CoProHMMs is astonishing, because, in contrast to most of the other approaches, they do not exploit conservation across different species. Hence, the inclusion of cross-species conservation into CoProHMMs and the proposed UTR model, and the integration of CoProHMMs into other approaches might be a worthwhile direction of future research.

4 Conclusions

miRNAs are involved in the regulation of many cellular processes, and the prediction of miRNA targets is one of the most active fields of bioinformatics. Here, we propose a novel statistical model called conditional profile HMM (CoProHMM) for learning the rules of miRNA-TS interaction from data. We demonstrate that CoProHMMs are capable of reconstructing patterns of miRNA-TS binding built into existing programs from predictions of these approaches.

Conservation is key feature of most miRNA target prediction approaches leading to higher precision at the expense of sensitivity. Interestingly, we find in a benchmark study that a simple UTR model utilizing CoProHMMs yields a competitive precision compared to leading approaches for predicting target genes, although it does not exploit conservation across species.

We anticipate that the number of experimentally verified TSs will rapidly increase in the next years. Only recently, [CZMD09, HLB$^+$10] have independently published novel biological data that shed light on miRNA targeting. Briefly, the two experimental approaches use in-vivo crosslinking, Ago2 immunoprecipitation and cDNA sequencing, and have been able to determine TSs of several miRNAs with high accuracy. Since the power of statistical approaches like CoProHMMs highly depends on the quality of the training data, we might speculate that the performance of CoProHMMs will even increase using these data. Additionally, CoProHMMs might be a suitable approach to extract new and refined rules of miRNA-TS binding from such verified TSs.

We make an implementation of the CoProHMMs and the UTR model available to the scientific community with the next release of the open source Java library Jstacs[6].

References

[AMP$^+$09] Panagiotis Alexiou, Manolis Maragkakis, Giorgos L. Papadopoulos, Martin Reczko, and Artemis G. Hatzigeorgiou. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, 25(23):3049–3055, 2009.

[6] http://www.jstacs.de

[BB01] Pierre Baldi and Søren Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, London, 2nd edition, 2001.

[BSRC05] Julius Brennecke, Alexander Stark, Robert B. Russell, and Stephen M. Cohen. Principles of MicroRNA–Target Recognition. *PLoS Biology*, 3(3), 2005.

[CdM05] Jesús Cerquides and Ramon López de Mántaras. Robust Bayesian Linear Classifier Ensembles. In *Proceedings of the 16th European Conference on Machine Learning*, volume 3720 of *Lecture Notes in Computer Science*, pages 72–83. Springer, 2005.

[CZMD09] Sung Wook Chi, Julie B. Zang, Aldo Mele, and Robert B. Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–486, 07 2009.

[EJG$^+$03] Anton Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora Marks. MicroRNA targets in Drosophila. *Genome Biology*, 5(1):R1, 2003.

[FFBB09] Robin C. Friedman, Kyle Kai-How Farh, Christopher B. Burge, and David P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, 2009.

[GJSvDE08] Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Research*, 36(suppl_1):D154–158, 2008.

[GKK$^+$07] Jan Grau, Jens Keilwagen, Alexander Kel, Ivo Grosse, and Stefan Posch. Supervised posteriors for DNA-motif classification. In Claudia Falter, Alexander Schliep, Joachim Selbig, Martin Vingron, and Dirk Walther, editors, *German Conference on Bioinformatics*, volume 115 of *Lecture Notes in Informatics (LNI) - Proceedings*, Bonn, 2007. Gesellschaft für Informatik.

[HCT$^+$08] Sheng-Da Hsu, Chia-Huei Chu, Ann-Ping Tsou, Shu-Jen Chen, Hua-Chien Chen, Paul Wei-Che Hsu, Yung-Hao Wong, Yi-Hsuan Chen, Gian-Hung Chen, and Hsien-Da Huang. miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. *Nucleic Acids Research*, 36(suppl_1):D165–169, 2008.

[HGC95] David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. In *Machine Learning*, pages 197–243, 1995.

[HLB$^+$10] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. 141(1):129–141, 04 2010.

[KBM$^+$94] Anders Krogh, Michael Brown, I. Saira Mian, Kimmen Sjölander, and David Haussler. Hidden Markov Models in Computational Biology : Applications to Protein Modeling. *Journal of Molecular Biology*, 235(5):1501 – 1531, 1994.

[KGP$^+$05] Azra Krek, Dominic Grun, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J Epstein, Philip MacMenamin, Isabelle da Piedade, Kristin C Gunsalus, Markus Stoffel, and Nikolaus Rajewsky. Combinatorial microRNA target predictions. *Nature Genetics*, 37(5):495–500, 05 2005.

[LBB05] Benjamin P. Lewis, Christopher B. Burge, and David P. Bartel. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell*, 120(1):15 – 20, 2005.

[LSJR$^+$03] Benjamin P. Lewis, I-hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. Prediction of Mammalian MicroRNA Targets. *Cell*, 115(7):787 – 798, 2003.

[Mac98] David J. C. MacKay. Choice of Basis for Laplace Approximation. *Machine Learning*, 33(1):77–86, 1998.

[MRS$^+$09] 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. *Nucleic Acids Research*, 37(suppl_2):W273–276, 2009.

[SST⁺08] Matthias Selbach, Bjorn Schwanhausser, Nadine Thierfelder, Zhuo Fang, Raya Khanin, and Nikolaus Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, 09 2008.

[XZC⁺09] Feifei Xiao, Zhixiang Zuo, Guoshuai Cai, Shuli Kang, Xiaolian Gao, and Tongbin Li. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Research*, 37(suppl_1):D105–110, 2009.

5.3. Editing of Epstein-Barr virus-encoded BART6 miRNAs controls their dicer targeting and consequently affects viral latency

Usually, top performing target prediction programs exploit information regarding evolutionary conservation of predicted miRNA binding sites. However, this informative feature might as well decrease prediction performance in specific cases. This for example might happen when miRNA targeting has a negative effect on the organism's survival and consequently the organism tends to avoid it. This is the case regarding viral miRNA targeting against a host organism. Taking this into account, in the following publication we describe a combined computational and experimental approach for viral miRNAs of Epstein-Barr virus where we found that ebv-miR-BART6-5p silence Dicer through multiple non conserved target sites located in the 3'UTR of Dicer mRNA and that mutation and A-to-I editing appear to be adaptive mechanisms that antagonize ebv-miR-BART6 activities consequently affecting viral latency. This work was published in Iizasa *et al* (Iizasa, Wulff et al. 2010).

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 285, NO. 43, pp. 33358–33370, October 22, 2010
© 2010 by The American Society for Biochemistry and Molecular Biology, Inc. Printed in the U.S.A.

Editing of Epstein-Barr Virus-encoded BART6 MicroRNAs Controls Their Dicer Targeting and Consequently Affects Viral Latency*[S]

Received for publication, April 26, 2010, and in revised form, August 16, 2010 Published, JBC Papers in Press, August 17, 2010, DOI 10.1074/jbc.M110.138362

Hisashi Iizasa[‡§], Bjorn-Erik Wulff[‡], Nageswara R. Alla[‡], Manolis Maragkakis[¶‖], Molly Megraw[**], Artemis Hatzigeorgiou[¶], Dai Iwakiri[§], Kenzo Takada[§], Andreas Wiedmer[‡], Louise Showe[‡], Paul Lieberman[‡], and Kazuko Nishikura[‡1]

From the [‡]The Wistar Institute, Philadelphia, Pennsylvania 19104, the [§]Institute for Genetic Medicine, Hokkaido University, Sapporo 060-0815, Japan, the [¶]Institute of Molecular Oncology, Biomedical Sciences Research Center Alexander Fleming, 16672 Vari-Athens, Greece, the [‖]Institute of Computer Science, Martin Luther University Halle-Wittenberg, 06120 Halle, Germany, and the [**]Department of Genetics, School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Certain primary transcripts of miRNA (pri-microRNAs) undergo RNA editing that converts adenosine to inosine. The Epstein-Barr virus (EBV) genome encodes multiple microRNA genes of its own. Here we report that primary transcripts of ebv-miR-BART6 (pri-miR-BART6) are edited in latently EBV-infected cells. Editing of wild-type pri-miR-BART6 RNAs dramatically reduced loading of miR-BART6-5p RNAs onto the microRNA-induced silencing complex. Editing of a mutation-containing pri-miR-BART6 found in Daudi Burkitt lymphoma and nasopharyngeal carcinoma C666-1 cell lines suppressed processing of miR-BART6 RNAs. Most importantly, miR-BART6-5p RNAs silence Dicer through multiple target sites located in the 3′-UTR of Dicer mRNA. The significance of miR-BART6 was further investigated in cells in various stages of latency. We found that miR-BART6-5p RNAs suppress the EBNA2 viral oncogene required for transition from immunologically less responsive type I and type II latency to the more immunoreactive type III latency as well as Zta and Rta viral proteins essential for lytic replication, revealing the regulatory function of miR-BART6 in EBV infection and latency. Mutation and A-to-I editing appear to be adaptive mechanisms that antagonize miR-BART6 activities.

MicroRNAs (miRNAs)[2] play important roles in many processes including development, differentiation, proliferation, and apoptosis (1, 2). Certain miRNAs act as tumor suppressors or oncogenes and are associated with many cancers (3). Primary transcripts of miRNA genes (pri-miRNAs) are processed sequentially by Drosha and Dicer (4, 5). Nuclear Drosha (6) together with its partner DGCR8 (7, 8) cleaves pri-miRNAs, releasing 60–70-nucleotide pre-miRNAs. Recognition of correctly processed pre-miRNAs and their nuclear export is carried out by exportin-5 and RanGTP (9). Cytoplasmic Dicer together with the double-stranded RNA (dsRNA)-binding protein TRBP then cleaves pre-miRNAs into 20–22-nucleotide siRNA-like duplexes (10, 11). In most cases one strand of the duplex (called the effective strand) serves as the mature miRNA, whereas the other strand (called passenger strand) is eliminated. After integration into the miRNA-induced silencing complex (miRISC), miRNAs block translation via partially complementary binding sites located in the 3′-UTRs of targeted mRNAs or guide the degradation of target mRNAs after binding, mainly via the 5′ half of the miRNA sequence, called the “seed sequence” (1, 4, 5).

Epstein-Barr Virus (EBV) causes mononucleosis during acute and lytic infection and also establishes a persistent and latent infection in the human host. Latently infected EBV has been demonstrated to be associated with a variety of human cancers such as Burkitt lymphoma, Hodgkin disease, and nasopharyngeal carcinoma (12, 13). Lytic infection and transition to distinctive states of latency (type I-III) are regulated by select viral genes and their interaction with the host immune system (12, 13). Virus genomes encode miRNAs of their own, and the first viral miRNA was identified in human B cells infected with EBV (14). A total of 23 EBV miRNA genes are known and located in the BHRF1 and BART (Bam H1 A rightward transcript) regions of the genome (15–17). These EBV miRNAs have been implicated in regulating the transition from lytic replication to latent infection and in attenuating antiviral immune responses (18). However, only a limited number of their targets have been identified so far. The viral miRNAs seem to target both viral and host cell genes (18). For instance, miR-BART2 targets the EBV DNA polymerase, BALF5, perhaps promoting entry of the virus to latency by slowing down viral replication at the transition point from lytic to latent infection (19). Down-regulation of the EBV protein LMP1 by three EBV miRNAs, miR-BART1–5p, miR-BART16, and miR-BART17–5p, has been reported (20). LMP1 produced during the EBV type II and III latency controls the NF-$\kappa\beta$ signaling pathway and growth and apoptosis of host cells. Targeting of

[*] This work was supported, in whole or in part, by National Institutes of Health Grants CA010815, GM040536, and HL099342. This work was also supported by the Ellison Medical Foundation (AG-SS-2281-09) and the Commonwealth Universal Research Enhancement Program, Pennsylvania Department of Health (to K. N.).

[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental Experimental Procedures, Table 1, and Figs. 1–6.

[1] To whom correspondence should be addressed: The Wistar Institute, 3601 Spruce St., Philadelphia, PA 19104-4268; Tel.: 215-898-3828; Fax: 215-898-3911; E-mail: kazuko@wistar.org.

[2] The abbreviations used are: miRNA, microRNA; ADAR, adenosine deaminases acting on RNA; miRISC, miRNA-induced silencing complex; EBV, Epstein-Barr virus; qRT, quantitative real-time.

host cell genes PUMA (p53-up-regulated modulator of apoptosis) by miR-BART5 (21) and CXC-chemokine ligand 11 (CXCL11) by miR-BHRF1–3 (22) have been reported. Down-regulation of PUMA may suppress apoptosis of virus-infected host cells (21), whereas suppression of CXCL11 may shield EBV-infected B cells from cytotoxic T cells (22).

One type of RNA editing involves the conversion of adenosine residues into inosine (A-to-I editing) in dsRNA through the action of adenosine deaminase acting on RNA (ADAR). Three ADAR gene family members (ADAR1–3) have been identified in humans and rodents (23, 24). The translation machinery reads an inosine as if it were guanosine, which could lead to codon changes (25). Thus, when A-to-I RNA editing occurs within a coding sequence, synthesis of proteins not directly encoded by the genome can result, as demonstrated with transcripts of glutamate receptor ion channels and 5-HT_{2C} serotonin receptors (26). However, the most common targets for A-to-I editing are non-coding RNAs that contain inverted repeats of repetitive elements such as Alu elements and LINEs located within introns and 3′-UTRs (27–30). The biological significance of non-coding, repetitive RNA editing is largely unknown. Furthermore, editing of certain pri-miRNAs has been reported (31, 32). A recent survey has revealed that ~20% of human pri-miRNAs are subject to A-to-I RNA editing catalyzed by ADAR1 and ADAR2 (33). Editing of pri-miRNAs modulates expression and function of miRNAs (33). For instance, A-to-I editing of several adenosine residues located near the Drosha cleavage sites of pri-miRNA-142 results in inhibition of the processing by Drosha and consequent down-regulation of mature miR-142 RNAs (34), whereas editing of two sites identified near the end loop of the pri-miR-151 hairpin structure inhibits the Dicer cleavage step (35). By contrast, editing of primary transcripts of the miR-376 cluster at two sites located within the seed sequence does not affect their processing but results in expression of mature-edited miR-376 RNAs with altered seed sequences and consequent silencing of a set of genes different from those targeted by unedited miR-376 RNAs (36).

In this study we set out to examine editing of EBV miRNAs in EBV-transformed lymphoblastoid GM607 cells, Burkitt lymphoma Daudi cells, and nasopharyngeal carcinoma C666-1 cells. Human lymphoblastoid cells such as GM607 cells in type III latency express a set of genes essential for this specific state of latency, such as EBNA2 and LMP1. By contrast, Daudi Burkitt lymphoma cells in the restricted sub-type of type III latency do not express EBNA2 due to the genomic deletion (37, 38). Viral infection in C666-1 nasopharyngeal carcinoma cells is associated with more restricted forms of type II latency, which expresses only a limited number of viral genes, representing a less immune-responsive state (38). We have found that primary transcripts of four EBV miRNAs, including miR-BART6, are subject to A-to-I editing. Moreover, we demonstrate that editing of pri-miR BART6 RNAs as well as mutations of miR-BART6 RNAs found in latently EBV-infected cells inhibits expression or their loading onto the functionally active miRISC. Most significantly, we found that miR-BART6 targets Dicer and affects the latent state of EBV viral infection. Regulation of the miR-BART6 expression and function through A-to-I editing and mutation may be critical for the establishment or maintenance of latent EBV infection.

EXPERIMENTAL PROCEDURES

Cell Culture—EBV-transformed lymphoblastoid cell line GM607 (GM00607) was obtained from Coriell Institute for Medical Research (Camden, NJ). Burkitt lymphoma cell line Daudi was obtained from American Type Culture Collection (Manassas, VA). Burkitt lymphoma Mutu I and Mutu III and nasopharyngeal carcinoma line C666-1 were used in our previous studies (39–41). These cell lines were cultured in RPMI1640 (Mediatech Inc., Manassas, VA), supplemented with 100 units/ml benzylpenicillin, 100 μg/ml of streptomycin sulfate (both from Invitrogen) and 10% fetal calf serum (FCS) (Tissue Culture Biological, Tulare, CA). HeLa and HEK293T cells were cultured in Dulbecco's modified Eagle's medium (Invitrogen) supplemented with 10% FCS.

Analysis of in Vitro Processed Pri-miRNA Products by Northern Blotting—Nonradioactive pri-miR-BART6 RNAs (10 fmol) were synthesized by *in vitro* transcription and processed by Drosha-DGCR8 (20 ng) and/or Dicer-TRBP complexes (20 ng) as described previously (34). Processed RNAs were electrophoresed on a 15% polyacrylamide, 8 M urea gel and transferred to a Hybond XL membrane (GE Healthcare) by electroblotting. Membranes were UV-cross-linked (StrataLinker; Stratagene, La Jolla, CA), and hybridized with 5′-^{32}P-labeled miRCURY locked nucleic acid probes (Exiqon Inc., Woburn, MA) and analyzed by Northern blotting. The hybridization buffer contained 50% formamide, 0.5% SDS, 5× saline/sodium phosphate/EDTA, 5× Denhardt's solution, and 20 μg/ml sheared, denatured, salmon sperm DNA. Hybridization was conducted at 34 °C. Membranes were washed by 2× SSC/0.1% SDS, and hybridized signals were quantified by a Typhoon Imager System.

Luciferase Reporter Constructs—The human Dicer 3′-UTR (1498 bp), which contains four miR-BART6-5p binding sites (supplemental Experimental Procedures), were amplified using human genomic DNA extracted from GM607 cells and specific primers hDicerFW (5′-GCT**ACTAGT**GATCTTTGGCTAAACACCCCAT-3′) and hDicerRV (5′-GCT**GTTTAAAC**CTCCAACAAAAGTGAAACGGC-3′). The PCR products were inserted into a luciferase reporter vector (pMIR-REPORT^{TM} Luciferase; Ambion) after digestion with Spe1 and Pml1.

Transfections of miR-BART6 RNAs—miR-BART6-5p and unedited miR-BART6–3p RNA duplexes were synthesized at Ambion (Pre-miR^{TM} miRNA). All transfections were carried out in triplicate as described previously (42). Briefly, HeLa cells were pre-plated in 24-well tissue culture plates. 200 ng of luciferase reporter plasmid and 200 ng of control vector pMIR-REPORT^{TM} β-galactosidase Control Plasmid (Ambion) were diluted into 50 μl of Opti-MEM (Invitrogen) with or without 10 pmol of miR-BART6 duplex or sequence unrelated control miRNA, cel-miR-67, or miR-376a followed by the addition of 3 μl of Lipofectamine 2000 (Invitrogen). The transfection mixture was incubated at room temperature for 5 min followed by the addition of DNA/miR-BART6 and further incubated at room temperature for 20 min. Then 100 μl of transfection mixture was added to the HeLa cells in 500 μl of the growth medium. Transfection efficiency monitored by using 5-carboxylfluorescein-labeled control siRNA (Ambion) was more than 80%. Forty-eight hours after transfection, the luciferase activity was measured using the Luciferase Assay System (Promega, Madison, WI) together with β-galactosidase activity by reading the absorbance at 415 nm in a plate reader with the β-Galactosidase Enzyme Assay System (Promega). Normalized luciferase values divided by the β-galactosidase activity were statistically compared among each group by Mann-Whitney *U* test.

Transfections of the miR-BART6-5p Antagomir—C666.1 or Mutu I cells (1.5 × 10^5 cells) were cultured in 24-well plates. The next day cells were transfected with 50 pmol of inhibitor of miR-BART6-5p (miScript miRNA inhibitor, Qiagen) or AllStars Negative Control siRNA (Qiagen) using 3 μl of Hiperfect Transfection reagent (Qiagen). After 72 h, total RNA was extracted and treated with DNase I. First-strand cDNA was synthesized using 1 μg of total RNA using miScript Reverse transcription kit (Qiagen) or Superscript III with random primer.

Transfections of the Dicer Targeting shRNA Expression Vector—To suppress Dicer expression, a short hairpin expression vector was used. Using BLOCK-iT RNAi Designer (Invitrogen), complementary DNA oligos were designed. For construction of Dicer shRNA plasmids, sense (5′-CACCGCAGCTCTGGATCATAATACCCGAAGGTATTATGATCCAGAGCTGC-3′) and antisense (5′-AAAAGCAGCTCTGGATCATAATACCTTCGGGTATTATGATCCAGAGCTGC-3′) strand oligos were synthesized. For construction of LacZ2.1 Control, sense (5′-CACCAAATCGCTGATTTGTGTAGTCGGAGACGACTACACAAATCAGCGA-3′) and antisense (5′-AAAATCGCTGATTTGTGTAGTCGT CTCCGACTA CACAAATCAGCGATTT-3′) strand oligos were synthesized. To generate a double-stranded DNA, these oligos were annealed and cloned into pENTER/H1/TO vector (Invitrogen). C666.1 or Mutu III cells (1 × 10^6) were transfected with 1 μg of vector DNA using CUY21Pro-Vitro (NEPA GENE., Co Ltd, Ichikawa, Japan). After 48 h, total RNA was extracted and treated with DNase I. First-strand cDNA was synthesized using 1 μg of total RNA using the miScript reverse transcription kit (Qiagen) or Superscript III with random primers. Transfection efficacy monitored by co-transfection of ptdTomato-C1 vector (Clontech) was ~70–80%.

Induction of Viral miRNA Expression in HEK293T Cells—The pri-miR-BART6 sequences were PCR-amplified using genomic DNA extracted from GM607 cells and a set of primers, LentiBART6FW (5′-GC**CTCGAG**TGACCTTGTTGGTACTTTAAGGTTG-3′) and LentiBART6-UneditedRV (5′-GC**GAATTC**TGGCCTTGAGTTACTCTAAGGCTA-3′) containing a thymidine residue at the +20 site or LentiBART6-Edited RV (5′-GCGAATTCTGGCCTTGAGTTACTCCAAGGCTA-3′) containing a cytidine residue (edited) at the +20 site. These PCR products were digested with XhoI and EcoRI (New England Biolabs, Ipswich, MA) and ligated into pTRIPZ vector (Open Biosystems, Huntsville, AL). pTRIPZ-derived lentiviruses were transfected into HEK293T in the presence of puromycin (Sigma). Permanently transfected cell lines were induced for pri-miR-BART6 expression with 2 μg/ml doxycycline (Sigma). Transfection efficiency and expression of pri-miRNA were determined by turboRFP expression. Protein and total RNA were extracted 48 h after DOX induction. Levels of mature miR-BART6-5p were examined by dideoxyoligonucleotide/primer-extension assay.

miRISC Loading Assay—The target probes were 5′-end ^{32}P-labeled with T4 polynucleotide kinase (New England Biolabs) and [γ-^{32}P]ATP. 5 fmol of ^{32}P-labeled miR-BART6-5p target RNA (5′-AACCUACUAUGGAUUGGACCAACCUUACCAAG-3′), BART6–3P-unedited target (5′-AACCUAAGCUAAGGCUAGUCCGAUCCCGCCAAG-3′), BART6–3P-edited target (5′-AACCUAGCCAAGGCUAGUCCGAUCCCCGCCAAG-3′), and pre-miR-BART6 RNAs, which had been cleaved from pri-miR-BART6 RNAs with Drosha-DGCR8 and gel-purified, were incubated with FLAG-tagged Ago2-complex made from permanently transfected HEK293 cells in a reaction buffer containing 1 unit/μl RNasin, 20 mM Tris-HCl (pH 7.6), 0.1 M NaCl, 10% glycerol, 2 mM DTT, 0.2 mM PMSF, 1 mM β-mercaptoethanol, 3.2 mM $MgCl_2$, 1 mM ATP, 20 mM creatine phosphate, and 1 units/μl creatine kinase at 37 °C for 90 min as described previously (43, 44). miRISC loading products (^{32}P-labeled cleaved target RNAs) were electrophoresed on a 15% polyacrylamide, 8 M urea gel, and quantified by Typhoon Imager.

RESULTS

A-to-I Editing Sites and Mutations Found in EBV Pri-miRNAs—We examined the primary transcripts of all 23 EBV miRNAs for A-to-I RNA editing in latently EBV-infected human lymphoblastoid GM607, Daudi Burkitt lymphoma, and C666-1 nasopharyngeal carcinoma cells. We found that pri-miR-BHRF1–1, pri-miR-BART6, pri-miR-BART8, and pri-miR-BART16 are edited at specific sites (Fig. 1*A* and supplemental Fig. 1*A*). Although the editing frequencies of pri-miR-BHRF1–1, pri-miR-BART8, and pri-miR-BART16 were relatively low (supplemental Fig. 1*B*), editing of pri-miR-BART6 in Daudi and GM607 cells at the +20 site reached 50 and 70%, respectively (Fig. 1*B*). Low levels of editing of pri-miR-BART6 RNAs were also detected in C666-1 cells (Fig. 1*B*). We found that the size of the end loop and the terminal stem of the pri-miR-BART6 of Daudi is smaller than that of GM607 cells (wild-type) due to deletion of three uridine nucleotides (Fig. 1*A*). The same deletion was detected in C666-1 cells, as reported previously (16).

Because involvement of enzymatically active ADAR1 and ADAR2 in the RNA editing mechanism has been established (23, 24, 45), we examined the expression of ADAR1 and ADAR2 in GM607, Daudi, and C666-1 cells by Western blotting analysis. Although no ADAR2 was detected, abundant expression of ADAR1 (both interferon-inducible p150 and constitutive p110 isoforms) (46) was found in all three cell lines (supplemental Fig. 2), indicating that ADAR1 is likely to be responsible for editing of pri-miR-BART6. However, we cannot exclude the possibility that ADAR2 may be also able to edit this site.

Processing of Pri-miR-BART6 Is Affected by Editing and Mutation—Many single nucleotide polymorphisms (SNPs) found in human miRNA genes affect biogenesis and function, suggesting

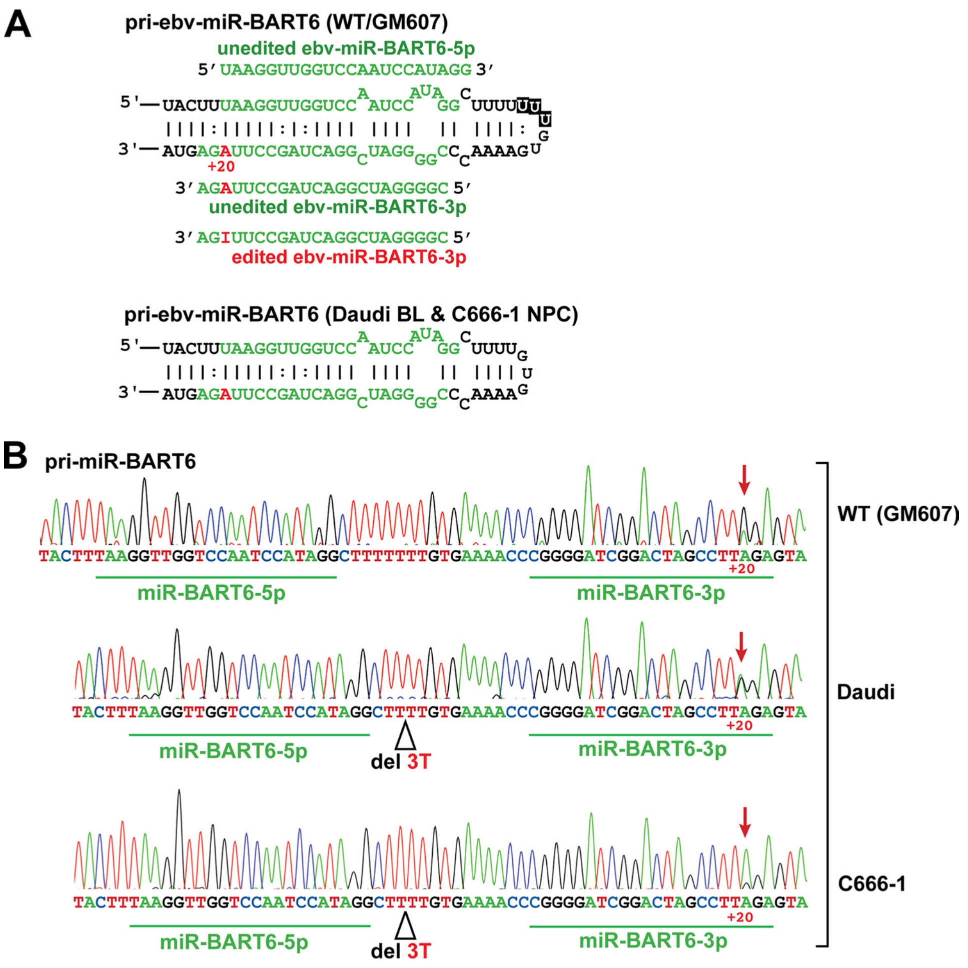

A
pri-ebv-miR-BART6 (WT/GM607)
unedited ebv-miR-BART6-5p
5′ UAAGGUUGGUCCAAUCCAUAGG 3′
+20
3′ AGAUUCCGAUCAGGCUAGGGGC 5′
unedited ebv-miR-BART6-3p
3′ AGIUUCCGAUCAGGCUAGGGGC 5′
edited ebv-miR-BART6-3p
pri-ebv-miR-BART6 (Daudi BL & C666-1 NPC)
B
pri-miR-BART6
WT (GM607)
Daudi
C666-1
miR-BART6-5p
miR-BART6-3p
del 3T
+20

FIGURE 1. **A-to-I RNA editing of pri-miR-BART6 RNAs.** *A*, shown are hairpin structures of pri-miR-BART6. Two different hairpin structures of pri-miR-BART6 (partial), the wild-type from GM607 cells and a mutant found in Daudi Burkitt and C666-1 cells, are shown. The editing site adenosine (+20 site), highlighted in *red*, is indicated by a *number* with the 5′ end of the mature miR-BART6–3p sequence counted as +1. The regions to be processed into the mature miRNAs (5p sense and 3p antisense strands) are highlighted in *green*. Mature miR-BART6-5p and both unedited and edited -3p RNAs are also shown. Three deleted U nucleotides are indicated in *black boxes* within the wild-type hairpin structure. *B*, DNA sequencing chromatograms of RT-PCR products derived from GM607, Daudi, and C666-1 pri-miR-BART6 RNAs are shown. The RNA editing site (+20) is detected as an A-to-G change in the cDNA sequencing chromatogram as indicated by *red arrows*. Three T nucleotides, deleted in pri-miR-BART6 from Daudi and C666-1 cells, are indicated. Editing frequency was estimated as a percentage estimated from the ratio of G peak over the sum of G and A peaks of the sequencing chromatogram. Two separate measurements were done, and identical results were obtained.

that they may be associated with diseases (47). Sequence variations in several EBV pri-miRNAs have also been reported (16), but the significance of most of these mutations has not been evaluated. We reasoned that editing at the +20 site and the mutations found in Daudi and C666-1 cells may affect the biogenesis of pri-miR-BART6 RNAs. An *in vitro* pri-miRNA processing assay using recombinant Drosha-DGCR8 and Dicer-TRBP complexes (33, 34, 36) was conducted with uniformly ^{32}P-labeled unedited and "edited" wild-type and Daudi (C666-1) pri-miR-BART6 RNAs, which were prepared by *in vitro* transcription. The edited pri-miRNAs had an A-to-G substitution at the +20 site. We had previously shown that the miRNA processing machinery recognizes A-to-G substitutions of pri-miRNAs as if they were A-to-I changes (34). The radioactive pri-, pre-, and mature miRNA products were quantitatively analyzed after fractionation on a polyacrylamide gel (Fig. 2, *A* and *B*). In addition, nonradioactive pri-miRNAs and their *in vitro* Dicer and/or Drosha cleavage assay products were also analyzed by Northern blotting using 5p- or 3p strand-specific oligonucleotide probes (Fig. 2*C*). The efficient conversion of both unedited and edited wild-type (GM607) pri-miR-BART6 to pre-miR-BART6 and mature miR-BART6 was detected, indicating that the editing of wild-type pri-miR-BART6 at the +20 site has no inhibitory effect on Drosha and Dicer cleavage (Fig. 2, *A* and *B*). Generation of both 5p and 3p mature miRNAs from unedited and edited wild-type pri-miR-BART6 was confirmed by Northern blotting analysis using strand-specific probes (Fig. 2*C*). Similarly, unedited Daudi (C666-1) pri-miR-BART6 RNAs were processed to pre- and mature miRNAs, although Dicer cleavage efficiency was reduced to ~70% of the unedited wild-type level, likely due to the deletion of three U residues (Fig. 2, *B* and *C*). However, Drosha cleavage of edited Daudi (C666-1) pri-miR-BART6 was completely blocked (Fig. 2, *A* and *C*). Binding of DGCR8 to Daudi (C666-1) pre-miR-BART6 seemed to be unaffected by editing, as seen from a set of electrophoresis mobility shift assay (EMSA) gels (supplemental Fig. 3). The nearly identical K_d values (~5 nM) for binding to unedited and edited pri-miR-BART6 RNAs were estimated from analysis of several EMSA gels. Thus, a combination of the deletion of three U residues and editing at the +20 site appears to inhibit Drosha cleavage of pri-miR-BART6 RNAs.

Targeting of Dicer by miR-BART6—The inhibitory effects of mutation and A-to-I editing on processing of pri-miR-BART6 RNAs into mature miRNAs may indicate that this viral miRNA plays a role in regulating EBV infection state in Daudi and C666-1 cells. For instance, suppression of miR-BART6 RNAs may be necessary for EBV to remain at a specific state of latency. Certain viral miRNAs have been shown to target genes of the host cell as well as genes of the virus itself (18). Using the DIANA-microT program (48) we, therefore, predicted *in silico* human and EBV target genes for miR-BART6-5p and miR-BART6–3p (both unedited and edited isoforms). The candidate target genes were pruned by a species conservation filter and also by accepting only genes that have multiple target sites within the 3′-UTRs. We found no strong target gene candidates containing multiple binding sites for

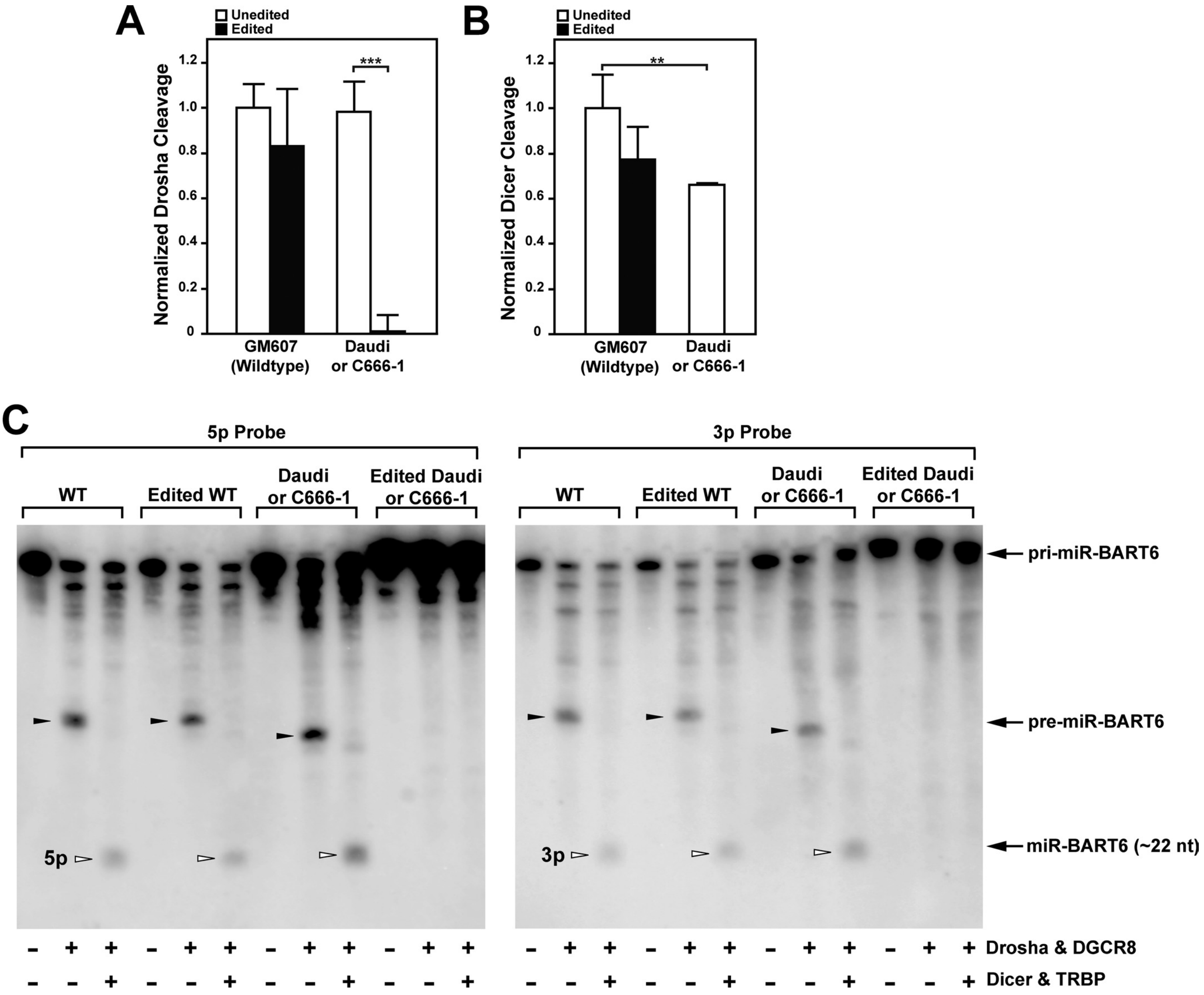


FIGURE 2. ***In vitro* processing of pri-miR-BART6 RNAs by miRNA processor complexes.** *A*, effect of editing on Drosha cleavage of wild-type and mutant pri-miR-BART6 RNAs was tested with uniformly ^{32}P-labeled pri-miR-BART6 RNAs. The mutant pri-miR-BART6 sequences of Daudi and C666-1 are identical. Thus, it is indicated as *Daudi or C666-1*. Unedited or edited pri-miR-BART6 RNAs (*i.e.* containing an A-to-G substitution at the +20 site) was subjected to the Drosha cleavage reaction using Drosha-DGCR8 complex. *B*, effect of editing on Dicer cleavage is shown. The Drosha-DGCR8 reaction products were subjected to the Dicer cleavage reaction using the Dicer-TRBP complex. *A* and *B*, three independent assays were done. Differences analyzed by Mann-Whitney *U* test: **, $p <$ 0.005; ***, $p < 0.001$. *Error bars*, S.E. ($n = 3$). *C*, Northern blotting analysis of *in vitro* processed miR-BART6 RNAs is shown. Nonradioactive pri-miR-BART6 RNAs processed *in vitro* by Drosha-DGCR8 and/or Dicer-TRBP complexes were analyzed by Northern blotting using a ^{32}P-labeled 5p- or 3p-strand specific oligo probe. Representative results for unedited and edited pri-miR-BART6 RNAs of wild-type (GM607) and mutant (Daudi) are shown.

miR-BART6-5p or -3p RNAs in the EBV genome. By contrast, the screening identified 14 human strong candidates for miR-BART6-5p and 3 human targets for miR-BART6–3p regardless of whether miR-BART6–3p RNAs were edited or unedited (supplemental Table 1). Because the +20 editing site of miR-BART6–3p is located outside of the seed sequence (Fig. 1*A*), it was anticipated that editing should not severely affect the selection of target genes.

Most interestingly, Dicer was one of the high-score targets for miR-BART6-5p (supplemental Table 1). Because of its importance and global effects on many genes via RNAi, we decided to further investigate the targeting of Dicer by miR-BART6-5p RNAs. Four target binding sites were identified within the ~1.5-kb region of human Dicer mRNA 3′-UTR (Fig. 3, *A* and *B*, and supplemental Experimental Procedures). Although a limited conservation of some 5p sites for elephant (site 1 and site 4) or armadillo (site 1 and site 2) was found, all four sites identified were otherwise unique to the human Dicer 3′-UTR and not evolutionarily conserved even for the chimpanzee Dicer 3′-UTR (data not shown). This is unusual for high score targets, which often have better species conservation, supporting their biological significance. In light of the fact that EBV specifically infects human, it is possible that miR-BART6-5p evolved to target Dicer specifically in human during the establishment of the EBV-host relationship.

In vitro validation experiments were conducted in HeLa cells (these cells are EBV-negative and, thus, lack pre-existing miR-BART6 RNAs) cotransfected with a luciferase

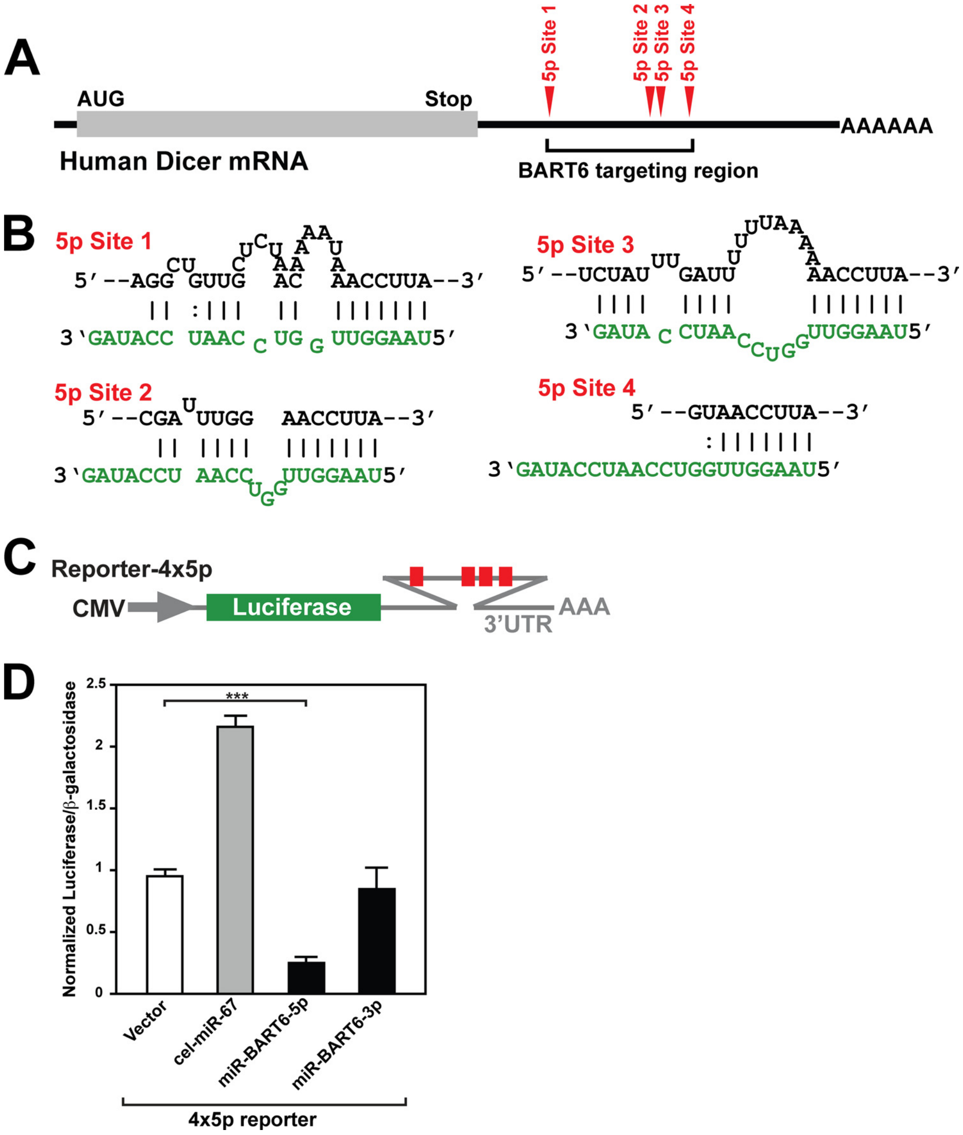


FIGURE 3. **Target sites for miR-BART6-5p RNAs identified in the 3′-UTR of human Dicer mRNA.** *A*, the locations of four miR-BART6-5p target sites located within the 3′-UTR of human Dicer mRNA are schematically presented. *B*, RNA duplex formation between the Dicer 3′-UTR target sites and miR-BART6-5p RNAs are diagrammed. *C*, shown is a diagram of the luciferase reporter plasmid containing the four 5p strand target sites. *D*, relative luciferase activities in HeLa cells cotransfected with the reporter vector containing 4 × 5p sites are shown. Two controls, the vector-only transfection, and cotransfection with the unrelated sequence *C. elegans* miR-67 were conducted. Expression levels of the luciferase reporter gene were normalized by expression levels of a cotransfected β-galactosidase reporter gene. Three independent assays were conducted. The luciferase activities were compared statistically by Mann-Whitney *U* tests. Significant differences between vector only and miR-BART6-5p or -3p cotransfected experiments are indicated by *asterisks*; ***, $p < 0.001$. *Error bars*, S.E. ($n = 3$).

reporter construct containing the Dicer 3′-UTR including the four target sites of miR-BART6-5p (reporter-4 × 5p) (Fig. 3*C*). The luciferase expression levels were clearly suppressed by miR-BART6-5p (4-fold) but not by miR-BART6–3p in HeLa cells that were cotransfected with the 4 × 5p vector (Fig. 3*D*). For an unknown reason, our negative control *Caenorhabditis elegans* miR-67 RNAs unexpectedly increased the luciferase expression (Fig. 3*D*). This is not due to nonspecific exhaustion of the miRNA-mediated silencing machinery by this control miRNA, as the other sequence-unrelated control miRNA, human miR-376a-5p, had no effect on the luciferase expression (data not shown). The results strongly indicate that the four Dicer mRNA 3′-UTR sites are indeed target sites of miR-BART6-5p RNAs. To validate the targeting of the Dicer mRNA by miR-BART6 RNAs *in vivo*, we then measured endogenous expression levels of Dicer in HeLa cells. We found a substantial reduction in the Dicer levels (3.5-fold) in HeLa cells transfected with miR-BART6-5p but not with control cel-miR-67 (Fig. 4*A*), confirming *in vivo* targeting of Dicer by miR-BART6-5p RNAs.

Suppression of miRISC Loading of miR-BART6-5p RNAs by Editing—To further confirm the *in vivo* silencing of Dicer by miR-BART6 RNAs, we prepared two tetracycline-inducible pri-miR-BART6 RNA expression constructs in a lentivirus vector system; one expressing unedited wild-type pri-miR-BART6 and the other expressing the edited pri-miR-BART6 containing an A-to-G substitution at the +20 editing site. HEK293 cells (also EBV-negative and, thus, lacking pre-existing miR-BART6-5p RNAs) were infected with the lentiviral constructs and subjected to conditional induction of pri-miR-BART6 and consequent mature miR-BART6 RNAs. Very low editing activities have been reported in HEK293 cells (49), and we confirmed that the pri-miR-BART6 RNAs derived from the unedited pri-miRNA expression construct were barely edited (<5%, data not shown). Dicer levels were reduced by 70% in HEK293 cells infected with the unedited pri-miR-BART6 construct compared with the vector control (Fig. 4*B*). The reduction in the Dicer levels was also detected in HEK293 cells infected with the edited pri-miR-BART6 construct. However, the extent of suppression was much less, by 25%. This may indicate that editing of the wild-type pri-miR-BART6 RNA has negative effects on the *in vivo* expression or functions of miR-BART6-5p RNAs, although no difference was noted between unedited and edited pri-miR-BART6 of wild-type (GM607) in *in vitro* pri-miRNA processing (Fig. 2). No significant difference in miR-BART6-5p RNA levels was detected between HEK293 cells infected with the unedited and edited pri-miR-BART6 expression constructs (supplemental Fig. 4), indicating that the stability and/or turnover of the mature miR-BART6-5p RNAs is unlikely to be affected by editing.

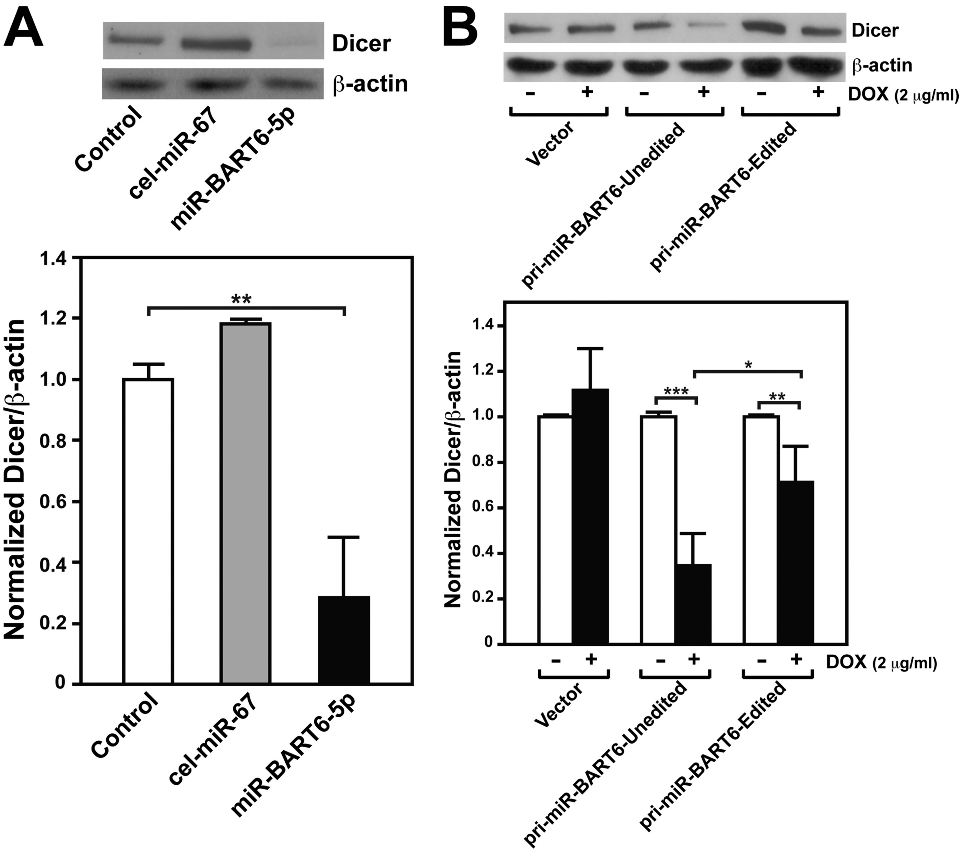


FIGURE 4. **Repression of Dicer by miR-BART6-5p RNAs.** *A*, Western blot analysis of Dicer expression levels in HeLa cells transfected with miR-BART6-5p RNAs is shown. Two control experiments were conducted; HeLa cells without transfection or transfected with a sequence-unrelated *C. elegans* miR-67. As a normalization control, β-actin levels were also monitored. A summary graph of normalized Dicer expression levels is also presented. *B*, shown is a Western blot analysis of Dicer expression in HEK293T cells infected with inducible lentivirus vectors for expression of unedited or edited (A-to-G substitution at the +20 site) pri-miR-BART6 RNAs. Expression of pri-miR-BART6 RNAs was induced with 2 μg/ml doxycycline (*DOX*). In the presence of doxycycline, the control vector directs the expression of non-silencing verified negative siRNAs (Open Biosystems). A summary graph of normalized Dicer expression levels is also shown. *A* and *B*, significant differences were analyzed by Mann-Whitney *U* tests: *, $p < 0.05$; **, $p < 0.005$; ***, $p < 0.001$. *Error bars*, S.E. ($n = 3$).

As one of the remaining possibilities, we thought that editing might affect the selection and loading of the "effective" strand onto the miRISC complex (50). We, therefore, examined the assembly of functional miRISC from recombinant FLAG-tagged Ago-2 complexed with Dicer and TRBP and either unedited or edited wild-type pre-miR-BART6 RNAs (Fig. 5*A*) as described previously (43, 44). We found that formation of the functional miRISC and consequent silencing (cleavage) of the 5p target RNA was indeed much more efficient with unedited pre-miR-BART6 than with edited pre-miR-BART6 (Fig. 5, *B* and *C*). Loading of miR-BART6–3p strand RNAs and consequent cleavage of their target RNA were extremely inefficient (Fig. 5, *B* and *C*), indicating that miR-BART6-5p is the major effective strand. The cloning frequency for miR-BART6-5p and -3p RNAs in GM607 cells also confirmed that the 5p strand is the effective strand. No sequence variations in the 5′ end sequence of the 5p strand was noted among miR-BART6 clones, indicating that editing at the +20 site does not affect the Drosha cleavage site (data not shown). Together our results clearly demonstrate that editing of the wild-type pri-miR-BART6, although not affecting the processing to pre- and mature miRNAs, inhibits the overall silencing effects of miR-BART6 RNAs. This is the first example of A-to-I editing of a pri-miRNA affecting miRISC loading.

Suppression of Many miRNAs by miR-BART6-5p—The dramatic reduction in the Dicer expression mediated by miR-BART6-5p (Fig. 4) suggests that it may affect the biogenesis of miRNAs globally. We, therefore, examined the effects of Dicer suppression on expression of other miRNAs by miRNA array analysis. The miRNA levels were examined in HeLa cells with substantially reduced Dicer levels after transfection with miR-BART6-5p RNAs (Fig. 4*A*). Once again, HeLa cells were used because of the absence of preexisting miR-BART6 RNAs. This study revealed that levels of at least 69 miRNAs were significantly reduced, and 14 of these miRNAs showed more than 2-fold suppression (supplemental Fig. 5*A*). Synthesis of these miRNAs may be particularly sensitive to the Dicer concentration. Interestingly, the expression of three miRNAs, miR-196b-5p, miR-205–5p, and miR-624–5p (supplemental Fig. 5*B*), was increased. Although we do not have a confirmed explanation for up-regulation of these three miRNAs, one possibility is that the genes regulating the expression of these miRNAs may be controlled negatively by other miRNAs whose levels are reduced. Our results demonstrate that suppression of Dicer mediated by miR-BART6-5p RNAs affects the expression of a large number of miRNAs.

Modulation of the EBV Latency State by miR-BART6-5p RNAs—We then asked whether Dicer silencing by miR-BART6-5p RNAs could control the EBV infection state. To examine this possibility, we first examined the relative expression levels of miR-BART6-5p strand RNAs and Dicer among GM607, Daudi, and C666-1 cells by qRT-PCR. Much higher levels of miR-BART6-5p were detected in C666-1 cells, which have much less editing than GM607 and Daudi cells (Fig. 6*A*). The low levels of miR-BART6-5p in GM607 and Daudi cells are consistent with the high editing rate of pri-miR-BART6 RNAs in these cells (Fig. 1*B*), which affects their processing (Fig. 2), miRISC formation (Fig. 5*C*), and consequently the levels of mature miR-BART6-5p RNA. As expected, Dicer levels were lowest in C666-1 cells, in inverse relation to the miR-BART6-5p levels (Fig. 6*B*). Accordingly, we decided to explore the significance of Dicer repression by miR-BART6-5p RNAs in C666-1 cells. We first attempted to antagonize the miR-BART6-5p RNAs expressed in C666-1 cells by transfection of a miR-BART6-5p antagomir. As expected, the miR-BART6-5p

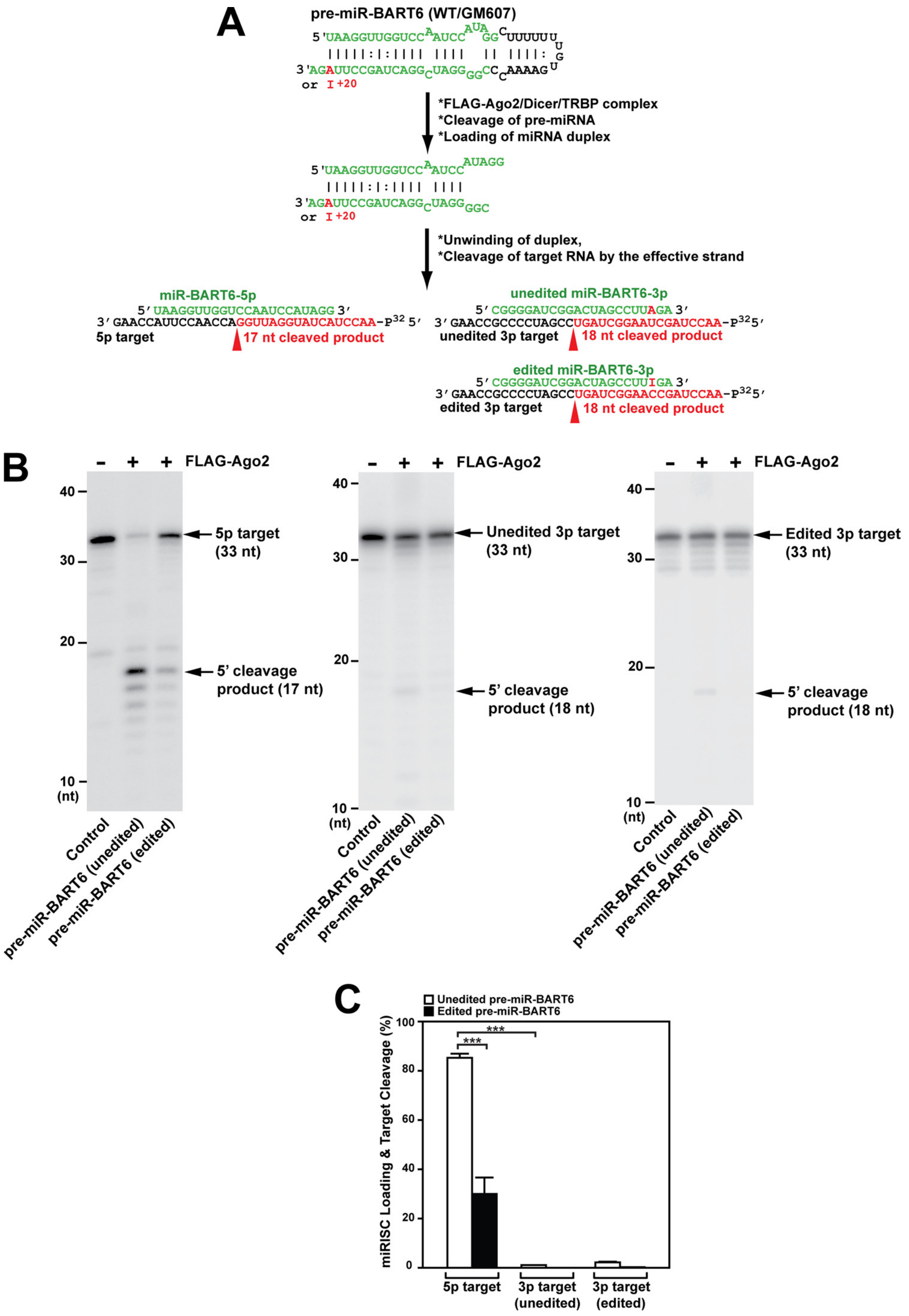


FIGURE 5. **Assembly of functional miRISCs with FLAG-Ago2 and pre-miR-BART6 RNAs.** *A*, a miRISC loading assay of pre-miR-BART6 is shown. Cleavage of the cognate target for miR-BART6-5p or -3p RNAs is schematically shown. The target RNA was 5′ ^{32}P-labeled. *B*, cleavage of the cognate target product (17 nucleotides) guided by miR-BART6-5p was substantially more efficient with unedited pre-miR-BART6 RNAs than with edited pre-miR-BART6 RNAs (*left panel*). Cleavage, although very inefficient, of both unedited and edited 3p target was detected only with unedited pre-miR-BART6 (*middle* and *right panels*). *C*, quantitative summary of miRISC loading experiments is presented. The cleavage efficiency was estimated by the ratio of the radioactivity of the correctly cleaved band over that of the uncleaved control band. Significant differences were analyzed by Mann-Whitney *U* tests: ***, $p < 0.001$. *Error bars*, S.E. ($n = 3$).

antagomir substantially decreased the miR-BART6-5p level (~20-fold) with a concomitant increase in the Dicer levels (~2-fold), indicating that miR-BART6-5p RNAs constantly suppress and maintain Dicer at the reduced levels in C666-1 cells (Fig. 7*A*). We then examined the relative expression levels of several genes known to be important for either lytic infection or the state of latency: EBNA1, EBNA2, LMP1, Zta, and Rta (12, 13). EBNA1 is detected in type I, II, and III latency, whereas EBNA2 and LMP1 are usually detected in type III latency (12, 13). EBNA2 is essential for the transformation of B lymphocytes and plays a central role in type III latency by up-regulating promoters of all latent EBV genes. Deficiency of the EBNA2 expression is known in type I and II latency (12, 13). A weak expression of LMP1 in type II latency and its deficiency in type I latency have been reported (12, 13). By contrast, Zta and Rta are essential for the initiation of the lytic EBV infection cycle (12, 13). By antagonizing miR-BART6-5p, Zta and Rta increased by 2–3-fold, indicating that miR-BART6-5p keeps these gene products under control. Furthermore, we noticed substantial up-regulation of EBNA2 oncogene expression (~5-fold) and LMP1 (~2-fold) by suppression of miR-BART6-5p RNAs, whereas no effects on EBNA1 were observed (Fig. 7*A*).

The activities of the three viral promoters Cp, Wp, and Qp were also monitored (Fig. 7*B*). Transcription from Cp and Wp is characteristic of type III latency (51), whereas the Qp promoter is used in EBV-infected cells undergoing type I or II latency (52, 53). We used qRT-PCR primers specific for RNAs initiating at Wp, Cp, or Qp (54). Significant up-regulation of type III latency-specific Cp and Wp promoter activities (5.4- and 11-fold, respectively) were detected in C666-1 cells transfected with miR-BART6-5p antagomir. On the other hand, Qp promoter activities associated with type I and type II latency were completely abolished; that is, not detectable in comparison to control.

We then attempted to silence Dicer directly by transfecting a Dicer targeting short hairpin RNA (shRNA) expression vector in C666-1 cells (Fig. 7*C*). This reduced Dicer levels by ~70%, indicating the efficiency of this Dicer targeting siRNA expression vector. As expected, changes in the marker genes were completely opposite to those noted in

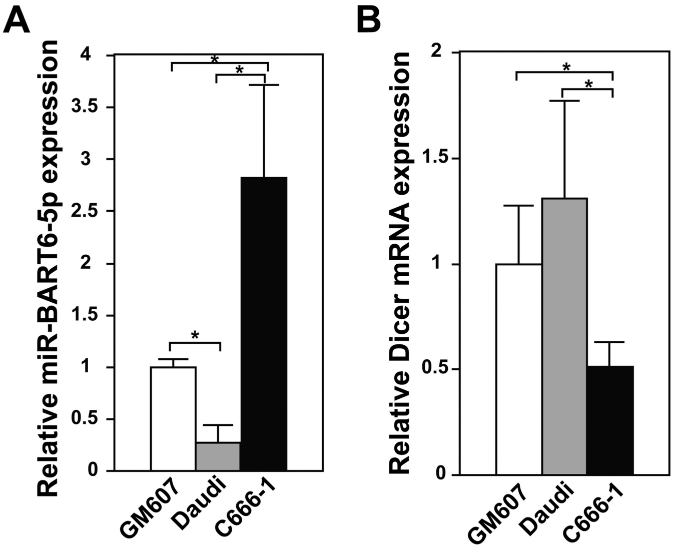


FIGURE 6. **Relative expression levels of miR-BART6-5p and Dicer in different cell lines.** *A*, miR-BART6-5p RNA levels were examined by qRT-PCR and normalized to β-actin mRNA level. Three independent assays were done. Significant differences were analyzed by Mann-Whitney *U* tests. *, $p < 0.05$. *Error bars*, S.E. ($n = 3$). *B*, Dicer mRNA levels were monitored by qRT-PCR and normalized to β-actin mRNA levels. Three independent assays were performed. Significant differences were analyzed by Mann-Whitney *U* tests. *, $p < 0.05$. *Error bars*, S.E. ($n = 3$).

C666-1 cells transfected with the miR-BART6-5p antagomir; that is, an ~7-fold reduction in the EBNA2 expression as well as substantial down-regulation for LMP1, Zta, and Rta (Fig. 7*C*), further confirming that control of these critical viral genes by miR-BART6-5p is mediated directly via its silencing effects on Dicer.

Finally, we examined genetically identical pairs of Burkitt lymphoma Mutu I and Mutu III cell lines, which are in type I and type III latency, respectively, to assess the function of miR-BART6-5p and Dicer silencing in B lymphoma cells (non- nasopharyngeal carcinoma cell lines). We found that the miR-BART6-5p level is higher in Mutu I than in Mutu III (supplemental Fig. 6*A*). By contrast, the Dicer level was lower in Mutu I than in Mutu III as expected (supplemental Fig. 6*B*). Low level A-to-I editing of pri-miR-BART6 RNAs was detected only in Mutu III cells, and no mutations were found in the miR-BART6 gene of Mutu I and Mutu III cells (data not shown). Thus, it is currently unknown why higher miR-BART6-5p expression is detected in Mutu I cells as compared with Mutu III cells. We first transfected Mutu I cells with miR-BART6-5p antagomir. As we observed in C666-1 cells, the antagomir effectively reduced miR-BART6-5p levels (~10-fold) and increased Dicer levels (1.9-fold). Furthermore, significant up-regulation of EBNA2, LMP1, Zta, and Rta, as well as Wp and Cp activation, was detected (Fig. 8, *A* and *B*). We then transfected Mutu III cells with the Dicer shRNA expression plasmid, which successfully repressed Dicer levels (~5-fold). Opposite effects of the miR-BART6-5p antagomir were detected; this, is, suppression of EBNA2, LMP1, Zta, and Rta (Fig. 8*C*). Up-regulation of Qp activities and down-regulation of Cp and Wp activities were also observed (Fig. 8*D*).

Together, these results suggest that Dicer suppression mediated via miR-BART6-5p RNAs maintains not only the type II latency of C666-1 cells but also the type I latency of Mutu I cells by suppressing lytic replication and also inhibiting transition of these cell lines to type III latency, a more immunoresponse-prone state of the viral infection cycle.

DISCUSSION

Editing Frequency of EBV miRNAs—A-to-I editing of a viral miRNA, KSHV-miR-K12-10 was first implicated because of identification of many cDNA clones corresponding to KSHV-miR-K12-10 RNAs containing an A-to-G substitution compared with the genomic sequence (55). Additional studies conducted more recently confirmed that this is indeed due to A-to-I editing at this site of the viral transcript harboring KSHV-miR-K12-10 by ADAR1 (56). Interestingly, the transcript could be processed into the viral miRNA as well as the mRNA coding for Kaposin A. A-to-I editing and consequent recoding of Kaposin A reduced its transforming activity (56). However, the significance of A-to-I editing of KSHV-miR-K12-10 RNAs remains unknown.

Apart from these reports on KSHV-miR-K12-10 RNAs, there has been no additional report on A-to-I editing of viral miRNAs. In this study we examined EBV miRNAs for A-to-I RNA editing in GM607 B lymphoblastoid cells, Daudi Burkitt lymphoma cells, and C666-1 nasopharyngeal carcinoma cells. We found that primary transcripts of four miRNAs, miR-BHRF1-I, miR-BART6, miR-BART8, and miR-BART16, undergo editing at specific sites. In view of ~20% of human miRNAs being subject to A-to-I editing (33), our findings of editing of 4 of 23 EBV miRNAs indicate that both cellular and viral miRNAs are subject to editing at about the same frequency. Among four EBV pri-miRNAs, we focused on pri-miR-BART6, which is highly edited at the +20 site of the 3p strand side of the hairpin dsRNA structure.

Suppression of miR-BART6 Expression and miRISC Assembly by A-to-I Editing—We have shown previously that A-to-I editing of pri-miRNAs can suppress their processing to pre-miRNAs by inhibiting Drosha cleavage in the nucleus (34) or suppress processing of pre-miRNAs to mature miRNAs by inhibiting Dicer cleavage in the cytoplasm (35). Furthermore, in some cases A-to-I editing of pri-miRNAs resulted in expression of miRNAs with an altered (edited) seed sequence and consequent silencing of a set of genes different from those targeted by the unedited version miRNAs (36).

In vitro pri-miR-BART6 processing studies revealed that a combination of A-to-I editing at the +20 site, and the three-U-residue deletion mutation, as observed in Daudi Burkitt lymphoma and C666-1 nasopharyngeal carcinoma cells, blocks the Drosha cleavage step completely. Editing of wild-type pri-miR-BART6 RNAs did not affect their processing. However, loading of miR-BART6-5p onto the functionally active miRISC was substantially inhibited by A-to-I editing at the +20 site. Editing of pri-miR-BART6 RNAs reported in this study is the first example in which editing suppresses loading of miRNA onto miRISC.

Selection of miR-BART6-5p as an Effective Strand—It has been reported that the relative stabilities of the base pairs at the 5′ ends of the duplex consisting of two miRNA strands determine the selection of the effective strand, which is loaded onto

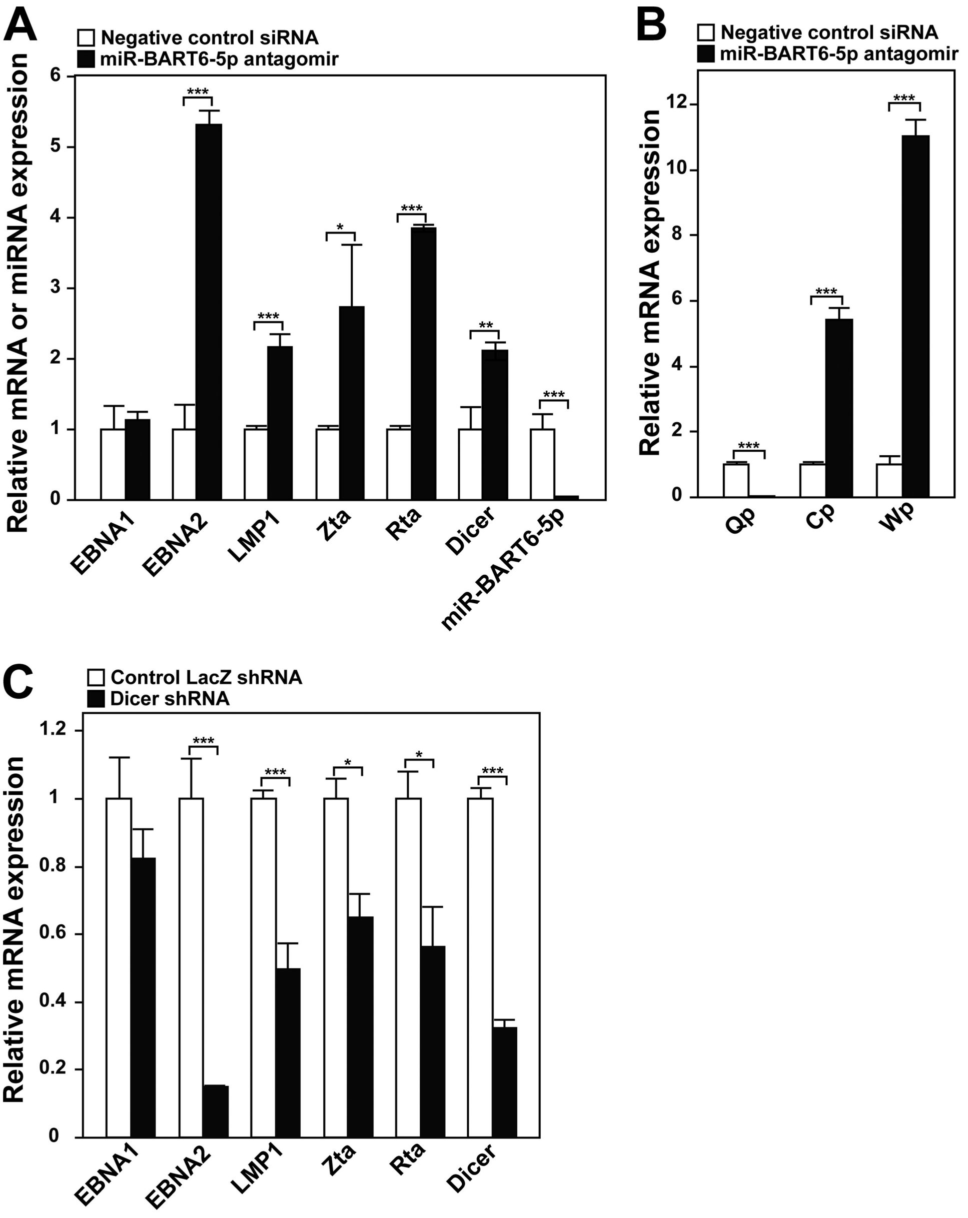


FIGURE 7. **Control of viral genes critical for the state of latency and lytic viral replication.** *A*, up-regulation of EBV genes critical for latency and viral replication by the miR-BART6-5p antagomir is shown. Expression of select viral genes including miR-BART6-5p in C666-1 cells transfected with the miR-BART6-5p antagomir or control (sequence unrelated Qiagen AllStars Negative Control siRNA) was examined by qRT-PCR. Three independent assays were done. Significant differences were analyzed by Mann-Whitney *U* tests. *, $p < 0.01$; **, $p < 0.005$; ***, $p < 0.001$. *Error bars*, S.E. ($n = 3$). *B*, shown are changes induced by the miR-BART6-5p antagomir in the viral promoters Qp, specific for the type I and type II latency, and Cp and Wp, specific for the type III latency. Transcripts initiated from Qp, Cp, and Wp were determined by qRT-PCR and compared with β-actin transcripts. Three independent assays were done. Significant differences were analyzed by Mann-Whitney *U* tests. **, $p < 0.005$; ***, $p < 0.001$. *Error bars*, S.E. ($n = 3$). *C*, repression of EBV genes after Dicer knock-down by shRNA. Expression of viral genes in C666-1 cells transfected with the Dicer targeting shRNA expression plasmid or control vector containing shRNA against LacZ was monitored by qRT-PCR. Three independent assays were conducted. Significant differences were analyzed by the Mann-Whitney *U* test. *, $p < 0.01$; ***, $p < 0.001$. *Error bars*, S.E. ($n = 3$).

miRISC and acts as the functional miRNA (50, 57). According to these studies, the strand whose 5′ end of the sense-antisense strand duplex is less stable is more frequently selected as the effective strand (50, 57). Interestingly, the miR-BART6 duplex consisting of 5p and 3p strands generated by Dicer cleavage predicts the selection of the 3p strand as the effective strand because of a relatively long fray present in the 5′ end of the 3p strand duplex (Fig. 5*A*). However, we noted that the 5p strand with the more stable 5′ end of the duplex was much more effectively loaded onto miRISC (Fig. 5*B*). More recently, major roles played by internal mismatched pairs in the selection of the effective strand for loading onto miRISC and also for unwinding of the duplex and consequent formation of the functional miRISC have been reported (58). According to the studies, central mismatches including G·U wobble pairs at positions 7–11 increase the formation of the miRNA duplex-miRISC, whereas the presence of an additional mismatch within the seed sequence and/or 3′-mid regions at positions 12–15 promotes unwinding of the duplex and formation of the mature functional miRISC containing the single-stranded effective miRNA (58). Interestingly, the miR-BART6-5p effective strand duplex contains these central (G·U at position 8), seed (G·U at position 6), and 3′-mid region (A·C at position 13) mismatched pairs (Fig. 5*A*), perhaps explaining at least partly why the 5p strand is more effective than the 3p strand.

Although the presence of an internal U·G or U·I wobble pair in place of a U·A Watson-Crick pair decreases the stability of the RNA duplex structure, a terminal U·G or U·I pair confers more stability, although subtle, to the RNA duplex than a U·A pair (59). Thus, replacement of a U·A base pair with U·I (U·G) wobble pair at the 5′ end of the 5p and 3p strand duplex due to editing at the +20 site is likely to increase the stability of the duplex, consistent with our observation that loading of miR-BART6-5p is much more efficient with unedited pre-miR-BART6 than with edited pre-miR-BART6 RNAs. Taken together, A-to-I editing at the +20 site suppresses the miRISC loading due to increased stability of the 5′ end of the 5p strand of the duplex.

Significance of Dicer Repression by miR-BART6 for the Viral Life Cycle—Most significantly, we provided evidence that miR-BART6-5p suppresses Dicer expression through binding to four target sites present within the 3′-UTR of the human Dicer mRNA. Interestingly, these four target sites are not conserved in the mouse or even chimpanzee Dicer mRNA, revealing that miR-BART6-5p RNAs target only human Dicer. In view of the

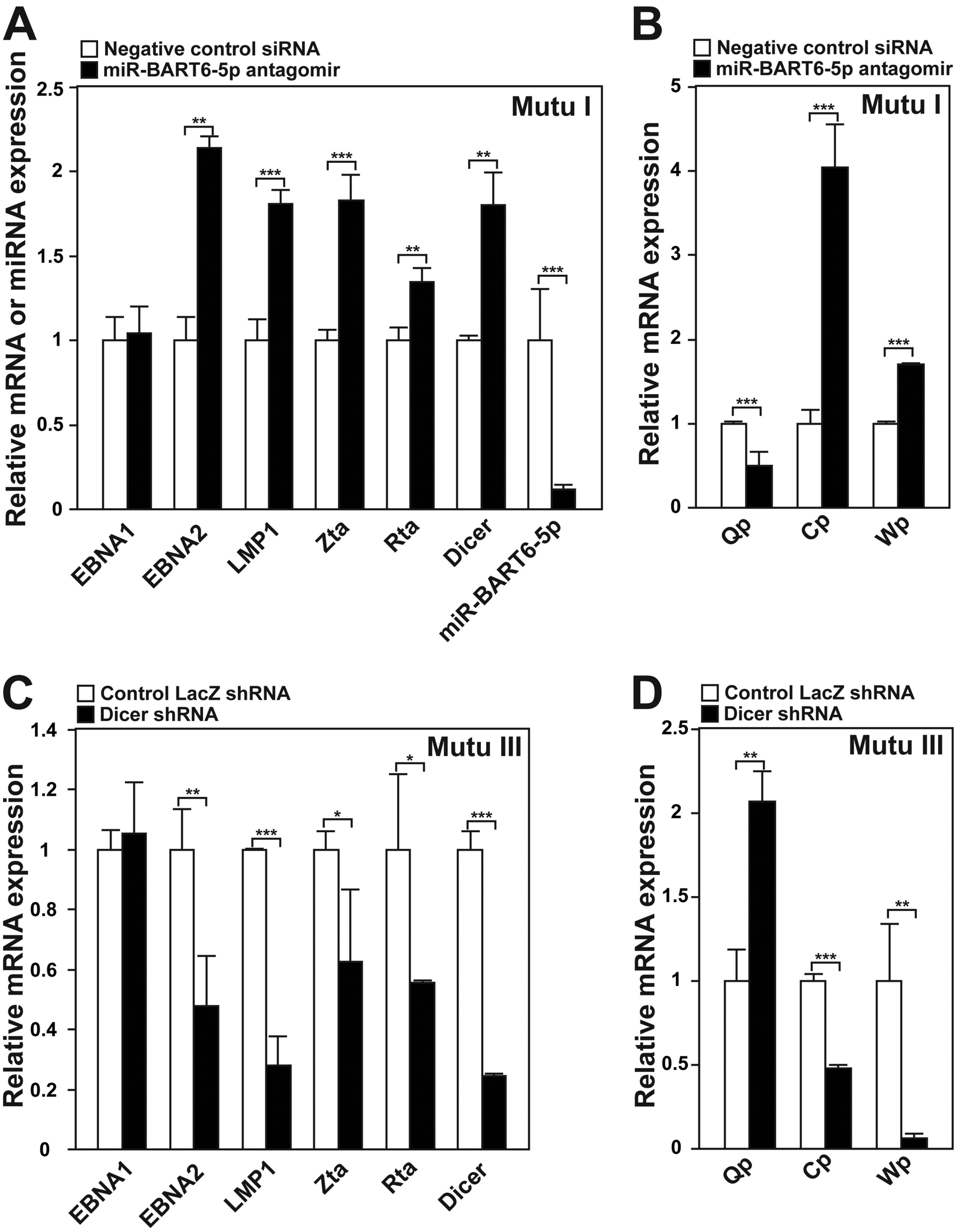


FIGURE 8. **The effects of miR-BART6-5p and Dicer silencing in B lymphoma cells.** *A*, shown is up-regulation of EBV genes critical for latency III and lytic replication in Mutu I cells transfected with the miR-BART6-5p antagomir or control siRNA. Dicer and miR-BART6-5p levels were also monitored. Three independent qRT-PCR assays were performed. Significant differences were analyzed by Mann-Whitney *U* tests. **, $p < 0.005$; ***, $p < 0.001$. *Error bars*, S.E. ($n = 3$). *B*, changes in viral promoter activities were induced in Mutu I cells by the miR-BART6-5p antagomir. Three independent qRT-PCR assays were done. Significant differences were analyzed by Mann-Whitney *U* tests. ***, $p < 0.001$. *Error bars*, S.E. ($n = 3$). *C*, down-regulation of EBV genes critical for latency III and lytic replication was detected in Mutu III cells transfected with the Dicer targeting shRNA expression plasmid or control vector containing shRNA against LacZ. Three independent qRT-PCR assays were done. Significant differences were analyzed by Mann-Whitney *U* tests. *, $p < 0.05$; **, <0.005; ***, $p < 0.001$. *Error bars*, S.E. ($n = 3$). *D*, changes in viral promoter activities were induced in Mutu III cells by Dicer knockdown. Three independent qRT-PCR assays were performed. Significant differences were analyzed by Mann-Whitney *U* tests. **, $p < 0.005$; ***, $p < 0.001$. *Error bars*, S.E. ($n = 3$).

EBV host specificity, *i.e.* EBV infections occur only in human, silencing of Dicer by miR-BART6-5p might have been established during the course of EBV evolution into a human-specific virus. It may be prudent to discount species conservation, usually used as one of the important parameters for target prediction programs, when target genes of a miRNA from a virus with narrow host range are screened.

Our miRNA array analysis revealed that Dicer suppression by miR-BART6-5p RNAs leads to suppression of many miRNAs. Because Dicer is required for processing of miR-BART6 itself, it is anticipated that a negative feedback loop may be made to tightly control Dicer and miR-BART6 as well as other viral and host cell miRNA levels. It has been reported that EBV infection of primary B cells results in a dramatic down-regulation of host cell miRNA expression, implying the presence of a suppressor of miRNA expression encoded by the virus (60). It was proposed that EBV may manipulate the expression of miRNAs as a major regulatory step in the viral life cycle, whereas host cells may potentially use miRNAs in response to EBV (18, 60). It appears that miR-BART6-5p likely is this viral miRNA suppressor and plays a critical role in the EBV virus life cycle by silencing Dicer and regulating the expression of miRNAs.

A global reduction in miRNA expression has been seen in many cancer cells (61). Suppression of Dicer by let-7 as well as by miR-103/107 and the consequent global reduction of miRNA synthesis have been reported (62–65). The cell proliferation rate is repressed by let-7, and it is proposed that let-7 acts as a master regulator of cell proliferation (66). Promotion of epithelial-to-mesenchymal transition and metastasis is controlled by miR-103/107 that down-regulates Dicer and consequently miR-200 (65). It appears that EBV has acquired miR-BART6 to mimic the powerful strategy of let-7 or miR-103/107 to down-regulate host cell miRNA production, which may be necessary to respond to host immune response and help EBV to stay in a specific state of latency and not initiate lytic viral replication. Suppression of miR-BART6-5p by antagomir indeed resulted in activation of EBNA2, LMP1, Zta, and Rta genes, critical for transition to type III latency or lytic replication, in Mutu I and C666-1 cells usually remaining in the less immune reactive type I and type II latency, respectively. In addition, the type III latency-specific Cp and Wp promoter activities were dramatically activated by miR-BART6-5p antagomir, whereas the type I and type II latency-specific Qp promoter activities were suppressed by the antagomir. We currently have no explanation how these promoter activities are up- or down-regulated. Involvement of B-cell-specific factors that activate the Wp promoter has been reported (67). On the other hand, many factors including E2F1, Rb, and LSD1 histone demethylase have been suggested to control the Cp promoter activities (68). Reduction of Dicer and consequent suppression of specific miRNAs that control these factors may be one possible mechanism to affect different viral promoter activities.

In conclusion, our results suggest the important roles played by EBV miR-BART6 RNAs in the regulation of viral replication and latency. Naturally occurring pri-miR-BART6 mutation and editing may be an adaptive selection to counteract the miR-BART6 function.

Acknowledgments—We thank L. Valente and W. Deng for providing purified Drosha-DGCR8, Dicer-TRBP, and Ago2-Dicer-TRBP complexes and J. M. Murray and B. Zinshteyn for comments.

REFERENCES

1. Bartel, D. P. (2004) *Cell* **116,** 281–297
2. Stefani, G., and Slack, F. J. (2008) *Nat. Rev. Mol. Cell Biol.* **9,** 219–230
3. Esquela-Kerscher, A., and Slack, F. J. (2006) *Nat. Rev. Cancer* **6,** 259–269
4. Kim, V. N., Han, J., and Siomi, M. C. (2009) *Nat. Rev. Mol. Cell Biol.* **10,** 126–139
5. Winter, J., Jung, S., Keller, S., Gregory, R. I., and Diederichs, S. (2009) *Nat. Cell Biol.* **11,** 228–234
6. Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., and Kim, V. N. (2003) *Nature* **425,** 415–419
7. Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F., and Hannon, G. J. (2004) *Nature* **432,** 231–235
8. Gregory, R. I., Yan, K. P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004) *Nature* **432,** 235–240
9. Lund, E., Güttinger, S., Calado, A., Dahlberg, J. E., and Kutay, U. (2004) *Science* **303,** 95–98
10. Chendrimada, T. P., Gregory, R. I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., and Shiekhattar, R. (2005) *Nature* **436,** 740–744
11. Förstemann, K., Tomari, Y., Du, T., Vagin, V. V., Denli, A. M., Bratu, D. P., Klattenhoff, C., Theurkauf, W. E., and Zamore, P. D. (2005) *PLoS Biol.* **3,** e236
12. Hislop, A. D., Taylor, G. S., Sauce, D., and Rickinson, A. B. (2007) *Annu. Rev. Immunol.* **25,** 587–617
13. Pagano, J. S., Blaser, M., Buendia, M. A., Damania, B., Khalili, K., Raab-Traub, N., and Roizman, B. (2004) *Semin. Cancer Biol.* **14,** 453–471
14. Pfeffer, S., Zavolan, M., Grässer, F. A., Chien, M., Russo, J. J., Ju, J., John, B., Enright, A. J., Marks, D., Sander, C., and Tuschl, T. (2004) *Science* **304,** 734–736
15. Cai, X., Schäfer, A., Lu, S., Bilello, J. P., Desrosiers, R. C., Edwards, R., Raab-Traub, N., and Cullen, B. R. (2006) *PLoS Pathog.* **2,** e23
16. Edwards, R. H., Marquitz, A. R., and Raab-Traub, N. (2008) *J. Virol.* **82,** 9094–9106
17. Grundhoff, A., Sullivan, C. S., and Ganem, D. (2006) *RNA* **12,** 733–750
18. Cullen, B. R. (2009) *Nature* **457,** 421–425
19. Barth, S., Pfuhl, T., Mamiani, A., Ehses, C., Roemer, K., Kremmer, E., Jäker, C., Höck, J., Meister, G., and Grässer, F. A. (2008) *Nucleic Acids Res.* **36,** 666–675
20. Lo, A. K., To, K. F., Lo, K. W., Lung, R. W., Hui, J. W., Liao, G., and Hayward, S. D. (2007) *Proc. Natl. Acad. Sci. U.S.A.* **104,** 16164–16169
21. Choy, E. Y., Siu, K. L., Kok, K. H., Lung, R. W., Tsang, C. M., To, K. F., Kwong, D. L., Tsao, S. W., and Jin, D. Y. (2008) *J. Exp. Med.* **205,** 2551–2560
22. Xia, T., O'Hara, A., Araujo, I., Barreto, J., Carvalho, E., Sapucaia, J. B., Ramos, J. C., Luz, E., Pedroso, C., Manrique, M., Toomey, N. L., Brites, C., Dittmer, D. P., and Harrington, W. J., Jr. (2008) *Cancer Res.* **68,** 1436–1442
23. Bass, B. L. (2002) *Annu. Rev. Biochem.* **71,** 817–846
24. Nishikura, K. (2006) *Nat. Rev. Mol. Cell Biol.* **7,** 919–931
25. Basilio, C., Wahba, A. J., Lengyel, P., Speyer, J. F., and Ochoa, S. (1962) *Proc. Natl. Acad. Sci. U.S.A.* **48,** 613–616
26. Jepson, J. E., and Reenan, R. A. (2008) *Biochim. Biophys. Acta* **1779,** 459–470
27. Athanasiadis, A., Rich, A., and Maas, S. (2004) *PLoS Biol.* **2,** e391
28. Blow, M., Futreal, P. A., Wooster, R., and Stratton, M. R. (2004) *Genome Res.* **14,** 2379–2387
29. Kim, D. D., Kim, T. T., Walsh, T., Kobayashi, Y., Matise, T. C., Buyske, S., and Gabriel, A. (2004) *Genome Res.* **14,** 1719–1725
30. Levanon, E. Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z. Y., Shoshan, A., Pollock, S. R., Sztybel, D., Olshansky, M., Rechavi, G., and Jantsch, M. F. (2004) *Nat. Biotechnol.* **22,** 1001–1005
31. Blow, M. J., Grocock, R. J., van Dongen, S., Enright, A. J., Dicks, E., Futreal, P. A., Wooster, R., and Stratton, M. R. (2006) *Genome Biol.* **7,** R27
32. Luciano, D. J., Mirsky, H., Vendetti, N. J., and Maas, S. (2004) *RNA* **10,** 1174–1177
33. Kawahara, Y., Megraw, M., Kreider, E., Iizasa, H., Valente, L., Hatzigeorgiou, A. G., and Nishikura, K. (2008) *Nucleic Acids Res.* **36,** 5270–5280
34. Yang, W., Chendrimada, T. P., Wang, Q., Higuchi, M., Seeburg, P. H., Shiekhattar, R., and Nishikura, K. (2006) *Nat. Struct. Mol. Biol.* **13,** 13–21
35. Kawahara, Y., Zinshteyn, B., Chendrimada, T. P., Shiekhattar, R., and Nishikura, K. (2007) *EMBO Rep.* **8,** 763–769
36. Kawahara, Y., Zinshteyn, B., Sethupathy, P., Iizasa, H., Hatzigeorgiou, A. G., and Nishikura, K. (2007) *Science* **315,** 1137–1140
37. Kelly, G. L., Milner, A. E., Tierney, R. J., Croom-Carter, D. S., Altmann, M., Hammerschmidt, W., Bell, A. I., and Rickinson, A. B. (2005) *J. Virol.* **79,** 10709–10717
38. Cheung, S. T., Huang, D. P., Hui, A. B., Lo, K. W., Ko, C. W., Tsang, Y. S., Wong, N., Whitney, B. M., and Lee, J. C. (1999) *Int. J. Cancer* **83,** 121–126
39. Iwakiri, D., Sheen, T. S., Chen, J. Y., Huang, D. P., and Takada, K. (2005) *Oncogene* **24,** 1767–1773
40. Kiss, C., Nishikawa, J., Takada, K., Trivedi, P., Klein, G., and Szekely, L. (2003) *Proc. Natl. Acad. Sci. U.S.A.* **100,** 4813–4818
41. Maruo, S., Nanbo, A., and Takada, K. (2001) *J. Virol.* **75,** 9977–9982
42. Cheng, A. M., Byrom, M. W., Shelton, J., and Ford, L. P. (2005) *Nucleic Acids Res.* **33,** 1290–1297
43. Gregory, R. I., Chendrimada, T. P., Cooch, N., and Shiekhattar, R. (2005) *Cell* **123,** 631–640
44. Maniataki, E., and Mourelatos, Z. (2005) *Genes Dev.* **19,** 2979–2990
45. Melcher, T., Maas, S., Herb, A., Sprengel, R., Seeburg, P. H., and Higuchi, M. (1996) *Nature* **379,** 460–464
46. Samuel, C. E. (2001) *Clin. Microbiol. Rev.* **14,** 778–809
47. Sun, G., Yan, J., Noltner, K., Feng, J., Li, H., Sarkis, D. A., Sommer, S. S., and Rossi, J. J. (2009) *RNA* **15,** 1640–1651
48. Maragkakis, M., Reczko, M., Simossis, V. A., Alexiou, P., Papadopoulos, G. L., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., Vergoulis, T., Koziris, N., Sellis, T., Tsanakas, P., and Hatzigeorgiou, A. G. (2009) *Nucleic Acids Res.* **37,** W273–W276
49. Dabiri, G. A., Lai, F., Drakas, R. A., and Nishikura, K. (1996) *EMBO J.* **15,** 34–45
50. Khvorova, A., Reynolds, A., and Jayasena, S. D. (2003) *Cell* **115,** 209–216
51. Pearson, G. R., Luka, J., Petti, L., Sample, J., Birkenbach, M., Braun, D., and Kieff, E. (1987) *Virology* **160,** 151–161
52. Schaefer, B. C., Strominger, J. L., and Speck, S. H. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92,** 10565–10569
53. Smith, P. R., and Griffin, B. E. (1992) *J. Virol.* **66,** 706–714
54. Luo, B., Wang, Y., Wang, X. F., Liang, H., Yan, L. P., Huang, B. H., and Zhao, P. (2005) *World J. Gastroenterol.* **11,** 629–633
55. Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grässer, F. A., van Dyk, L. F., Ho, C. K., Shuman, S., Chien, M., Russo, J. J., Ju, J., Randall, G., Lindenbach, B. D., Rice, C. M., Simon, V., Ho, D. D., Zavolan, M., and Tuschl, T. (2005) *Nat. Methods* **2,** 269–276
56. Gandy, S. Z., Linnstaedt, S. D., Muralidhar, S., Cashman, K. A., Rosenthal,

L. J., and Casey, J. L. (2007) *J. Virol.* **81,** 13544–13551

57. Schwarz, D. S., Hutvágner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P. D. (2003) *Cell* **115,** 199–208
58. Kawamata, T., Seitz, H., and Tomari, Y. (2009) *Nat. Struct. Mol. Biol.* **16,** 953–960
59. Strobel, S. A., Cech, T. R., Usman, N., and Beigelman, L. (1994) *Biochemistry* **33,** 13824–13835
60. Godshalk, S. E., Bhaduri-McIntosh, S., and Slack, F. J. (2008) *Cell Cycle* **7,** 3595–3600
61. Kumar, M. S., Lu, J., Mercer, K. L., Golub, T. R., and Jacks, T. (2007) *Nat. Genet.* **39,** 673–677
62. Forman, J. J., Legesse-Miller, A., and Coller, H. A. (2008) *Proc. Natl. Acad. Sci. U.S.A.* **105,** 14879–14884
63. Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008) *Nature* **455,** 58–63
64. Tokumaru, S., Suzuki, M., Yamada, H., Nagino, M., and Takahashi, T. (2008) *Carcinogenesis* **29,** 2073–2077
65. Martello, G., Rosato, A., Ferrari, F., Manfrin, A., Cordenonsi, M., Dupont, S., Enzo, E., Guzzardo, V., Rondina, M., Spruce, T., Parenti, A. R., Daidone, M. G., Bicciato, S., and Piccolo, S. (2010) *Cell* **141,** 1195–1207
66. Johnson, C. D., Esquela-Kerscher, A., Stefani, G., Byrom, M., Kelnar, K., Ovcharenko, D., Wilson, M., Wang, X., Shelton, J., Shingara, J., Chin, L., Brown, D., and Slack, F. J. (2007) *Cancer Res.* **67,** 7713–7722
67. Tierney, R., Nagra, J., Hutchings, I., Shannon-Lowe, C., Altmann, M., Hammerschmidt, W., Rickinson, A., and Bell, A. (2007) *J. Virol.* **81,** 10092–10100
68. Tempera, I., and Lieberman, P. M. (2010) *Biochim. Biophys. Acta* **1799,** 236–245

5.4. microRNA targeting in coding regions: a computational and experimental study of functionality

In the following manuscript we present the work regarding miRNA targeting in the CDS. Chi *et al* (Chi, Zang et al. 2009) released a set of biological data which allowed the development of a miRNA target prediction program based on machine learning techniques. We used these data and the data of Hafner *et al* (Hafner, Landthaler et al. 2010) for the development of another release of microT denoted as microT-CDS. It is a miRNA target prediction program that stands out not only because it is a purely data driven approach but also because it succeeds in assessing miRNA targeting both in the 3'UTR and the coding sequence of genes. Importantly, we show that targeting in the coding sequence is not only functional but also confers an important biological meaning since evolutionary pressure might enforce the presence of sites on the CDS in cases when there is restricted targeting space in the 3'UTR. Also we show that inclusion of targets in the coding sequence increases prediction sensitivity by more than 10% while also increasing prediction precision.

microRNA targeting in coding regions: a computational and experimental study of functionality

Martin Reczko[1,2,*], Manolis Maragkakis[1,3,*], Panagiotis Alexiou[1,4], Ivo Grosse[3]
and Artemis G. Hatzigeorgiou[1,5,§]

[1]Institute of Molecular Oncology, Biomedical Sciences Research Center ‘Alexander Fleming’, Vari, Greece, [2]Synaptic Ltd., Heraklion, Greece, [3]Institute of Computer Science, Martin Luther University Halle-Wittenberg, 06120 Halle, Germany, [4]School of Biology, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece, [5]Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, USA

[*]These authors contributed equally to this work
[§]Corresponding author

Corresponding author contact information: Hatzigeorgiou A., email: hatzigeorgiou@fleming.gr

Abstract

Experimental evidence has accumulated showing that microRNA binding sites within protein coding sequences are functional in controlling gene expression. Here we report a computational analysis of such miRNA target sites, based on features extracted from existing high throughput immunoprecipitation and sequencing data. The analysis is performed independently for the coding sequence and the 3’UTR and reveals different sets of features and models for the two regions. The two models are combined into a novel computational model for microRNA target genes, DIANA microT-CDS, which achieves significantly higher sensitivity at a similar precision level compared to other widely used programs and a model that implements target sites only in the 3’UTR. Importantly, further analysis indicates that genes with shorter 3’UTRs are preferentially targeted in the coding sequence.

Introduction

MicroRNAs (miRNAs) are small endogenous RNA molecules that play a key role in development and diseases through post-transcriptional regulation of gene expression. They are part of the RNA-Induced Silencing Complex (RISC) and guide it to specific miRNA Recognition Elements (MREs) on the mRNA molecules of target genes and lead either to translational repression and/or messenger RNA (mRNA) degradation (1).

Although most of the MREs have been found in the 3' Untranslated Region (3'UTR) of protein coding genes, there are individual reports of MREs located in the coding sequence (CDS) of target genes with evidence for their relation to important biological functions (2). In Duursma et al. (3) it is shown that miR-148 represses specific splice variants of DNA methyltransferase 3b (Dnmt3b) gene expression by targeting its coding sequence reporting that this mechanism might play a role in determining the relative abundance of different splice variants. Forman et al. (4) suggest that four let-7 miRNA target sites within the coding sequence of the miRNA-processing enzyme Dicer establish a mechanism for a miRNA/Dicer auto-regulatory feedback loop. In Elcheva et al. (5) it is shown that the coding region of beta-transducin repeat containing protein 1 is regulated by miR-183. Takagi et al. (6) showed that Hepatocyte Nuclear Factor 4 alpha (HNF4a) is down-regulated by miR-24 targeting its coding region. The expression of miR-24 is regulated by cellular stress, thus affecting the metabolism and cellular biology. Finally, Abdelmohsen et al. (7) showed that miR-519 represses the translation of the RNA-binding protein Hu antigen R (HuR) which in turn reduces HuR-regulated gene expression and cell division.

Also recently, high throughput data allowed for the direct identification of MREs on the target genes (8-9). In Hafner et al. through immunoprecipitation of the miRNA containing ribonucleoprotein complexes and sequencing of the associated RNA fragments (PAR-CLIP), it is shown that miRNAs tend to bind in approximately equal proportions on the 3'UTR as well as on the CDS of target mRNAs. The authors also suggest that miRNA targeting in the CDS has indeed a measurable effect on miRNA mediated mRNA degradation.

Up to now, most miRNA target prediction programs limit their search for MREs only within the 3'UTR (10). Here we present a novel approach that allows us for the first time to refine miRNA targeting both on the 3'UTR and the CDS by modeling the potential interaction between these two targeting mechanisms. The method is based on the analysis of a verified set of MREs against a negative set of MREs as defined through the PAR-CLIP data from Hafner et al. and introduces several novel features that have an effect on miRNA targeting mechanism.

The analysis is performed independently for the MREs on the two gene regions (3’UTR and CDS) to account both for the possibility of differing targeting mechanisms as well as for the possibility of differing MRE functionality (Figure 1).

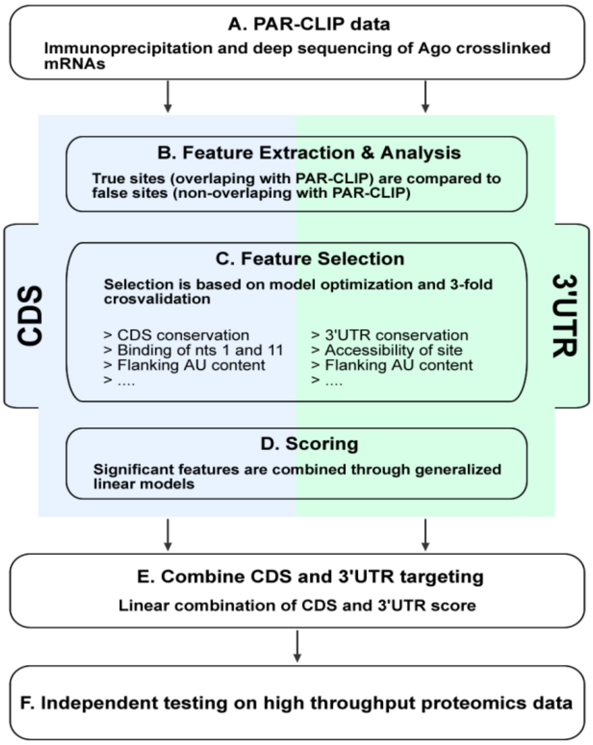

A. PAR-CLIP data
Immunoprecipitation and deep sequencing of Ago crosslinked mRNAs
B. Feature Extraction & Analysis
True sites (overlaping with PAR-CLIP) are compared to false sites (non-overlaping with PAR-CLIP)
CDS
3'UTR
C. Feature Selection
Selection is based on model optimization and 3-fold crosvalidation
> CDS conservation
> Binding of nts 1 and 11
> Flanking AU content
> ....
> 3'UTR conservation
> Accessibility of site
> Flanking AU content
> ....
D. Scoring
Significant features are combined through generalized linear models
E. Combine CDS and 3'UTR targeting
Linear combination of CDS and 3'UTR score
F. Independent testing on high throughput proteomics data

Figure 1: Flowchart of the analysis on the PAR-CLIP data. The MREs specified by the PAR-CLIP data are divided in two categories according to the genomic region in which they lie on (A). For these two sets several features are extracted and the most informative of them are selected by comparing true MREs with false MREs (B). The selection is performed through a three-fold crossvalidation model optimization (C). For each identified miRNA MRE the selected features (depending on the gene region it lies on) are combined into a MRE score through generalized linear models (D). For each gene we define the CDS score and the 3’UTR score which are calculated by summing the MRE scores that lie on each genomic region respectively. These two scores are linearly combined into a final score (E). To test for the overall performance of this scoring approach we performed a completely independent test on the high- throughput proteomics data of Selbach et al. (F).

Results

Target sites in coding regions result in significantly more sensitive target prediction.

The overall performance of the developed algorithm is tested on a data set completely independent from the PAR-CLIP training data. The test set is derived from the measurements of protein level changes after the transfection of five miRNAs in HeLa cells as provided by Selbach et al. (11). All genes with a logarithmic protein downregulation exceeding 0.2 are considered as targeted. Approximately half of these genes do not carry a single corresponding miRNA seed (nucleotides 2-7 from the 5'end of the miRNA) match in their 3'UTR sequences and are consequently not recognized by any computational miRNA target prediction program currently available. The combined (CDS & 3'UTR) model presented here increases sensitivity more than 12% from 52% to 65% in comparison to the 3'UTR region model, keeping specificity at the same level of 32% (see supplementary Fig. S1). To test the significance of the additional CDS model we compared the predicted results with a partly random predictor where for each miRNA, the scores of the two models are shuffled by combining the 3'UTR score of each gene with a randomly selected CDS score from a target gene of the same miRNA. The performance of this partly randomized predictor is significantly lower than the combined model (supplementary Fig. S1).

The combined model is additionally compared to other widely used miRNA target prediction programs such as TargetScan 5.0 (12), PicTar (13) and RNA22 (14) as well as a seed measure, whose predictions are defined through miRNA seed matches on the 3'UTR of genes and has been shown to be more sensitive than any other published prediction program (10) (see Materials and Methods). The seed predictions are scored based on the number of seed matches identified on each gene. Figure 2 reports the performance of the above mentioned programs at different levels of sensitivity. At lower specificity values a high sensitivity increase is observed, outperforming also the seed measure.

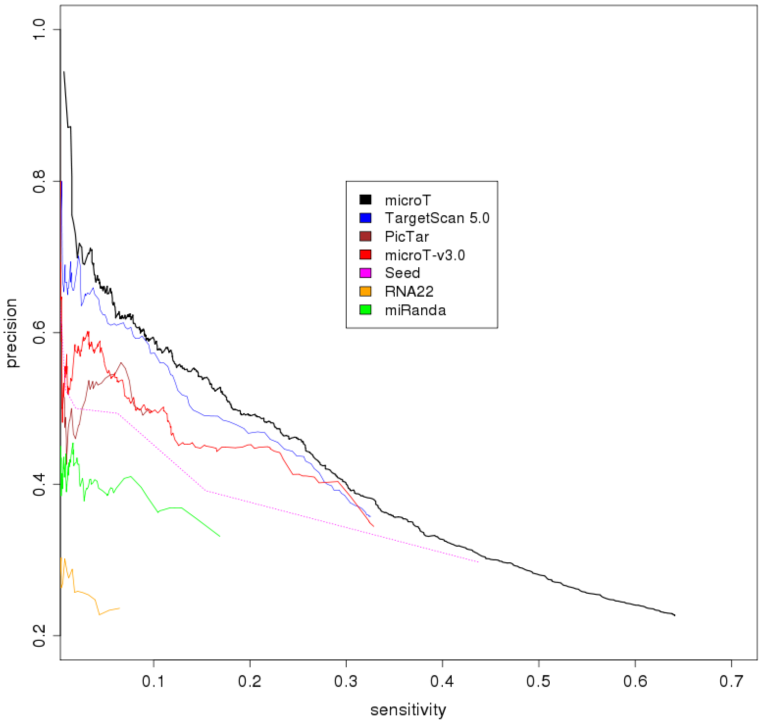


Figure 2: Precision of the predictions for different target prediction methods for increasing sensitivities (pROC analysis), tested on the data set from (Selbach et al., 2008)

In order to investigate the significance of the improvement of DIANA-microT-CDS to the next best performing program (TargetScan 5.0) we sample out of the 16164 measured miRNA:gene interactions in the

Selbach et al. data a random subsets of 8000 interactions each. A statistically significant improvement of prediction precision (p-value < 10^{-11}, Wilcoxon test) is found for the combined model in comparison to TargetScan on these 100 randomly collected gene sets. Measuring the area under the receiver operating curve (AUC), a significantly better average AUC value of 0.668 for microT in comparison to 0.615 for TargetScan is observed (p-value < 10^{-15}, Wilcoxon test). Evaluating in the same way the average protein downregulation of the predicted targets, we find a significantly higher downregulation for the predictions of microT-CDS compared to TargetScan 5.0 (p-value < 10^{-12}, Wilcoxon test).

The overlap between the targets predicted by the two programs on the same dataset (Selbach et al.) is ranging between 50% and 70% at a specific precision level (Figure 3), which implies that large number of correct targets is predicted only by microT-CDS. Particularly at lower precision levels the number of correct predictions is almost doubled using microT-CDS. A test of the microT-CDS algorithm on the five mentioned individual cases of experimentally verified CDS targeting returns three positive cases (for the genes: Dnmt3b,Dicer and HNF4a). This is in agreement with our estimated sensitivity and is currently the only available computational prediction for this type of sites.

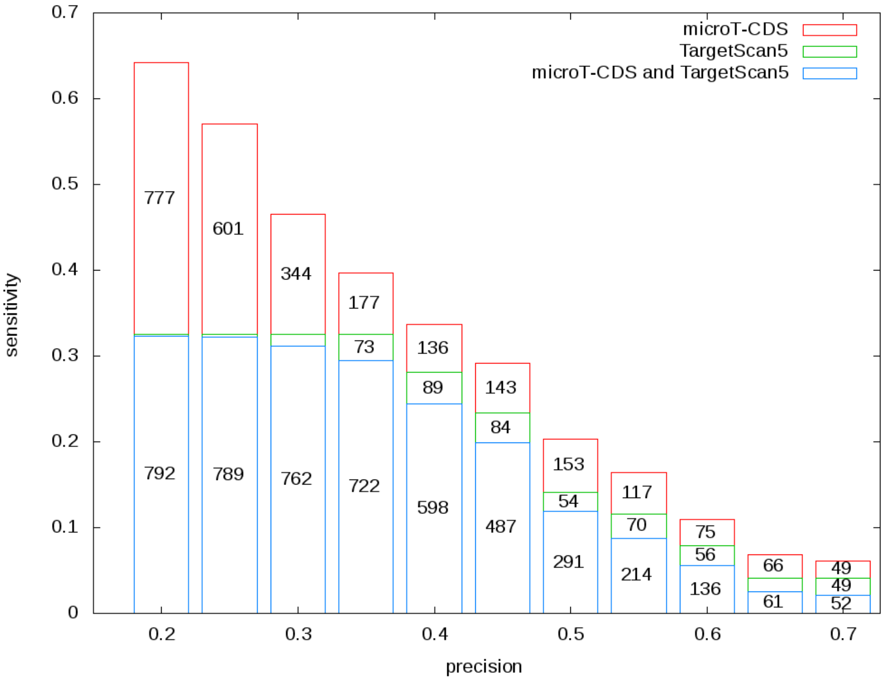


Figure 3: Comparison of the number of targets correctly predicted for the 3 sets: predicted only by microT-CDS, predicted only by TargetScan 5.0 and predicted by both programs. The comparison evaluates the 2447 known targets in the Selbach et al. dataset at specific score thresholds corresponding to different prediction precision levels.

Another independent test to evaluate the performance of our program in the detection of CDS target sites is performed on the high-throughput HITS-CLIP dataset of (Chi et al., 2009). The Argonaute-mRNA binding sites corresponding to mouse microRNA targets are used here. Of the top 20 expressed microRNAs in this experiment, 12 are not in the set of microRNAs used for the development of our algorithm. Of the genes targeted by these 12 microRNAs, 2156 have HITS-CLIP clusters only in the CDS and are not targeted in the 3'UTR. The seed location of 645 of these 2156 sites is correctly predicted by DIANA-microT-CDS. After multiple randomizations of the locations of the predicted sites, only 23.6 match to real binding sites. The ratio of true over randomly predicted sites is thus estimated as larger than 27.

Genes with shorter 3'UTR have significantly more targets in coding regions

To gain more insight into the mechanism underlying CDS targeting we investigated relations between CDS and 3'UTR targeting in the above mentioned dataset. Therefore we compared the CDS target scores with the

3' UTR length of the same target protein coding genes and could observe a significant preferred occurrence of MREs in the coding sequence for genes with 3'UTR sequences shorter than 500 nucleotides long (Wilcoxon test, p-value < 0.05) (Figure 4).

Such preference could not be observed for the group of genes that are measured as not targeted by miRNAs in the same proteomics experiment. We further tested the evidence of our observation by randomly combining the CDS scores of targeted and non targeted genes with the 3'UTR scores of the same group of genes respectively. We could again for both cases not observe a preference for CDS targeting in genes with short 3'UTRs as seen in the real genes score measurements (Figure 4). Similarly when analyzing the miRNA target genes as observed from 13 microarray experiments (see Materials and Methods) we observed that genes identified as targeted only on the CDS have significantly shorter 3'UTR sequences than genes targeted only on the 3'UTR (p-value < 10^{-13}, Wilcoxon test). These findings suggest that evolutionary pressure might enforce the presence of additional sites on the CDS in cases where there is restricted space on the 3'UTR.

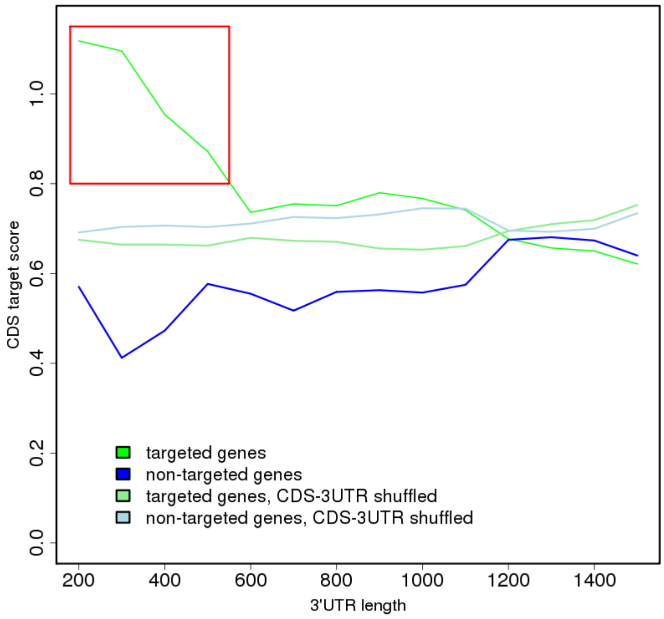


Figure 4: Preferential occurrence of MREs in coding sequence for short 3'UTRs. Comparing the sum of the predicted site scores in coding sequence (CDS score) against various 3'UTR sizes of targeted (green line) and non-targeted (blue line) genes on an independent test set reveals a significantly higher number of sites in coding regions for genes with 3'UTR lengths shorter than 500 nt (red box, Wilcoxon test, p-value < 0.05).

Discussion

High throughput proteomics experiments (11,15) that measure changes for thousands of genes both on the mRNA and the protein level reveal that approximately half of the genes whose expression is increased/decreased after miRNA transfection/knockout do not carry a single corresponding miRNA seed match in their 3'UTR sequence. The program introduced here enables the recognition of 12% of these down regulated genes as additional targets of miRNAs, having their targets in coding regions. The contribution of additional target sites located in the CDS is further verified in additional tests on the microarray experiments measuring the effect of over- or under expression of 6 miRNAs not contained in the training set. Comparing our algorithm when using only target sites in the 3'UTR with the algorithm using all sites on this data, the sensitivity of detecting verified targeted genes when using the same score cut-off is increased from 42.7% to 46.8% by more than 4%, while the false positive predictions and the precision of the predictions remains at the same level. This corresponds to 25 correctly predicted additional targets in this set of 600 verified targets.

The analysis of the recent data for miRNA associated protein immunoprecipitation and the subsequent RNA sequencing has been the base for the development of a program that uses several features which differ from those used by other miRNA target prediction programs. Generally, evolutionary conservation is a strong indication for MRE functionality (12,16-17). However, the coding sequences of genes usually have a significantly higher background conservation level than 3'UTR sequences due to their underlying amino acid content. We incorporate therefore a specific feature for conservation of MREs in coding regions, exploiting the conservation of synonymous codons.

The analysis described here reveals also that functional MREs in the CDS preferentially require a stronger binding than MREs in the 3'UTR. MREs in coding regions require a perfect binding along the miRNA seed region and mismatches disrupt their functionality. A feature analysis for MREs in 3'UTRs reveals a number of novel significant features, such as the requirement for increased accessibility in the mRNA secondary structure at the beginning of an MRE.

In several cases the synergistic effect of two features is shown to be more informative than the two features used independently. For example the higher mRNA AU content in the region surrounding an MRE (18) when combined with the free energy of the binding complex (p-value < 10-15, Wald test) gains higher significance than any of these features alone. Interestingly, this gain suffices to eliminate this AU content as an independent feature.

Also, it is shown that evolutionary pressure might enforce the presence of sites on the CDS in cases when there is restricted available space for targeting in the 3'UTR. All prediction results of microT-CDS are available through the DIANA web server (19) at www.microrna.gr/microT-CDS.

Methods

Datasets

PAR-CLIP data

The PAR-CLIP data (Figure 1A) is downloaded from the supplementary material of Hafner et al..

Microarray data

Microarray data are downloaded from ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae) and from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo). The data sets used are E-GEOD-12091 (20) (mir-26b), E-GEOD-12092 (20) (mir-98), E-GEOD-6207 (21) (miR-124), E-GEOD-958618 (miR-335), GSM155604 (22) (miR-106b), GSM210897 (18) (miR-7), GSM210898 (18) (miR-9), GSM210901 (18) (miR-122a), GSM210903 (18) (miR-128a), GSM210904 (18) (miR-132), GSM210909 (18) (miR-142), GSM210911 (18) (miR-148b), GSM210913 (18) (miR-181a).

Proteomics data

Changes in protein levels resulting from overexpressing miRNAs hsa-mir-1, hsa-mir16, hsa-mir30a, hsa-mir155 and hsa-let-7b (as estimated in Selbach et al.) are downloaded from http://psilac.mdc-berlin.de. RefSeq protein IDs are converted to corresponding Ensembl Gene IDs (Ensembl release 54). There are only 120 RefSeq protein IDs that corresponded to multiple Ensembl IDs, and of these 20 corresponded to multiple Ensembl IDs with different 3'UTR lengths. For these cases, the Ensembl ID corresponding to the longest 3'UTR is used. In total 16164 measurements for potential miRNA:mRNA interactions are identified, of which 2447 have a logarithmic protein downregulation exceeding 0.2 and are considered true targets and 13717 are considered false targets.

HITS-CLIP data

The HITS-CLIP data is downloaded from the supplementary material of Chi et al (9).

miRNA sequences

The miRNA sequences used are downloaded from miRBase Build 13.

Gene sequences

The CDS and 3'UTR sequences are the longest annotated transcript for each gene and are downloaded from Ensembl build 54.

Multiple Alignments

Multiple genome alignments are downloaded from UCSC Genome Browser. Human (hg18) alignment to the following 16 vertebrate genomes are used: panTro1, rheMac2, rn4, mm8, oryCun1, bosTau2, canFam2, dasNov1, loxAfr1, echTel1, monDom4, galGal2, xenTro1, tetNig1, fr1, danRer3 and Mouse (mm9) alignment to the following 16 vertebrate genomes are used: rn4, oryCun1, hg18, panTro2, rheMac, canFam, bosTau3, dasNov1, loxAfr1, echTel, monDom4, galGal3, xenTro2, tetNig, fr2, danRer5.

miRNA target prediction of other programs

For TargetScan we download the source code from http://www.targetscan.org/vert_50 and execute it for all human miRNAs against all 3'UTR sequences also used by microT to provide an accurate comparison with this program. The results obtained for TargetScan on our 3'UTR set perform slightly better than the results downloaded from the TargetScan server. The predictions for all other programs are derived from Alexiou et al. (10).

Feature Extraction

Alignment for putative MRE identification

A dynamic programming algorithm identifies the best alignment between the miRNA extended seed sequence and every 9 nucleotide window on the 3'UTR. The alignment is initially restricted so as the pairing of the miRNA extended seed with the 9 nt window begins at position 1 or 2 of the miRNA extended seed. A minimum of four consecutive Watson-Crick (WC) binding nucleotides is required starting at position 1 or 2 of the miRNA extended seed. A single G:U wobble pair is allowed for binding sites with more than 6 consecutive WC binding nucleotides. A single bulge or mismatch is allowed for binding sites with eight WC binding nucleotides.

Primary analysis of PAR-CLIP data and training set construction

The PAR-CLIP data produced in Hafner et al. consists of genomic coordinates specifying potential positions of MREs (8). Each putative MRE position is further refined through the existence of a T to C mutation in the sequenced tags as reported in the Hafner et al. To identify the miRNA involved in each MRE, the sequences of all identified genomic locations of the PAR-CLIP data are aligned against the miRNA sequence of the top 100 expressed miRNAs. These aligned locations are putative MREs and are further filtered to keep only those that are located closer than 5 nucleotides to the T to C mutation. In case there are more than one putative miRNA bindings in the same region then only the MRE with the higher number of binding nucleotides is retained. This set of MREs is defined as the true set. On the other hand, the false set consists of all aligned locations which do not overlap with the PAR-CLIP data. To take into account the probability part of the false set to correspond to miRNAs or genes which are indeed functional but are not expressed in the particular tissue of the PAR-CLIP experiment we have only retained aligned locations for the top 100 expressed miRNAs in the experiment and genes that already contained at least one true MRE. Overall, out of the 17310 PAR-CLIP peaks throughout the genome, 5075 overlap with an MRE in the 3'UTR and 6057 overlap with an MRE in the CDS.

Detection of binding categories with significant PAR CLIP reads enrichment

The binding category of a putative MRE is determined through the alignment procedures described above. All binding categories are then separated based on whether the mRNA nucleotide opposite the first nucleotide of the miRNA is an A or not and whether it is a matching nucleotide or not. This procedure defines 64 different binding categories which are then compared between the true and false set of MREs as defined in the PAR-CLIP data set (Figure 1B). This comparison is performed through a logistic regression between the binding categories and the presence or absence of the corresponding MRE in the true or false set of the PAR-CLIP data. The estimated regression coefficient (values in supplementary table S1) is thereafter used as a feature in the generalized linear models to characterize the overall efficiency of each MRE and is denoted as the "binding category weight" feature. An example category is labeled "8mer+3'pairing 1st:mismatch+NotA" and corresponds to 8 matches between the miRNA extended seed and the mRNA plus additional bindings in the 3'end and the first nucleotide opposite the 5' end of the mRNA is not a match nor is an Adenosine.

Conservation measure of the MRE sequence in CDS

The CDS conservation scoring method is based on a recently proposed approach (4) of calculating excess sequence conservation above the one required for amino acid conservation. The underlying concept is that functional MREs in the CDS are expected to preferentially conserve those nucleotides that would have no effect on the amino acid outcome, but would interfere with miRNA targeting (see Methods). The 30-way genomic alignments (UCSC) for the coding regions for all mRNAs have been downloaded and for each pairwise alignment of the reference species to any other species, we calculate the probabilities that the sequence of a triplet (or partial triplet) is conserved, given that the amino acid it codes for is or is not conserved. Each predicted MRE is scored by adding these probabilities from all pairwise alignments of the triplets (partial or full) that cover the MRE, normalized by the maximum score that could have been achieved by this same MRE. This feature is denoted as "CDS conservation".

Conservation measure of the MRE sequence in 3'UTRs

The algorithm assesses the evolutionary conservation of a MRE in the 3'UTR by calculating a conservation score based on 16 species. To compensate for the overall degree of conservation in the whole 3'UTR, the conservation score for each MRE is defined as the ratio of the number of species in which the binding positions of the extended seed region are conserved versus the respective number using the maximal number of species having any conservation in the whole 3'UTR region. This feature is denoted as "conservation".

Detection of significantly accessible locations within MREs

A logistic regression between the presence or absence of reads in the PAR-CLIP data and the accessibility of the 3'UTR sequence as calculated with the Sfold algorithm (23) using each of the 40 nucleotides upstream and 10 nucleotide downstream of the start of each MRE as a feature is performed to identify any significant targeting feature related to accessibility. The largest region with a reasonably significant contribution (p-value < 0.1, Wald test) and consistent direction of the contribution at all positions ranges across positions -1, 1 and 2. The sum of accessibilities in this region, denoted as "MRE accessibility (-1 to 2)", is used as a feature in the following.

Other MRE features

Two of the three features identified in Grimson et al. (18), the MRE flanking AU content denoted as "flanking AU content" and the distance of the MRE to the closest 3'UTR end denoted as "distance to closest 3'UTR end" are used. Additionally, the distance between adjacent MREs denoted as "adjacent MRE distance", the free energy of binding as calculated with RNAhybrid (24) denoted as "free energy" and the resulting binding pattern of the 29 nucleotides of the 3'UTR along the MRE denoted as "bnt1" to "bnt29", are also evaluated as features. All second order interactions between all features are automatically generated and selected using F-tests.

Feature selection

To determine an optimal feature set using crossvalidation, the PAR-CLIP data set is split into 3 disjoint subsets, stratified for positive and negative sites. On each subset a logistic regression using the features described above is performed and a feature selection procedure minimizing the Akaike information criterion (AIC) using the stepAIC implementation in the MASS (25) package for R determines an optimal set of features. For this initial set of features, the capability of each single feature to separate the complete PAR-CLIP data into sites with reads and sites without reads is tested using a Wilcoxon test and only features showing significant ($p\text{-value} < 0.05$) separation are retained. This feature selection procedure is performed independently for sites in the CDS and sites in the 3'UTR (Figure 1C). The full list of selected CDS and 3'UTR features is provided in supplementary table S2.

Training and scoring

Using the identified significant features, different machine learning methods like support vector machines, neural networks, random forests and generalized linear models (GLM) (25) are compared for the calculation of an MRE score. The best performance in crossvalidation is obtained using GLMs. Each gene region (CDS or 3'UTR) has a separate model. The regression coefficients for all features and their significances are in supplementary table S2. The scores for all MREs identified on a region are summed into a region score (Figure 1D).

Combining CDS and 3'UTR targeting

For the optimal combination of the two region scores we train another generalized linear model using data from the 13 different microarray experiments measuring mRNA expression changes when a miRNA is either transfected or knocked out (defined in the data section). The genes in each data set are sorted according to expression fold change compared to the control and the top and bottom 100 genes from each experiment are used as the true and false examples for training the generalized linear model (Figure 1E).

Acknowledgements

We thank Nectarios Koziris from the National Technical University of Athens for providing their computer cluster to conduct experiments and support all DIANA-microT-CDS web services.

References

1. Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, 136, 215-233.
2. Tay, Y., Zhang, J., Thomson, A.M., Lim, B. and Rigoutsos, I. (2008) MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, 455, 1124-1128.

3. Duursma, A.M., Kedde, M., Schrier, M., le Sage, C. and Agami, R. (2008) miR-148 targets human DNMT3b protein coding region. *RNA*, 14, 872-877.
4. Forman, J.J., Legesse-Miller, A. and Coller, H.A. (2008) A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. *Proc Natl Acad Sci U S A*, 105, 14879-14884.
5. Elcheva, I., Goswami, S., Noubissi, F.K. and Spiegelman, V.S. (2009) CRD-BP protects the coding region of betaTrCP1 mRNA from miR-183-mediated degradation. *Mol Cell*, 35, 240-246.
6. Takagi, S., Nakajima, M., Kida, K., Yamaura, Y., Fukami, T. and Yokoi, T. (2010) MicroRNAs regulate human hepatocyte nuclear factor 4alpha, modulating the expression of metabolic enzymes and cell cycle. *J Biol Chem*, 285, 4415-4422.
7. Abdelmohsen, K., Srikantan, S., Kuwano, Y. and Gorospe, M. (2008) miR-519 reduces cell proliferation by lowering RNA-binding protein HuR levels. *Proc Natl Acad Sci U S A*, 105, 20297-20302.
8. Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141, 129-141.
9. Chi, S.W., Zang, J.B., Mele, A. and Darnell, R.B. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460, 479-486.
10. Alexiou, P., Maragkakis, M., Papadopoulos, G.L., Reczko, M. and Hatzigeorgiou, A.G. (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, 25, 3049-3055.
11. Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R. and Rajewsky, N. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455, 58-63.
12. Friedman, R.C., Farh, K.K., Burge, C.B. and Bartel, D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19, 92-105.
13. Lall, S., Grun, D., Krek, A., Chen, K., Wang, Y.L., Dewey, C.N., Sood, P., Colombo, T., Bray, N., Macmenamin, P. *et al.* (2006) A genome-wide map of conserved microRNA targets in C. elegans. *Curr Biol*, 16, 460-471.
14. Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.S., Tam, W.L., Thomson, A.M., Lim, B. and Rigoutsos, I. (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, 126, 1203-1217.
15. Baek, D., Villen, J., Shin, C., Camargo, F.D., Gygi, S.P. and Bartel, D.P. (2008) The impact of microRNAs on protein output. *Nature*, 455, 64-71.
16. Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P. and Burge, C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, 115, 787-798.
17. Kiriakidou, M., Nelson, P.T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z. and Hatzigeorgiou, A. (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes Dev*, 18, 1165-1178.
18. Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P. and Bartel, D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 27, 91-105.
19. Maragkakis, M., Vergoulis, T., Alexiou, P., Reczko, M., Plomaritou, K., Gousis, M., Kourtis, K., Koziris, N., Dalamagas, T. and Hatzigeorgiou, A.G. (2011) DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association. *Nucleic Acids Res*.
20. Gennarino, V.A., Sardiello, M., Avellino, R., Meola, N., Maselli, V., Anand, S., Cutillo, L., Ballabio, A. and Banfi, S. (2009) MicroRNA target prediction by expression analysis of host genes. *Genome Res*, 19, 481-490.
21. Wang, X. (2006) Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic Acids Res*, 34, 1646-1652.
22. Linsley, P.S., Schelter, J., Burchard, J., Kibukawa, M., Martin, M.M., Bartz, S.R., Johnson, J.M., Cummins, J.M., Raymond, C.K., Dai, H. *et al.* (2007) Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Mol Cell Biol*, 27, 2240-2252.
23. Ding, Y., Chan, C.Y. and Lawrence, C.E. (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res*, 32, W135-141.
24. Rehmsmeier, M., Steffen, P., Hochsmann, M. and Giegerich, R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, 10, 1507-1517.
25. Venables, W. and Ripley, B. (2002) *Modern Applied Statistics with S.* Springer.

Supplementary Material

Supplementary Table 1: Significant miRNA recognition element (MRE) binding categories. A logistic regression between the binding categories of the aligned MREs and the presence or absence of the corresponding MRE in the true or false set of the PAR-CLIP data reveals 21 binding categories with a significant association for MREs in the 3'UTR and 10 binding categories for MREs in the CDS (p-value < 0.05, Wald test).

Binding category	**3'UTR**		**CDS**	
	Regession coefficient*	**p-value**	**Regession coefficient**	**p-value**
9mer 1st:match+notA	**1.9533**	2.00^{-16}	1.08279	1.19^{-10}
8mer 1st:match+notA	**1.9555**	2.00^{-16}	0.98775	2.00^{-16}
7mer 1st:match+notA	1.1084	2.00^{-16}		
6mer 1st:match+notA	0.7788	2.00^{-16}		
8mer 1st:mismatch+A	*2.2699*	2.00^{-16}		
7mer 1st:mismatch+A	**1.9852**	2.00^{-16}	0.68969	7.27^{-9}
6mer 1st:mismatch+A	**1.0179**	2.00^{-16}		
9mer 1st:match+A	**3.08**	2.00^{-16}		
8mer 1st:match+A	**2.7861**	2.00^{-16}	1.26186	2.00^{-16}
7mer 1st:match+A	**1.7453**	2.00^{-16}	0.76037	2.00^{-16}
8mer 1st:mismatch+NotA	**2.0928**	2.00^{-16}	1.08677	2.00^{-16}
7mer 1st:mismatch+NotA	**1.4572**	2.00^{-16}	0.3905	3.60^{-9}
6mer 1st:mismatch+NotA	0.7199	2.00^{-16}		
6mer 1st:match+A	**0.484**	8.67^{-10}		
9mer+3'pairing 1st:match+A			2.33208	2.47^{-2}
7mer+3'pairing 1st:mismatch+A	**5.1386**	2.78^{-5}		
7mer+3'pairing 1st:match+A	**4.4454**	7.16^{-5}		
8mer+wobble 1st:mismatch+NotA	1.1752	1.37^{-4}		
8mer+wobble 1st:match+A	**1.425**	1.68^{-3}	0.71923	3.33^{-2}
8mer+mismatch 1st:match+A	1.3601	7.19^{-3}		
8mer+mismatch 1st:match+NotA	1.4836	1.10^{-2}		
8mer+3'pairing 1st:mismatch+NotA	**2.3352**	2.16^{-2}	2.12444	3.60^{-3}

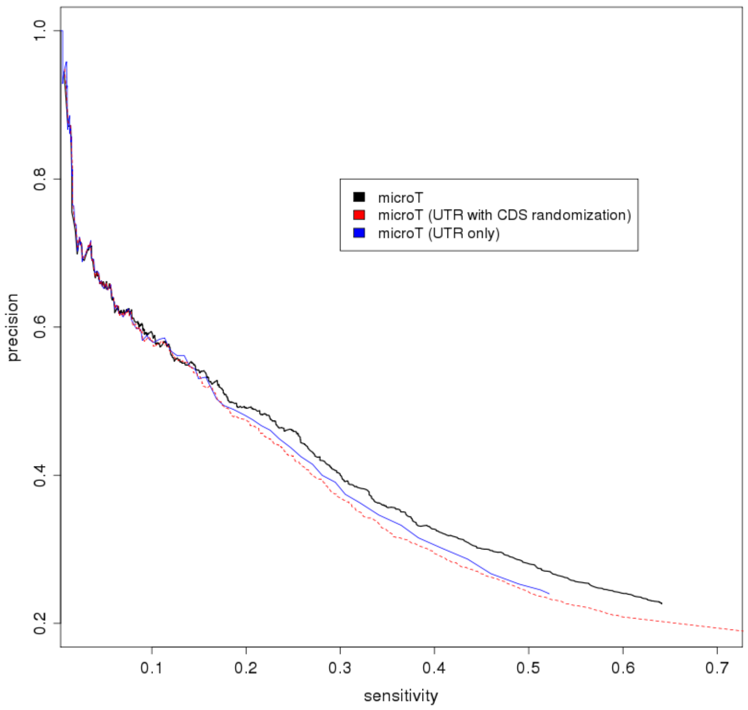
*3'UTR binding category also used in TargetScan are in bold.

Supplementary Table 2. The final set of features with their regression coefficients and significances.

features for CDS binding	**coefficient**	**Wald test**[*]	**Wilcoxon test**[**]
distance to closest CDS end	$-5.496 * 10^{-4}$	$3.4581 * 10^{-11}$	$2.45 * 10^{-9}$
CDS conservation	0.1628	$6.8078 * 10^{-5}$	$9.28 * 10^{-11}$
binding category weight	0.5055	0.0388447	$2.22 * 10^{-17}$
adjacent MRE distance	-0.003369	0.004249	0.00449
free energy	0.09122	$1.53 * 10^{-8}$	$8.23 * 10^{-6}$
synergistic features for CDS binding			
bnt11*bnt1	-0.3861	0.0198	$4.62 * 10^{-4}$
flanking AU content.free energy	-0.2837	$<2 * 10^{-16}$	$5.68 * 10^{-29}$
features for 3'UTR binding			
distance to closest 3'UTR end	-0.00148	$1.65 * 10^{-12}$	$5.7 * 10^{-98}$
binding category weight	0.385	0.00221	$2.49 * 10^{-196}$
free energy	0.0698	$3.66 * 10^{-9}$	$3.4 * 10^{-10}$
MRE accessibility (-1 to 2)	0.1395	$1.64 * 10^{-4}$	$4.85 * 10^{-19}$
conservation	1.801	$4.53 * 10^{-7}$	$1.46 * 10^{-129}$
synergistic features for 3'UTR binding			
flanking AU content*free energy	-0.322	$<2.0 * 10^{-16}$	$4.31 * 10^{-210}$
binding category weight*conservation	0.672	0.00713	$1.06 * 10^{-244}$
distance to closest 3'UTR end * adjacent MRE distance	$3.52 * 10^{-6}$	$6.85 * 10^{-4}$	$7.96 * 10^{-63}$

[*] significance of the feature in the regression obtained from a Wald test is shown as averages after threefold crossvalidation on the PAR-CLIP training data

[**] two-tailed Wilcoxon test for training data discrimination of the feature. Note that this test evaluates each feature independently, while the Wald-test evaluates whether the feature adds a significant contribution to the regression using all features.



Supplementary Figure 1: Precision of the predictions for the combined CDS and 3'UTR prediction ("microT"), the prediction using the 3'UTR only ("microT (UTR only)") and the randomization between the CDS and 3'UTR predictions ("microT (UTR with CDS randomization)").

6. microRNA FUNCTIONAL ANALYSIS

The goal of most miRNA analyses is to associate particular miRNAs to certain functions in a biological context. The identification of such associations may provide potential targets for diagnosis and therapy in human diseases. In this chapter I present four bioinformatics tools designed for the analysis of miRNA related biological data and produce results that will assist in understanding miRNA function. Presented in chronological order these tools consist of a Web server with extensive information, wide connectivity to biological resources and functional analysis through automated bibliographic searches, a Web application for the assessment of miRNA involvement in biological pathways denoted as DIANA-mirPath, a Web application for the identification of miRNAs involved in the differential expression of genes denoted as DIANA-mirExTra and an updated Web server with predictions for two widely studied species: *D. melanogaster* and *C. elegans* as well as associations of miRNAs to publications related to diseases.

6.1. DIANA-microT web server: elucidating miRNA functions through target prediction

In the miRNA field, information has been expanding in an increasing way in the last years. For this, the development of tools such as a target prediction program which provide primary data is not on its own sufficient and there is need for applications, primarily web based, which will serve as an interface between bioinformatics tools and researchers. These applications need to be able to organize the available information and present it in an intuitive and integrated way. In the following publication I present a Web server which provides extensive information and wide connectivity to online biological resources in a user friendly interface. Target gene and miRNA functions are elucidated through automated bibliographic searches and functional information is extracted through KEGG (Kanehisa, Goto et al. 2004) pathways. Also, the server offers links to nomenclature, sequence and protein databases and users are facilitated by being able to search for targeted genes using different nomenclatures or functional features. Additionally, since miRNA target prediction is a computationally intensive task, I developed an infrastructure in the computer cluster of the National Technical University of Athens to enable users to execute prediction for custom miRNA sequences, as part of the Web server. The work was published in Maragkakis *et al* (Maragkakis, Reczko et al. 2009)

Published online 30 April 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W273–W276***
doi:10.1093/nar/gkp292

DIANA-microT web server: elucidating microRNA functions through target prediction

M. Maragkakis[1], M. Reczko[1,6], V. A. Simossis[1], P. Alexiou[1], G. L. Papadopoulos[1], T. Dalamagas[2], G. Giannopoulos[2,3], G. Goumas[4], E. Koukis[4], K. Kourtis[4], T. Vergoulis[2,3], N. Koziris[4], T. Sellis[2,3], P. Tsanakas[4] and A. G. Hatzigeorgiou[1,5,]*

[1]Department of Molecular Oncology, Biomedical Sciences Research Center 'Alexander Fleming', Vari, [2]Department for the Management of Information Systems, ''Athena'' Research Center, Athens, [3]Knowledge and Database Systems Lab, Department of Computer Science, School of Electrical and Computer Engineering, National Technical University of Athens, [4]Computing Systems Laboratory, Department of Computer Science, School of Electrical and Computer Engineering, National Technical University of Athens, [5]Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA, USA and [6]Synaptic Ltd., Heraklion, Greece

Received February 22, 2009; Accepted April 14, 2009

ABSTRACT

Computational microRNA (miRNA) target prediction is one of the key means for deciphering the role of miRNAs in development and disease. Here, we present the DIANA-microT web server as the user interface to the DIANA-microT 3.0 miRNA target prediction algorithm. The web server provides extensive information for predicted miRNA:target gene interactions with a user-friendly interface, providing extensive connectivity to online biological resources. Target gene and miRNA functions may be elucidated through automated bibliographic searches and functional information is accessible through Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The web server offers links to nomenclature, sequence and protein databases, and users are facilitated by being able to search for targeted genes using different nomenclatures or functional features, such as the genes possible involvement in biological pathways. The target prediction algorithm supports parameters calculated individually for each miRNA:target gene interaction and provides a signal-to-noise ratio and a precision score that helps in the evaluation of the significance of the predicted results. Using a set of miRNA targets recently identified through the pSILAC method, the performance of several computational target prediction programs was assessed. DIANA-microT 3.0 achieved there with 66% the highest ratio of correctly predicted targets over all predicted targets. The DIANA-microT web server is freely available at www.microrna.gr/microT.

INTRODUCTION

MicroRNAs (miRNAs) are approximately 22-nt long endogenously expressed RNA molecules which regulate gene expression, preferentially by binding to the 3′-untranslated region (3′-UTR) of protein coding genes (1) and have been found to confer a novel layer of genetic regulation in a wide range of biological processes. Since their initial identification in 1993 (2), there have been several efforts for the identification of miRNA targeted genes (miTGs), but biological experiments have uncovered only a small fraction of all miTGs. Due to this, computational target prediction remains one of the key means to analyze the role of miRNAs in biological processes.

In the last 5 years, more than two dozen miRNA target prediction programs have been published (3). Most of these programs are mainly based on sequence alignment of the miRNA seed region (nucleotides 2–7 from the 5′-end of the miRNA) to the 3′-UTR of candidate target genes leading to the identification of putative binding sites. Their specificity is usually increased by exploiting the commonly observed evolutionary conservation of the binding sites or by using additional features such as structural accessibility (4,5), nucleotide composition (6) as well as location of the binding sites within the 3′-UTR (7). Recently, Selbach *et al.* (12) determined the complement of all the genes targeted by five miRNAs induced independently in HeLa cells using microarrays and pulsed stable isotope labeling with amino acids in cell

*To whom correspondence should be addressed. Tel: +30 210 9656310, Ext. 190; Fax: +30 210 9653934; Email: hatzigeorgiou@fleming.gr culture (pSILAC). Based on this dataset, they performed a comparative assessment of several commonly used target prediction programs which showed that only three [*DIANA-microT 3.0*, *PicTar* (9) and *TargetScanS* (13)] achieved precision levels (the fraction of the predicted targets that were actually downregulated) >60%. *DIANA-microT 3.0* predicted 294 targets total out of which 194 were correct and thus reached a precision of 66%.

The *DIANA-microT 3.0* algorithm is based on parameters that are calculated individually for each miRNA, and for each miRNA recognition element (MRE), depending on binding and conservation levels. The total predicted score of a miRNA:target gene interaction is the weighted sum of conserved and unconserved MREs of a gene. We also provide a signal-to-noise ratio (SNR) and a precision score specific for each interaction that can be used as a helpful confidence estimation of the 'correctness' and the false positive rate of each predicted miTG. This information can be easily looked up on the user-friendly *DIANA microT* web server where prediction results are organized in expandable tabs to group the available information, reduce the presentation complexity and show additional prediction details only on demand. Cases where a predicted interaction is registered as experimentally supported or predicted by other programs are also noted. The server offers an efficient search engine allowing multiple gene nomenclatures or queries based on gene involvement in specific biological pathways. The analysis of predicted interactions is supported by significance evaluation measures, extensive linkage to several online biological resources and automated bibliographic searches in PubMed. The server also supports prediction requests based on user-defined miRNA sequences and is integrated in a platform with two further miRNA functional analysis tools: *mirPath*, a pathway analysis tool of predicted targets and *mirExTra*, a miRNA analysis based on differential expressed mRNA profiles.

METHODS AND RESULTS

The DIANA microT web server

The web server may be accessed through a search engine with several options. The upper search box is used for browsing target genes predicted for a single miRNA. In this field, the miRNA name may be provided explicitly or partially. The second search box is used for identifying miRNAs which might be targeting a specific gene. In this case, the gene may be provided either based on Ensembl gene ID, RefSeq gene ID, common name or as part of the Ensembl description. If the search criteria correspond to more than one possible match, a list of alternatives is presented to the user to choose from. The lower search box combines the two search criteria offering the capability to identify if a specified miRNA targets a specified gene. For presenting the results, the web server results page (Figure 1) is divided in two parts. In the upper region, the user may find information concerning the provided search term; whereas, the prediction results are presented in the lower part.

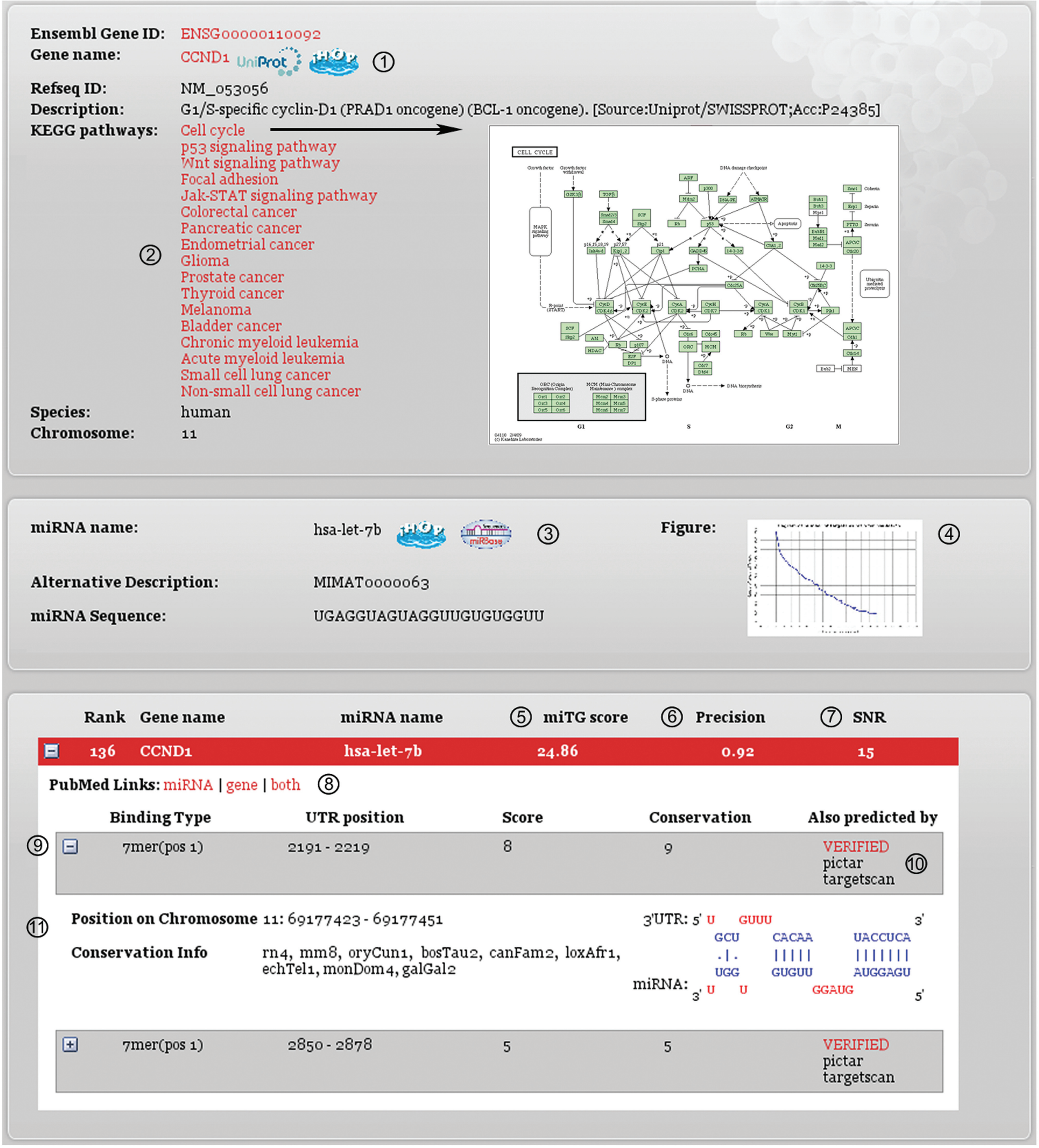
Figure 1 presents a typical results page based on a combined search for a miRNA and a gene. To assess the significance of the predicted interactions, the web server offers evaluation measures such as the precision score and the SNR. The information for each MRE score including conservation and binding structure of the MRE:mRNA interaction is also provided. Cases where an interaction is registered in the database of experimentally supported miRNA targets [TarBase, (8)] are highlighted with a link to the database. Moreover, all the interactions which are also predicted by PicTar (9) or TargetScan 4.2 (6) are noted in the web page. For each predicted interaction, the results page offers extensive linkage to multiple online biological resources [UniProt, Ensembl, miRBase, iHOP and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (11)] as well as automated bibliographic searches in PubMed for the miRNA, the target gene or the combination of the two.

DIANA-microT 3.0 algorithm description

A typical miRNA is approximately 22-nt long, but the nucleotides close to the 5′-end of the miRNA are crucial for recognizing a target sequence and binding to it. Usually, a strong binding [at least seven consecutive Watson–Crick (WC) base pairing nucleotides] between the first 9 nt from the 5′-end of the miRNA sequence (here called as the miRNA *driver* sequence) and the target gene is required for sufficient repression of protein production. However, there is experimental (10) evidence that a weaker binding, involving only six consecutively paired nucleotides or including G:U wobble pairs, can also repress protein production if there is additional binding between the miRNA 3′-end and the target gene.

The DIANA-microT 3.0 algorithm considers as MREs, those UTR sites that have 7-, 8- or 9-nt long consecutive WC base pairing with the miRNA, starting from position 1 or 2 from the 5′end of the miRNA. For sites with additional base pairing involving the 3′-end of the miRNA, a single G:U wobble pair or binding of only six consecutive nucleotides to the driver sequence are also allowed. Using as features the MRE binding type and the MRE conservation profile, all identified MREs are scored through comparative analysis versus a set of MREs identified based on mock miRNA sequences. The overall miTG score is calculated as the weighted sum of the scores of all identified MREs on the 3′-UTR. The algorithm uses up to 27 species to assess the MRE conservation profile taking into account both conserved and nonconserved MREs for the estimation of the final miTG score.

For the evaluation of each miRNAs predicted interactions, the program compares them to those predicted for a set of mock miRNAs. Mock miRNAs are independently created for each real miRNA and are designed to have approximately the same number of predicted targets as the real miRNA. This allows for the calculation of miRNA-specific SNR at different miTG score cut-offs as well as for the estimation of a precision score that provides an indication of the false positive rate of a particular miTG interaction.



Ensembl Gene ID: ENSG00000110092
Gene name: CCND1 UniProt iHOP ①
Refseq ID: NM_053056
Description: G1/S-specific cyclin-D1 (PRAD1 oncogene) (BCL-1 oncogene). [Source:Uniprot/SWISSPROT;Acc:P24385]
KEGG pathways: ②
Cell cycle
p53 signaling pathway
Wnt signaling pathway
Focal adhesion
Jak-STAT signaling pathway
Colorectal cancer
Pancreatic cancer
Endometrial cancer
Glioma
Prostate cancer
Thyroid cancer
Melanoma
Bladder cancer
Chronic myeloid leukemia
Acute myeloid leukemia
Small cell lung cancer
Non-small cell lung cancer
Species: human
Chromosome: 11

miRNA name: hsa-let-7b iHOP miRBase ③
Figure: ④
Alternative Description: MIMAT0000063
miRNA Sequence: UGAGGUAGUAGGUUGUGUGGUU

Rank	Gene name	miRNA name	⑤ miTG score	⑥ Precision	⑦ SNR
⊟ 136	CCND1	hsa-let-7b	24.86	0.92	15

PubMed Links: miRNA | gene | both ⑧

	Binding Type	UTR position	Score	Conservation	Also predicted by
⑨ ⊟	7mer(pos 1)	2191 - 2219	8	9	VERIFIED pictar targetscan ⑩

⑪
Position on Chromosome 11: 69177423 - 69177451
Conservation Info rn4, mm8, oryCun1, bosTau2, canFam2, loxAfr1, echTel1, monDom4, galGal2

```
3'UTR: 5' U    GUUU                  3'
          GCU  CACAA   UACCUCA
          .|.  |||||   |||||||
          UGG  GUGUU   AUGGAGU
miRNA: 3' U  U      GGAUG            5'
```

	Binding Type	UTR position	Score	Conservation	Also predicted by
⊞	7mer(pos 1)	2850 - 2878	5	5	VERIFIED pictar targetscan

Figure 1. DIANA-microT web server results page. The key features have been marked in the figure and are explained below. (1) Gene names and corresponding links to UniProt (protein information) and iHOP (functional and bibliographic information). (2) KEGG pathways in which the gene of interest is involved. (3) MiRNA names and corresponding links to miRBase (sequence information) and iHOP. (4) A graph showing the SNR of the miRNA. The SNR is calculated by the DIANA-microT algorithm and is based on a comparative analysis of the real miRNA versus a set of mock miRNAs. (5) The prediction score. Higher miTG scores correspond to higher possibility of correct prediction. (6) SNR score of the interaction. Greater values correspond to better distinction from the mock background. This attribute must be examined in combination with the SNR diagram provided for each miRNA. (7) The precision score of the interaction. This score ranges from 0 to 1, and it estimates the significance of the prediction. (8) Literature links that perform an automated search in PubMed for the gene, the miRNA or for the combination of the two. (9) Binding site info: (a) binding type indicates the number of the binding nucleotides in the 5′-end of the miRNA; (b) UTR position indicates the position of the binding site on the 3′-UTR; (c) score indicates the contribution of each binding site to the overall miTG score; and (d) conservation indicates the number of species in which the binding site is conserved. (10) This field indicates if the interaction may be found in the database of experimentally supported targets (TarBase) or if it has additionally been predicted by another target prediction program (TargetScan or Pictar). (11) Additional binding site info: (a) position on chromosome shows the position of the binding site on the chromosome; (b) conservation info indicates the species in which the binding site is conserved; and (c) graphic representation of the miRNA binding on the 3′-UTR.

Target prediction support for novel miRNA sequences

The DIANA-microT server also supports prediction requests for user-defined miRNA sequences. The results of the *de novo* predictions are stored in a database from which they can later be retrieved and presented to the user who is provided with a unique key via email notification. Support for target prediction based on user-defined sequences remains a computationally intensive task even though the DIANA-microT 3.0 prediction algorithm is mainly based on dynamic programming routines. For this reason, all miRNA target prediction requests are supported by a 256 core cluster consisting of 32 nodes which succeeds close to linear speedup and is hosted at the National Technical University of Athens (NTUA).

Integration of further analysis tools mirPath and mirExTra

In a typical case, the miRNA involved in a biological process is known and there is a need to predict its targets. However, the reverse search may also be relevant in some cases where, for instance, high-throughput data from cDNA arrays indicating changes in the expression of protein coding genes is available. In this case, the putative targets are known whereas the miRNA targeting them is unknown. To this end, an additional pre-processing tool for target prediction *(mirExTra)* is also available that is able to uncover miRNAs that may be involved in the changes of the transcriptome by processing a list of differentially expressed protein coding genes and a list of genes whose expression is unchanged. The program identifies hexamers that correspond to the driver region of a miRNA starting at position 1 and 2, which are significantly overrepresented in the input list of the overexpressed genes relative to those whose expression levels are constant under the same conditions. The web server is also combined with a post-processing analysis tool of predicted targets *(mirPath)* regarding their role in biological pathways. To this end, KEGG pathways that are enriched in a group of miTGs are identified and the results are visualized by highlighting the miTGs in the pathway.

CONCLUSION

The miRNA target prediction experiment by Selbach *et al.* (12) revealed the problem of the large fraction of under predicted or falsely predicted target genes. With lower score thresholds sensitivity can be increased, while trading off specificity and variable score thresholds can help to find best combination of these two measures. It is therefore crucial to give the user the possibility to modify this threshold and simultaneously present all relevant information facilitating the interpretation, the evaluation or even the experimental verification of predicted interactions. We found that most miRNA target prediction programs are insufficient in this respect, even when providing a graphical user interface for their results. Additionally, the search and identification of interactions of interest is complicated by the existence of different gene nomenclatures and may discourage researchers from trying to further elucidate the effects of miRNAs in biological processes. New miRNAs are identified nearly every month and this rate is increasing through the use of the new deep sequencing technologies. MiRNAs may also undergo editing and change the majority of their targets (14). Our approach, the DIANA-microT web server, has been designed with these challenges in mind and provides a user-friendly interface also for unannotated miRNAs by a precise target prediction algorithm.

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
2. Lee,R.C., Feinbaum,R.L. and Ambros,V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
3. Sethupathy,P., Megraw,M. and Hatzigeorgiou,A.G. (2006) A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat. Methods*, **3**, 881–886.
4. Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
5. Long,D., Lee,R., Williams,P., Chan,C.Y., Ambros,V. and Ding,Y. (2007) Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol.*, **14**, 287–294.
6. Grimson,A., Farh,K.K., Johnston,W.K., Garrett-Engele,P., Lim,L.P. and Bartel,D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
7. Gaidatzis,D., van Nimwegen,E., Hausser,J. and Zavolan,M. (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, **8**, 69.
8. Papadopoulos,G.L., Reczko,M., Simossis,V.A., Sethupathy,P. and Hatzigeorgiou,A.G. (2008) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res*, **37(Database issue)**, D155–D158.
9. Lall,S., Grun,D., Krek,A., Chen,K., Wang,Y.L., Dewey,C.N., Sood,P., Colombo,T., Bray,N., Macmenamin,P. *et al*. (2006) A genome-wide map of conserved microRNA targets in C. elegans. *Curr. Biol.*, **16**, 460–471.
10. Brennecke,J., Stark,A., Russell,R.B. and Cohen,S.M. (2005) Principles of microRNA-target recognition. *PLoS Biol.*, **3**, e85.
11. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
12. Selbach,M., Schwanhausser,B., Thierfelder,N., Fang,Z., Khanin,R. and Rajewsky,N. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.
13. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
14. Kawahara,B., Zinshteyn,B., Sethupathy,P., Iizasa,H., Hatzigeorgiou,A.G. and Nishikura,K. (2007) Dictation of silencing targets by adenosine-to-inosine editing of microRNAs. *Science*, **315**, 1137–1140.

6.2. DIANA-mirPath: Integrating human and mouse miRNAs in pathways

In the following publication we present a tool to identify molecular pathways potentially affected by the expression of single or multiple miRNAs. This tool, named DIANA-mirPath, is a functional analysis tool incorporating miRNA targets in biological pathways. It is a Web based application whose algorithm consists of an enrichment analysis of miRNA target genes within manually designed biological pathways. The combinatorial effect of co-expressed miRNAs in the modulation of a given pathway is taken into account through the analysis of multiple miRNAs simultaneously. This work was published in Papadopoulos *et al* (Papadopoulos, Alexiou et al. 2009).

BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 15 2009, pages 1991–1993
doi:10.1093/bioinformatics/btp299

Systems biology

DIANA-mirPath: Integrating human and mouse microRNAs in pathways

G. L. Papadopoulos[1,*], P. Alexiou[1], M. Maragkakis[1], M. Reczko[1,3] and A. G. Hatzigeorgiou[1,2,*]

[1]Institute of Molecular Oncology, Biomedical Sciences Research Center “Alexander Fleming”, 16602 Varkiza, Greece, [2]Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA, USA and [3]Synaptic Ltd., Heraklion, Greece

Received on March 9, 2009; revised on April 23, 2009; accepted on April 28, 2009

Advance Access publication May 12, 2009

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: DIANA-mirPath is a web-based computational tool developed to identify molecular pathways potentially altered by the expression of single or multiple microRNAs. The software performs an enrichment analysis of multiple microRNA target genes comparing each set of microRNA targets to all known KEGG pathways. The combinatorial effect of co-expressed microRNAs in the modulation of a given pathway is taken into account by the simultaneous analysis of multiple microRNAs. The graphical output of the program provides an overview of the parts of the pathway modulated by microRNAs, facilitating the interpretation and presentation of the analysis results.

Availability: The software is available at http://microrna.gr/mirpath and is free for all users with no login or download requirement.

Contact: papadopoulos@fleming.gr or hatzigeorgiou@fleming.gr

1 INTRODUCTION

Post-transcriptional regulation of protein coding genes is emerging as one of the new frontiers in modern cellular biology. MicroRNAs (miRNAs) are ∼22-nt long non-coding RNAs that play an important role as fine regulators of cellular processes through specific post-transcriptional repression of protein coding genes (Filipowicz *et al*., 2008). MiRNAs have been shown to factor into several physiological and pathological human conditions such as stem cell differentiation (Li and Gregory, 2008), immune response (Bi *et al*., 2009), blood lineage and transformation (Garzon and Croce, 2008), tumor development (Esquela-Kerscher and Slack, 2006) and metastasis (Lujambio *et al*., 2008).

MiRNAs are functionally related with both signaling (Cui *et al*., 2006) and metabolic (Tibiche and Wang, 2008) networks and also extensively interact with transcription factors (Yu *et al*., 2008) through distinct topological patterns, integrating transcriptional and post-transcriptional mechanisms in biological regulatory networks. Despite the growing evidence for miRNA involvement in central biological processes (Zhang and Su, 2009), the systematic integration of miRNAs in biological pathways remains rather incomplete. Currently there are only two miRNA-specific functional analysis tools available. MiRGator (Nam *et al*., 2008) performs a miRNA functional analysis by mapping the predicted targets of a single miRNA in pathways. The source of miRNA target genes used in the analysis may be any of three target prediction programs [TargetScanS (Lewis *et al*., 2005), PicTar (Krek *et al*., 2005) and miRanda (John *et al*., 2004)]. Results are presented in a tabular format sorted by the enrichment *P*-value of each pathway. MiRDB (Wang, 2008) is a miRNA target prediction program which additionally offers precompiled information regarding miRNAs enrichment in a single pathway.

Here we introduce DIANA-mirPath, a web-based application that performs an enrichment analysis of predicted target genes of one or more miRNAs in biological pathways. It is known that miRNAs have multiple target genes and there is strong evidence that some miRNAs can act in concert with each other in order to modulate a molecular pathway (Ivanovska and Cleary, 2008). The combinatorial effect of co-expressed miRNAs in the modulation of a given pathway is addressed by our tool through the simultaneous analysis of multiple miRNAs. MiRNA target genes implicated in a given pathway are graphically annotated on the pathway map providing a direct overview of the miRNA modulated parts, facilitating the interpretation and presentation of miRNA-dependent regulation of biological pathways.

2 METHODS

The input of DIANA-mirPath is a list of miRNA target genes, defined in a user-friendly web interface by simply selecting the miRNA name and the target prediction software of preference. Retrieval of miRNA target genes is automated for the three miRNA target prediction programs that achieved precision levels higher than 60% in a recent comparison (Selbach *et al*., 2008): DIANA-microT (Maragkakis *et al*., 2009), PicTar (Krek *et al*., 2005) and TargetScan (Lewis *et al*., 2005). Alternatively any list of human or mouse miRNA target genes compiled by the user can be used as input by the application. DIANA-mirPath performs an enrichment analysis of the input datasets by comparing each set of genes to all available biological pathways provided by the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) (Kanehisa and Goto, 2000). KEGG is a database resource that provides knowledge about several genomes as well as their relationships to biological systems and has been utilized as a systematic knowledge base for molecular and network

[*]To whom correspondence should be addressed.

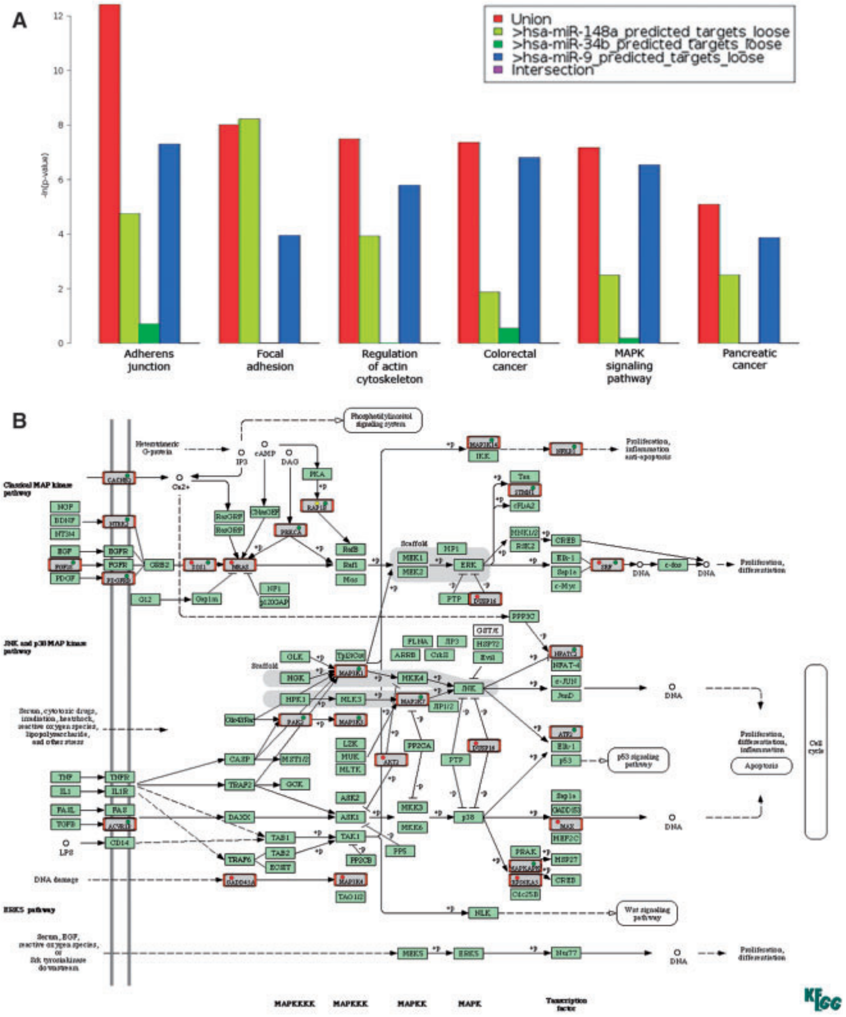


Fig. 1. DIANA-mirPath analysis, based on DIANA-microT 3.0 predictions, applied to explore altered biological processes by the epigenetically mediated silencing of miR-148a, miR-34b and miR-9, associated with human cancer metastasis. (**A**) The combinatorial effect of the miRNA signature is visible in the bar plot graph of the –ln *P* values. The Union dataset –ln *P*s (red bars) are higher than the –ln *P* values obtained for each single miRNA (yellow, green and blue bars) in most of the top targeted pathways. (**B**) The graphical annotation of the MAPK pathway produced by DIANA-mirPath. Targets of different miRNAs are differentiated by a coloured dot in the top of the highlight rectangle for a maximum of three miRNAs, in case of larger input datasets a mouse over option displays the names of miRNAs targeting the selected gene.

biology. Particularly the KEGG PATHWAY Database provides wiring diagrams of interaction and reaction networks between genes. The input dataset enrichment analysis is performed by a Pearson's chi-squared test $\{\chi^2 = \Sigma[(O - E)^2/E]\}$, where O (Observed) is the number of genes in the input dataset found to participate in a given pathway and E (Expected) is the number of genes expected by chance, given the pathway and input list size, to be member of that pathway. The input dataset enrichment in each KEGG Pathway is represented by the negative natural logarithm of the P-value (–ln P). The algorithm also performs an enrichment analysis of the Union and Intersection sets.

The enrichment P-value of the Union dataset in a specific pathway will reflect the coordinated downregulation of the pathway by all co-expressed miRNAs whereas the Intersection dataset gives an overview of the cooperative downregulation of single genes by all of the expressed miRNAs. A bar plot graph of the enrichment –ln P values is produced to facilitate the comparison of each pathway enrichment in different datasets (Fig. 1A). In the DIANA-mirPath output page all pathways are sorted according to a descending enrichment statistical score (–ln P) along with the number and names of each miRNAs target genes involved in each KEGG Pathway. MiRNA target genes found to be implicated in a given pathway are graphically annotated as an overlay of the pathway wiring diagram provided by the KEGG database and single genes or datasets can be independently highlighted by the user to facilitate the identification of genes or datasets of interest directly on the pathway map.

3 CONCLUSION

DIANA-mirPath is developed in order to estimate the impact of co-expressed miRNAs in biological pathways. As a representative scenario we apply DIANA-mirPath in the functional analysis of miRNAs associated with human metastatic cancer cells. In Lujambio *et al*. (2008), a DNA methylation-associated silencing of tumor suppressor miRNAs (miR-148a, miR-34b/c and miR-9) was found to contribute to the development of human cancer metastasis. In the same study, transfection of these miRNAs into the metastatic cell lines resulted in a lower capability of migration and less tumor growth. A functional analysis of this miRNA signature performed with DIANA-mirPath identifies both mitogenic and motility pathways to be extensively downregulated by the combined action of these three miRNAs. Top rated pathways involved cell–matrix and cell–cell adhesions, are known to play essential roles in cell motility, invasion and proliferation. Furthermore, the MAPK cascade (Fig. 1B), a highly conserved module that is involved in cell proliferation, differentiation and migration is also found to be significantly modulated by the presence or absence of these miRNAs. In the aforementioned case DIANA-mirPath is able to give a systemic explanation of the two observed phenotypes. In accordance with the particular emphasis given to the analysis of the coordinated modulation of a biological process by co-regulated microRNAs in the development of this tool, the example indicates that the global effect of the downregulated miRNAs might not only depend on single central target genes (i.e. well characterized oncogenes or tumor suppressor genes) but also through modulation of multiple components of proliferative and motility related pathways resulting on a more extended and coordinated downregulation. Given the lack of systematic integration of miRNAs in biological pathways we believe that the development of a tool like DIANA-mirPath can be a substantial aid in the planning and the interpretation of wet lab experiments aiming to infer systemic functions in miRNA expression signatures.

Funding: Aristeia Award from General Secretary Research and Technology, Greece.

Conflict of Interest: none declared.

REFERENCES

Bi,Y. *et al*. (2009) MicroRNAs: novel regulators during the immune response. *J. Cell Physiol.*, **218**, 467–472.

Cui,Q. *et al*. (2006) Principles of microRNA regulation of a human cellular signaling network. *Mol. Syst. Biol.*, **2**, 46.

Esquela-Kerscher,A. and Slack,F.J. (2006) Oncomirs—microRNAs with a role in cancer. *Nat. Rev. Cancer*, **6**, 259–269.

Filipowicz,W. *et al*. (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.*, **9**, 102–114.

Garzon,R. and Croce,C.M. (2008) MicroRNAs in normal and malignant hematopoiesis. *Curr. Opin. Hematol.*, **15**, 352–358.

Ivanovska,I. and Cleary,M.A. (2008) Combinatorial microRNAs: working together to make a difference. *Cell Cycle*, **7**, 3137–3142.

John,B. *et al*. (2004) Human MicroRNA targets. *PLoS Biol.*, **2**, e363.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Krek,A. *et al*. (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, 37, 495–500.

Lewis,B.P. *et al*. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.

Li,Q. and Gregory,R.I. (2008) MicroRNA regulation of stem cell fate. *Cell Stem Cell*, **2**, 195–196.

Lujambio,A. *et al.* (2008) A microRNA DNA methylation signature for human cancer metastasis. *Proc. Natl. Acad. Sci. USA*, **105**, 13556–13561.

Maragkakis,M. *et al*. (2009) DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res.*, [Epub ahead of print, May 14, 2009]

Nam,S. *et al*. (2008) miRGator: an integrated system for functional annotation of microRNAs. *Nucleic Acids Res.*, **36**, D159–D164.

Selbach,M. *et al*. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.

Tibiche,C. and Wang,E. (2008) MicroRNA regulatory patterns on the human metabolic network. *Open Syst. Biol. J.*, **1**, 1–8.

Wang,X. (2008) miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA*, **14**, 1012–1017.

Yu,X. *et al.* (2008) Analysis of regulatory network topology reveals functionally distinct classes of microRNAs. *Nucleic Acids Res*., **36**, 6494–6503.

Zhang,R. and Su,B. (2009) Small but influential: the role of microRNAs on gene regulatory network and 3'UTR evolution. *J. Genet. Genomics*, **36**, 1–6.

6.3. The DIANA-mirExTra web server: from gene expression data to miRNA function

High-throughput gene expression experiments are widely used to identify the role of genes involved in biological conditions of interest. Similarly, the identification of miRNAs and the genes they regulate may provide potential ways for diagnosis and therapy in human diseases. Although miRNA expression levels may not be routinely measured in high-throughput experiments, a possible involvement of miRNAs in the deregulation of gene expression can be computationally predicted and quantified through analysis of overrepresented motifs in the 3′UTR sequences of deregulated genes. For this, in the following publication we present DIANA-mirExTra which allows the comparison of frequencies of miRNA associated motifs between sets of genes that can lead to the identification of miRNAs responsible for the deregulation of large numbers of genes. I have also customized this program to be able to run in the computer cluster mentioned earlier allowing users to run the program through a Web interface. This work was published in Alexiou *et al* (Alexiou, Maragkakis et al. 2010).

The DIANA-mirExTra Web Server: From Gene Expression Data to MicroRNA Function

Panagiotis Alexiou[1,2], Manolis Maragkakis[1,3], Giorgio L. Papadopoulos[1], Victor A. Simmosis[1], Lin Zhang[4], Artemis G. Hatzigeorgiou[1,5]*

[1] Biomedical Sciences Research Center "Alexander Fleming", Institute of Molecular Oncology, Varkiza, Greece, [2] School of Biology, Aristotle University of Thessaloniki, Thessaloniki, Greece, [3] Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle, Germany, [4] Ovarian Cancer Research Center, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, [5] Computer and Information Sciences, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Abstract

Background: High-throughput gene expression experiments are widely used to identify the role of genes involved in biological conditions of interest. MicroRNAs (miRNA) are regulatory molecules that have been functionally associated with several developmental programs and their deregulation with diverse diseases including cancer.

Methodology/Principal Findings: Although miRNA expression levels may not be routinely measured in high-throughput experiments, a possible involvement of miRNAs in the deregulation of gene expression can be computationally predicted and quantified through analysis of overrepresented motifs in the deregulated genes 3′ untranslated region (3′UTR) sequences. Here, we introduce a user-friendly web-server, DIANA-mirExTra (www.microrna.gr/mirextra) that allows the comparison of frequencies of miRNA associated motifs between sets of genes that can lead to the identification of miRNAs responsible for the deregulation of large numbers of genes. To this end, we have investigated different approaches and measures, and have practically implemented them on experimental data.

Conclusions/Significance: On several datasets of miRNA overexpression and repression experiments, our proposed approaches have successfully identified the deregulated miRNA. Beyond the prediction of miRNAs responsible for the deregulation of transcripts, the web-server provides extensive links to DIANA-mirPath, a functional analysis tool incorporating miRNA targets in biological pathways. Additionally, in case information about miRNA expression changes is provided, the results can be filtered to display the analysis for miRNAs of interest only.

Citation: Alexiou P, Maragkakis M, Papadopoulos GL, Simmosis VA, Zhang L, et al. (2010) The DIANA-mirExTra Web Server: From Gene Expression Data to MicroRNA Function. PLoS ONE 5(2): e9171. doi:10.1371/journal.pone.0009171

Editor: Eshel Ben-Jacob, Tel Aviv University, Israel

Received October 27, 2009; **Accepted** January 23, 2010; **Published** February 11, 2010

Funding: Aristeia Award from General Secretary Research and Technology, Greece. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: hatzigeorgiou@fleming.gr

Introduction

MicroRNAs (miRNA) are short, approximately 22 nucleotides long, endogenously expressed RNA molecules that regulate gene expression by binding, in a sequence specific manner, to the 3′ UnTranslated Region (3′UTR) of messenger RNA (mRNA) molecules [1]. MiRNAs are not only present but can also be abundant in eukaryotic cells, controlling a wide variety of target genes [2]. In the past few years, miRNAs have been associated to the regulation of a wide range of biological processes [3].

High-throughput methods for gene expression profiling are being massively used in recent years. Such methods strive to describe specific transcriptomic states of a cell and can identify changes in expression levels between cell states of interest. Since miRNAs often regulate large numbers of mRNAs [4], there are cases where deregulated miRNAs are responsible for a large part of gene expression changes. MicroRNA expression levels may or may not be experimentally measured in such experiments. However even if miRNAs that are down- or upregulated are known, there is always the possibility that only a subgroup of those miRNAs would be responsible for the changes in the transcriptome.

Such miRNAs may be identified via computational analysis, based on the fact that miRNAs target mRNA transcripts in a sequence dependent manner (Figure 1). Although it is known that miRNAs usually bind to specific sites in the 3′UTR region of targeted mRNA transcripts, the accurate identification of all miRNA target genes has not been possible yet. MiRNA binding sequences often tend to be overrepresented in sets of miRNA regulated genes compared to a random selection of genes [4,5]. Different methods have been previously used to identify over- or under- expressed miRNAs through changes in the levels of their target genes. Essentially, the procedure followed by all such approaches is to identify differentially expressed genes, identify motifs that are overrepresented in these genes and then connect these motifs back to miRNAs. In an analysis performed by Lim et al [4] a motif discovery tool, MEME (Multiple Em for Motif Elicitation)[6], was used in order to identify motifs of six or more nucleotides in length

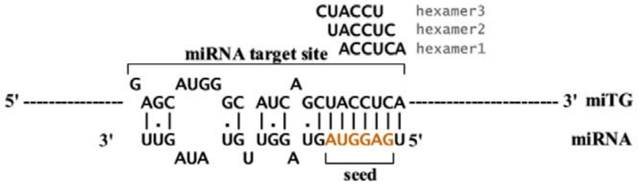

CUACCU hexamer3
UACCUC hexamer2
miRNA target site ACCUCA hexamer1
G AUGG A
5' AGC GC AUC GCUACCUCA 3' miTG
3' UUG UG UGG UGAUGGAGU 5' miRNA
AUA U A
seed

Figure 1. A miRNA molecule binds to a miRNA target gene (miTG). Hexamers 1,2 and 3 correspond to six nucleotide long sequences on the 3′UTR complementary to the first nucleotides of the miRNA . Hexamer 2 is the sequence complementary to the 'seed' of the miRNA, which has been suggested as the most important region for miRNA:miTG binding.
doi:10.1371/journal.pone.0009171.g001

that were significantly overrepresented in 3′UTR sequences of genes downregulated after hsa-miR-1 overexpression, compared to random 3′UTR sequences. The hexamer corresponding to position 2–7 of hsa-miR-1 was identified as the most significantly overrepresented motif.

In a similar experiment, Krutzfeld and colleagues [5] investigated the role of miRNA mmu-miR-122a in gene expression by neutralizing the miRNA through antagomirs and measuring the gene expression in wild type and knockdown cells. In a more sophisticated approach, they used the Wilcoxon Rank Sum test to compare hexamer frequencies between deregulated and unchanged genes between the two conditions. This analysis revealed that the frequency of the motif corresponding to the seed of mmu-miR-122 was significantly overrepresented in the 3′UTRs of upregulated genes and underrepresented in the 3′UTRs of downregulated genes.

Following this discovery, two freely available programs have been developed that perform similar computational analyses. MiReduce [7,8], uses the correlation of the genome wide mRNA log fold changes of genes against the motif content of their 3′UTRs. Each motif contained in the 3′UTR contributes linearly to the fold change prediction. The method iteratively calculates which motifs contribute most to the level of change of genes. Sylamer [9] is another software package that identifies overrepresented occurrences of sequences in a ranked list of genes using the hypergeometric p-value distribution. This approach calculates frequencies for hexamers 1, 2 and 3 as well as 7mers (positions 1–7, 2–8) and 8mers (positions 1–8, 2–9) and involves corrections for nucleotide biases. The p-values of each motif are compared to all other motifs. From the user point both programs have to be downloaded and compiled and include a limited data format as input. MiReduce outputs text files whereas Sylamer includes a java based graphical interface.

Given the broad impact of miRNAs in different development stages and diseases we have felt the emerging need for a tool that provides such investigations in a fast and user-friendly way. We believe that it is imperative that such a resource be platform independent and easy to use. A web-based implementation seems as the obvious choice. In this light, we have developed DIANA-mirExTra, an interactive and fully web based application that can be easily used by non-experts. Besides a motif analysis, the web server offers the option to use evolutionary information in order to refine results. Additionally, it allows the use of different nomenclatures for gene names as input and provides direct links to miRNA target prediction and functional analysis applications.

Results

The basic analysis flow of DIANA-mirExTra (www.microrna.gr/mirextra) is outlined in Figure 2. In the following section we will discuss each step of the algorithm in detail

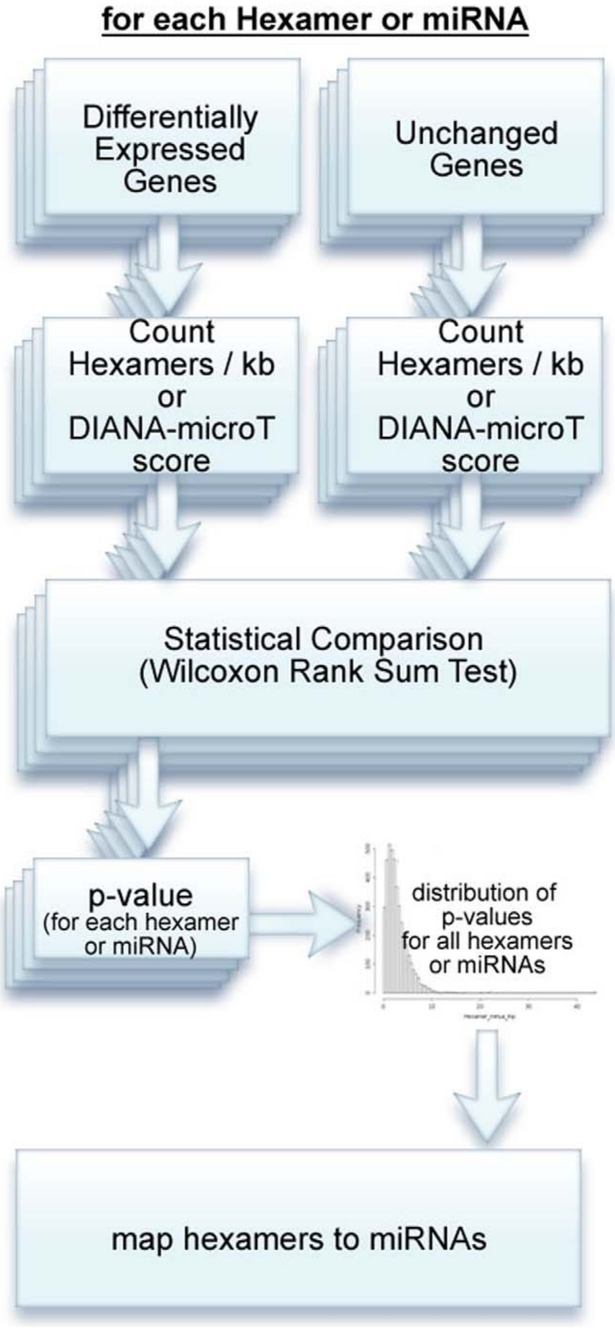

for each Hexamer or miRNA
Differentially Expressed Genes
Unchanged Genes
Count Hexamers / kb or DIANA-microT score
Count Hexamers / kb or DIANA-microT score
Statistical Comparison (Wilcoxon Rank Sum Test)
p-value (for each hexamer or miRNA)
distribution of p-values for all hexamers or miRNAs
map hexamers to miRNAs

Figure 2. Overview of the algorithm. For each possible hexamer, the occurrences on the 3′UTRs of changed and unchanged genes are counted. The counts are compared using the Wilcoxon Rank Sum Test and a p-value produced. The distribution of p-values is plotted in a histogram. Hexamers are mapped back to known miRNA sequences (see Figure 1). When DIANA-microT target prediction scores are used, the Wilcoxon Rank Sum Test is performed between scores of changed and unchanged genes. A p-value is calculated for each miRNA and a corresponding histogram is produced. The histogram and sorted p-values are returned to the user in the Results page (see Figure 4).
doi:10.1371/journal.pone.0009171.g002

Input Data

The input to the web-server is two sets of genes (changed and unchanged genes). The user is given the options to use a form in the webpage or to upload a file with the relative gene names. Gene names can be provided in any of a wide range of commonly used nomenclatures (Ensemble gene and transcript IDs, RefSeq IDs, HUGO, Affymetrix probe codes) and are automatically translated to Ensemble Gene IDs. The Ensemble database is the base for the sequences and gene names used by the program. The first list contains genes whose expression levels have been found to be significantly changed in a high-throughput experiment. The second list consists of background genes, which are usually genes that did not significantly change their expression levels. Optionally, an unchanged list may not be provided, and all genes not present in the first list will serve as the background set. Instead of a gene list the user may provide a list of genes with associated fold change values (or any other metric used in high-throughput experiments) be provided instead. In the latter case the changed and unchanged gene lists are produced by sorting all genes according to the metric provided and using a user-defined number of genes as "changed". Optionally, the user may use a miRNA filter, using a list of miRNAs of interest to calculate results only for hexamers corresponding to these miRNAs. This option simplifies the results page, and is especially useful when a miRNA expression measurement has been performed along the gene expression experiment.

AU Normalization on Microarray Data

When the input data is provided as microarray fold change levels, a single nucleotide composition bias may arise [10]. Single nucleotide AU normalization has been shown to improve the identification of miRNA signatures from microarray data. DIANA-mirExTra optionally provides such normalization as shown in Figure 3. When a bias is present the AU normalization option will diminish the correlation between AU composition and gene expression changes (Figure 3a, 3b). Moreover, when a bias is not present, the AU correction step will not significantly affect input values (Figure 3c, 3d).

Wilcoxon Test

After the input gene lists have been determined, we proceed to compare the distributions of all possible hexamers on the 3′UTR sequences between them. A one-sided Wilcoxon Rank Sum test is used in order to identify hexamers that are present significantly more often in the set of changed genes compared to the background of unchanged genes, as has been previously proposed [5]. A probability value (p-value) for each motif is calculated

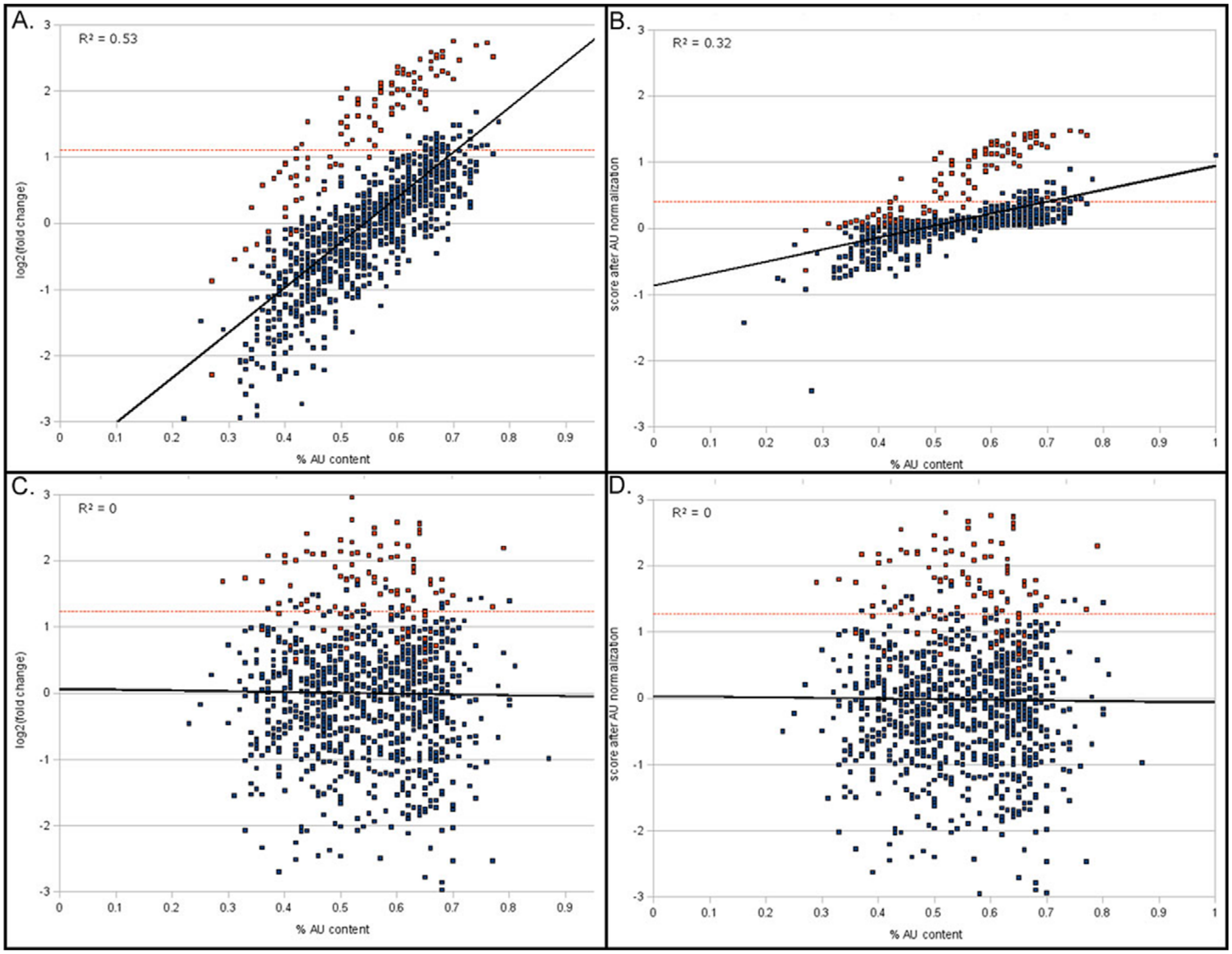


Figure 3. Results of AU correction. For 1000 genes, out of which 100 are upregulated (red points) and 900 are stable (blue points) the log2(fold change) is plotted against the percentage of As or Us in the 3′UTRs of genes. The top panels (A,B) show data with a linear AU bias and the bottom panels (C,D) show data with no AU bias. The left panels (A,C) show original data and the right panels (B,D) show data after AU correction. An optimal linear fit (black line) passes through the data with a correlation coefficient (R^2) denoted for each panel. A dotted red line denotes the 100 genes with the highest log2(fold change) values.
doi:10.1371/journal.pone.0009171.g003

signifying the probability that the changed and unchanged sets are produced by the same distribution and the differences between them are due to chance alone. As a more intuitive measure, the equivalent negative natural logarithm of the p-value (-lnp) is generally used. A histogram of the distribution of -lnp values of all motifs is provided in the results page so that the user may visually evaluate the significance of the results for a motif or miRNA of interest (Figure 4). Hexamers are mapped back onto the first 8 nucleotides of a miRNA (Figure 1), known to be the most important for the miRNA:mRNA binding [11,12].

Combination of Hexamers

The hexamer starting at position 2 of the miRNA, frequently called the 'seed' hexamer (Figure 1), can be used for an approximate identification of miRNA binding sites, with identification precision similar to some dedicated target prediction algorithms (Selbach et al. 2008). However, more than one miRNAs may share the same seed hexamer. We investigated whether it is possible to distinguish between similar miRNAs by using the p-values of flanking hexamers 1 and 3. Weighted -lnp values of three hexamers corresponding to each miRNA were summed using different weights to produce a total hexamer score (Figure 5). As a result, DIANA-mirExTra provides a combinatorial hexamer score in which the -lnp value of hexamer 1 is multiplied by a weight of 0.6 and added to the -lnp value of hexamer 2, and hexamer 3 is not taken This approach allows a single score per miRNA that takes into account the whole active region of the 8 first nucleotides of the miRNA.

Conserved Hexamers

Hexamers corresponding to miRNAs represent an extremely loose definition of miRNA target sites. Arguably most of the hexamers present on the 3′UTR of a gene will not be parts of active miRNA target sites. Interspecies conservation has been extensively used by miRNA target prediction programs in order to refine predictions of putative miRNA target sites. Conservation of hexamers between human and mouse sequences can be optionally used in DIANA-mirExTra for a stricter and more precise definition of miRNA target sites. This option prevents a part of randomly occurring hexamers from being counted as miRNA targets, but will be intrinsically biased towards miRNAs strongly conserved between the human and murine genomes.

Use of Target Prediction

Another option provided by DIANA-mirExTra is the use of miRNA target prediction scores instead of hexamer frequencies on a 3′UTR. A one-sided Wilcoxon Rank Sum test is performed for each miRNA, between the target prediction scores of the list of 'changed' genes versus the target prediction scores of the list of 'unchanged' genes. Target prediction scores are calculated by DIANA-microT [13,14], an advanced miRNA target prediction program that takes into account diverse features such as evolutionary conservation in several species and weights for different types of binding sites.

Meta-Analysis: Integration with DIANA-mirPath

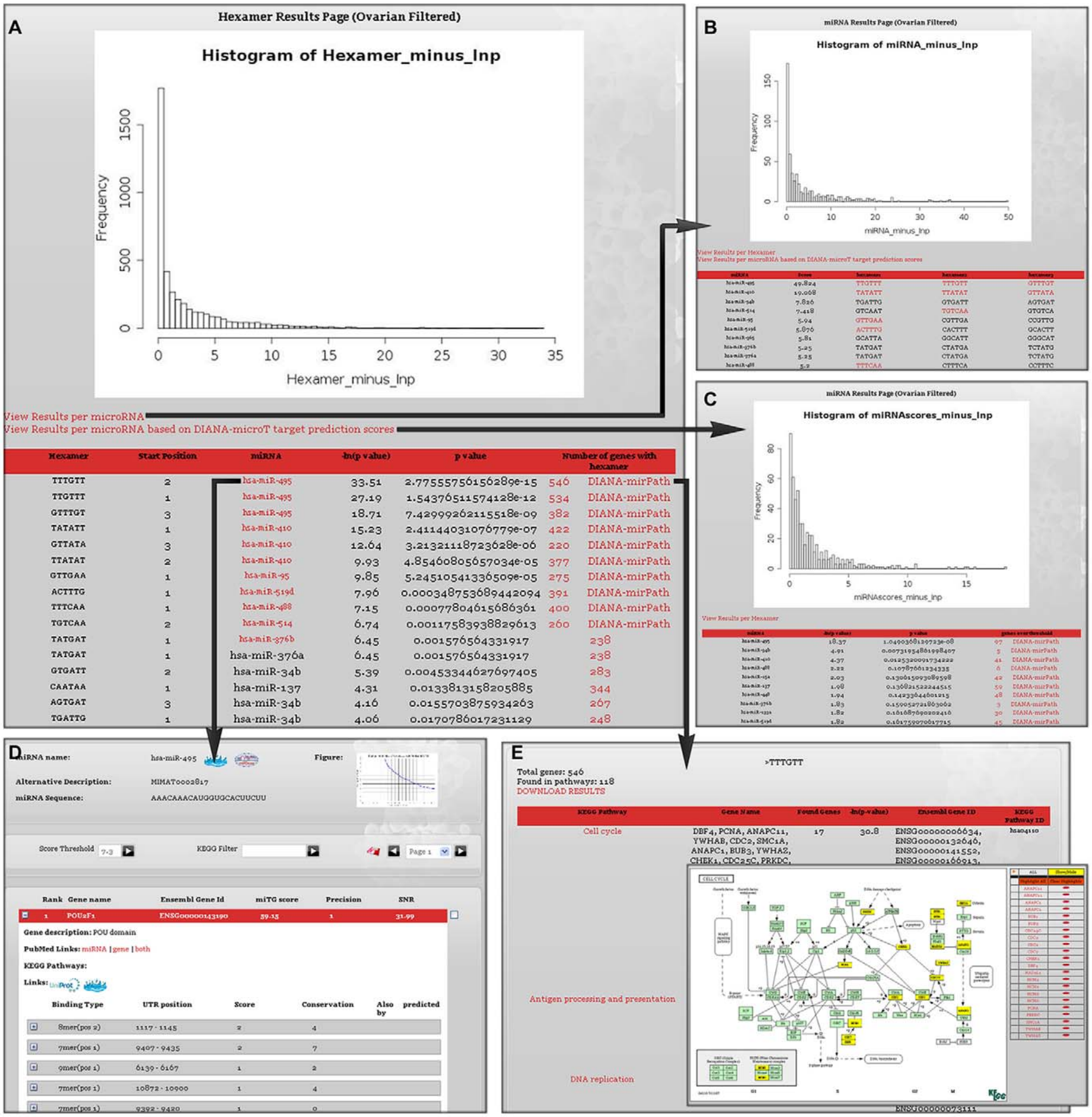
After results are produced, a link to the results page is returned to the user via email. Runs typically take approximately 10 minutes. The main DIANA-mirExTra Results Page (Figure 4a) shows p-values associated with each hexamer sorted in order of significance. A histogram of the -lnp values of all possible hexamers allows the user to evaluate the significance of the p-values of a given motif. Links to Results pages for combined motifs and target prediction score results allow the user to navigate to these pages (Figure 4b,4d). For the targets of each miRNA belonging to the set of 'changed' genes a link to functional analysis using DIANA-mirPath [15] is provided (Figure 4c). DIANA-mirPath is a tool that identifies KEGG pathways [16,17] enriched in the genes of interest. Such functional analysis may help to elucidate the biological function of a miRNA implicated in the condition of interest.

Evaluation

DIANA-mirExTra was tested on several experimental datasets in which a single miRNA has been artificially deregulated, and mRNA levels measured using microarrays. In such a high throughput experiment [4], human miRNA hsa-miR-1 was overexpressed in HeLa cells and the mRNA levels of protein coding genes were measured by microarray before and after the introduction of the miRNA. Using a set of 82 genes identified as downregulated in the original paper, we have identified the three hexamers associated with hsa-miR-1 as the most significantly overrepresented hexamers and the combined score of hsa-miR-1 as the top ranking score. In the same paper a similar experiment was performed with the overexpression of hsa-miR-124 in HeLa cells. All three hexamers corresponding to hsa-miR-124 achieved the maximum -lnp value and consequently the combined score of hsa-miR-124 was also the top-ranking one. In other experiments involving the repression of miRNA functionality using 'antagomirs' [5] and miR-155 deficient mice, DIANA-mirExTra has correctly identified the repressed murine miRNA in both occasions (mmu-miR-122a and mmu-miR-155) using microarray data. For both experiments the miRNA in question is found as top of the combined scores list, with a large difference in combined score to the second miRNA.

Beyond expression microarray data, DIANA-mirExTra was also tested on high-throughput protein data. In a recent set of experiments [18], a large number of proteins were identified as downregulated after overexpression of each of five miRNAs (let-7b, miR-155, miR-16, miR-1, miR-30a) and pulsed stable isotope labeling with amino acids in cell culture (pSILAC) assays. DIANA-mirExTra was used to identify the implicated miRNA in each of these cases. The hexamer in position 2 has been found as the top ranking hexamer with the maximum possible -lnp value in all datasets. All results pages for datasets mentioned above can be openly accessed online at http://diana.cslab.ece.ntua.gr/hexamers/prec_results.php.

An early version of DIANA-mirExTra has been used in order to identify multiple miRNAs involved in the progression from early to late stage Epithelial Ovarian Cancer (EOC) [19]. Among other experiments, 76 EOC specimens (8 early and 68 late stage EOC) were analyzed using microarrays and 948 genes were identified as significantly upregulated in late stage EOC. A further 15212 genes were considered as unchanged between the two cancer stages. Using this data, the DIANA-mirExTra algorithm was effectively used to predict twelve miRNAs as significant candidates possibly contributing to late-stage EOC. Five of these twelve miRNAs were located on a specific miRNA gene cluster (*Dlk1 – Gtl2* domain on chr14) suggesting that this miRNA cluster could possibly be involved with EOC progression to the late stage. Further experiments showed that the miRNA gene cluster identified by DIANA-mirExTra is commonly altered in EOC and possibly other human epithelial tumors, thus validating the involvement of these miRNAs in EOC progression. Additionally, a link was established between down-regulation of the expression of miRNAs encoded in the *Dlk1 – Gtl2* cluster and higher tumor proliferation leading to shorter patient survival times. The functional analysis of predicted target genes for the top microRNAs responsible for the transition identified the "Cell Cycle"pathway as significantly



Hexamer Results Page (Ovarian Filtered)

View Results per microRNA
View Results per microRNA based on DIANA-microT target prediction scores

Hexamer	Start Position	miRNA	-ln(p value)	p value	Number of genes with hexamer	
TTTGTT	2	hsa-miR-495	33.51	2.77555756156289e-15	546	DIANA-mirPath
TTGTTT	1	hsa-miR-495	27.19	1.54376511574128e-12	534	DIANA-mirPath
GTTTGT	3	hsa-miR-495	18.71	7.42999262115518e-09	382	DIANA-mirPath
TATATT	1	hsa-miR-410	15.23	2.41144031076779e-07	422	DIANA-mirPath
GTTATA	3	hsa-miR-410	12.64	3.21321118723628e-06	220	DIANA-mirPath
TTATAT	2	hsa-miR-410	9.93	4.85460805657034e-05	377	DIANA-mirPath
GTTGAA	1	hsa-miR-95	9.85	5.24510541336509e-05	275	DIANA-mirPath
ACTTTG	1	hsa-miR-519d	7.96	0.000348753689442094	391	DIANA-mirPath
TTTCAA	1	hsa-miR-488	7.15	0.000778046156863661	400	DIANA-mirPath
TGTCAA	2	hsa-miR-514	6.74	0.00117583938829613	260	DIANA-mirPath
TATGAT	1	hsa-miR-376b	6.45	0.001576564331917		238
TATGAT	1	hsa-miR-376a	6.45	0.001576564331917		238
GTGATT	2	hsa-miR-34b	5.39	0.00453344627697405		283
CAATAA	1	hsa-miR-137	4.31	0.0133813158205885		344
AGTGAT	3	hsa-miR-34b	4.16	0.0155703875934263		267
TGATTG	1	hsa-miR-34b	4.06	0.0170786017231129		248

miRNA Results Page (Ovarian Filtered)

View Results per Hexamer
View Results per microRNA based on DIANA-microT target prediction scores

miRNA	Score	hexamer1	hexamer2	hexamer3
hsa-miR-495	49.824	TTGTTT	TTTGTT	GTTTGT
hsa-miR-410	19.068	TATATT	TTATAT	GTTATA
hsa-miR-34b	7.826	TGATTG	GTGATT	AGTGAT
hsa-miR-514	7.418	GTCAAT	TGTCAA	GTGTCA
hsa-miR-95	5.94	GTTGAA	CGTTGA	CCGTTG
hsa-miR-519d	5.876	ACTTTG	CACTTT	GCACTT
hsa-miR-96	5.81	GCATTA	GGCATT	GGGCAT
hsa-miR-376b	5.25	TATGAT	CTATGA	TCTATG
hsa-miR-376a	5.25	TATGAT	CTATGA	TCTATG
hsa-miR-488	5.2	TTTCAA	CTTTCA	CCTTTC

miRNA Results Page (Ovarian Filtered)

View Results per Hexamer

miRNA	-ln(p value)	p value	genes over threshold	
hsa-miR-495	18.37	1.04903681297236e-08	97	DIANA-mirPath
hsa-miR-34b	4.91	0.00731954861998407	5	DIANA-mirPath
hsa-miR-410	4.37	0.0125320091734222	41	DIANA-mirPath
hsa-miR-488	2.22	0.107876612343335	6	DIANA-mirPath
hsa-miR-15a	2.03	0.130615093089598	42	DIANA-mirPath
hsa-miR-137	1.98	0.136821522244515	59	DIANA-mirPath
hsa-miR-448	1.94	0.142336446601215	48	DIANA-mirPath
hsa-miR-376b	1.83	0.159052721863062	3	DIANA-mirPath
hsa-miR-133a	1.82	0.161687690202416	30	DIANA-mirPath
hsa-miR-519d	1.82	0.161759070617715	45	DIANA-mirPath

miRNA name: hsa-miR-495
Alternative Description: MIMAT0002817
miRNA Sequence: AAACAAACAUGGUGCACUUCUU
Figure:

Score Threshold 7.3 KEGG Filter Page 1

Rank	Gene name	Ensembl Gene Id	miTG score	Precision	SNR
1	POU2F1	ENSG00000143190	59.15	1	31.99

Gene description: POU domain
PubMed Links: miRNA | gene | both
KEGG Pathways:
Links: UniProt

Binding Type	UTR position	Score	Conservation	Also predicted by
8mer(pos 2)	1117 - 1145	2	4	
7mer(pos 1)	9407 - 9435	2	7	
9mer(pos 1)	6139 - 6167	1	2	
7mer(pos 1)	10872 - 10900	1	4	
7mer(pos 1)	9392 - 9420	1	0	

>TTTGTT

Total genes: 546
Found in pathways: 118
DOWNLOAD RESULTS

KEGG Pathway	Gene Name	Found Genes	-ln(p-value)	Ensembl Gene ID	KEGG Pathway ID
Cell cycle	DBF4, PCNA, ANAPC11, YWHAB, CDC2, SMC1A, ANAPC1, BUB3, YWHAZ, CHEK1, CDC25C, PRKDC,	17	30.8	ENSG00000006634, ENSG00000132646, ENSG00000141552, ENSG00000166913,	hsa04110
Antigen processing and presentation					
DNA replication				ENSG00000073111	

Figure 4. Results Page and links (Epithelial Ovarian Cancer). Genes upregulated and miRNAs downregulated in late stage Epithelial Ovarian Cancer (EOC) compared to early stage EOC were run through DIANA-mirExTra. The main results page (A) consists of two parts. At the top of the page is the histogram of the distribution of –lnp values for all possible hexamers and at the bottom, the sorted list of hexamers that can be mapped on deregulated miRNAs with corresponding p-values. The same hexamer can be shown multiple times if it can be mapped on more than one miRNAs. Hexamers are sorted according to p-value and negative natural logarithm (-lnp value). Following the link "View Results per microRNA" the user is taken to a page (B) showing miRNAs sorted according to a combinatorial score produced by the values of hexamers 1 and 2. The link "View Results per microRNA based on DIANA-microT target prediction scores" leads to a similar results page (C) that uses as a measure the scores of each gene according to miRNA target prediction program DIANA-microT. (D) Genes that contain at least one of the top ten hexamers are marked in the results page of DIANA-microT. The DIANA-microT results page for each miRNA can be found following the link on the miRNA name from the first results page. Additionally, links to DIANA-mirPath lead to a page (E) showing functional analysis results using this program. Genes containing the hexamer of interest (A), or targeted by the miRNA of interest (C) are mapped on KEGG pathways and the most significantly overrepresented pathways can be identified by their corresponding p-values.

doi:10.1371/journal.pone.0009171.g004

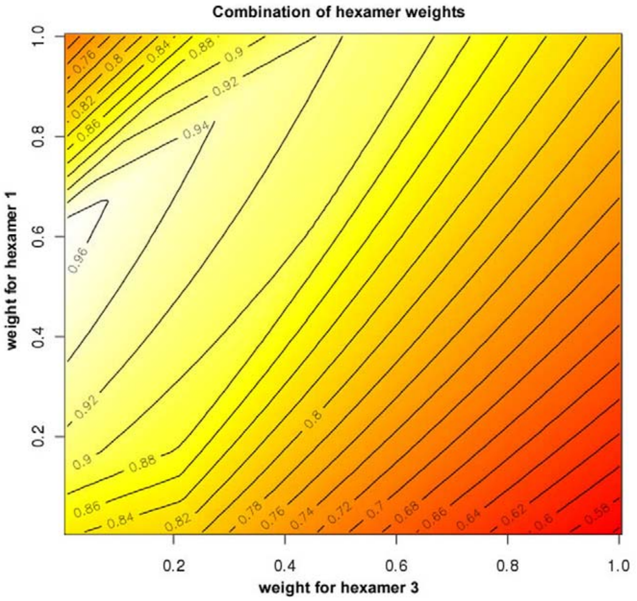


Figure 5. Combination of weighted -lnp values of the three hexamers. The weight for hexamer 1 is on the Y axis, for hexamer 2 is held constant at a value of 1, and for hexamer 3 is on the X axis. The mean normalized difference of the correct miRNA versus the next highest miRNA was maximized for 5 datasets of knocked out miRNAs (see Methods). The optimal weights combination for hexamers 1 and 3 were identified as 0.6 and 0 respectively. The value for hexamer 3 is still given in the Results page (see Figure 4) although it is not used for the combined score calculation.
doi:10.1371/journal.pone.0009171.g005

related with these genes. Other cancer related pathways were strongly related to the sets of genes suggesting ways in which miRNAs may affect EOC.

Discussion

The identification of miRNAs affecting the deregulation of genes is the primary objective of DIANA-mirExTra. Once miRNAs of interest are identified, the user can directly view predicted targets for these miRNAs as produced by DIANA-microT 3.0 [13,14]. However, the way in which this deregulation may contribute to disease development or other processes of interest can be elucidated through functional analysis of the results. DIANA-mirExTra moves towards this direction through its direct integration with a functional analysis tool, DIANA-mirPath, suggesting biological pathways in which targets of a miRNA of interest are more probable to be involved.

With our implementation of the algorithms proposed here in a user-friendly web server we strive to allow users without expertise in data analysis to use our algorithms easily and effectively. In other relevant available software packages, that first need to be downloaded and installed locally, 3′UTR and miRNA sequences have to be provided by the user in a program-specific format. In DIANA-mirExTra sequences are automatically downloaded by the Ensembl database [20] and linked to several widely used nomenclatures. This allows the direct use of the program without the prior download of bulky sequence files and without the need to process such files to fit a predetermined format. Additionally, the program is run in a web browser, without the need for download and compilation of source code. Results are stored in an online server and are accessible from anywhere and at all times. All submitted jobs are run remotely on a dedicated computational cluster, and allow users with low computational power to use the program without experiencing long running times or memory problems

Using the simplest hexamers, the user opts for a loose definition of a miRNA target gene and may be able to identify processes not deeply conserved in other species. The option to use hexamers conserved between human and mouse provides a refinement of results for processes and miRNAs that are conserved between the two species. The stricter approach of using predicted microRNA targets as motifs takes into account conservation in several species as well as miRNA specific characteristics and could be biased towards more deeply conserved miRNAs.

Given the important role that miRNA regulation plays in several cell processes, a routine check of miRNA involvement should be encouraged even if there is no reason for it to be suspected. A simple and intuitive online tool such as DIANA-mirExTra is the obvious choice for such routine checks as it does not need complicated installation, processing of external datasets or high computational power on the user end.

Materials and Methods

MicroRNA and 3′UTR Sequences

MicroRNA sequences used in all predictions for DIANA-microT [13,14] are taken from miRBase Build 10.0 [21]. 3′UTR sequences used are the longest annotated 3′UTRs from Ensembl 48 [20]. Name conversions to Ensemble gene names are done based on alternative names provided from Ensembl 48. Multiple genome alignments are downloaded from UCSC Genome Browser [22]. Human (hg18) alignment to 16 vertebrate genomes and Mouse (mm9) alignment to 29 vertebrate genomes are used.

Hexamers

Non-overlapping six nucleotide long motifs (hexamers) are counted on the 3′UTR sequence of protein coding genes provided by Ensembl. The count of hexamers is divided by the length of the 3′UTR sequence to calculate normalized counts (hexamers/nt).

Combination of Hexamers

The difference between the score of the 'correct' miRNA and the next best miRNA that did not have all three same hexamers was calculated and divided by the score of the 'correct' miRNA. The sum of these differences for five protein data sets [18] was maximized. The sum was calculated for all combinations of weights for hexamer 1 and 3 in 0.01 intervals for values between 0 and 1 (Figure 5). Keeping the weight for the 'seed' hexamer constant at 1, we have determined that for a weight of hexamer 1 set to 0.6, no value of hexamer 3 will improve the identification of the correct miRNA. Therefore DIANA-mirExTra provides a combinatorial hexamer score in which the -lnp value of hexamer 1 is multiplied by a weight of 0.6 and added to the -lnp value of hexamer 2. Hexamer 3 is not taken into account for the calculation of the combinatorial hexamer score.

Conservation

There is the option to use only hexamers perfectly conserved on the 3′UTRs of human and mouse based on multiple species alignments downloaded from UCSC Genome Browser.

Wilcoxon Rank Sum Test

The statistical package R is used to perform the Wilcoxon Rank Sum Test between counts or scores of 'changed' and 'unchanged' genes. The function wilcox.exact(exactranktests) is used for the one-sided test. The maximum p-value that this method may produce is 10^{-19} which is equal to -lnp = 43.74

AU Bias Correction

When microarray data with fold change values are used as input, an optional AU content intensity bias removal step is allowed as described by Elkon and Agami [10]. The statistical package R is used for the correction, and specifically the scatter plot smoothing function lowess using default parameters. Artificial data plotted in Figure 3 consists of 1000 values with a linear correlation to AU composition (Figure 3a, Figure 3b) or no correlation to AU composition (Figure 3c, Figure 3d). The difference of the means between the 100 ''upregulated'' genes (red spots) and the 900 ''unchanged'' genes (blue spots) is the same between Figure 3a and Figure 3c. Normally distributed noise has been added to both sets. Several other artificially produced examples with varying differences and levels of AU bias were produced (data not shown) with similar results.

Acknowledgments

We would like to thank Martin Reczko for reading the manuscript and comments.

Author Contributions

Conceived and designed the experiments: AH. Performed the experiments: PA LZ. Analyzed the data: PA MM GLP VAS. Contributed reagents/materials/analysis tools: VAS. Wrote the paper: PA AH.

References

1. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116: 281–297.
2. Friedman RC, Farh KK, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 19: 92–105.
3. Zhang R, Su B (2009) Small but influential: the role of microRNAs on gene regulatory network and 3′UTR evolution. J Genet Genomics 36: 1–6.
4. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, et al. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 433: 769–773.
5. Krutzfeldt J, Rajewsky N, Braich R, Rajeev KG, Tuschl T, et al. (2005) Silencing of microRNAs in vivo with 'antagomirs'. Nature 438: 685–689.
6. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34: W369–373.
7. Bussemaker HJ, Li H, Siggia ED (2001) Regulatory element detection using correlation with expression. Nat Genet 27: 167–171.
8. Sood P, Krek A, Zavolan M, Macino G, Rajewsky N (2006) Cell-type-specific signatures of microRNAs on target mRNA expression. Proc Natl Acad Sci U S A 103: 2746–2751.
9. van Dongen S, Abreu-Goodger C, Enright AJ (2008) Detecting microRNA binding and siRNA off-target effects from expression data. Nat Methods 5: 1023–1025.
10. Elkon R, Agami R (2008) Removal of AU bias from microarray mRNA expression data enhances computational identification of active microRNAs. PLoS Comput Biol 4: e1000189.
11. Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, et al. (2004) A combined computational-experimental approach predicts human microRNA targets. Genes Dev 18: 1165–1178.
12. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120: 15–20.
13. Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, et al. (2009) DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res 37: W273–276.
14. Maragkakis M, Alexiou P, Papadopoulos GL, Reczko M, Dalamagas T, et al. (2009) Accurate microRNA target prediction correlates with protein repression levels. BMC Bioinformatics 10: 295.
15. Papadopoulos GL, Alexiou P, Maragkakis M, Reczko M, Hatzigeorgiou AG (2009) DIANA-mirPath: Integrating human and mouse microRNAs in pathways. Bioinformatics 25: 1991–1993.
16. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.
17. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, et al. (2008) KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res 36: W423–426.
18. Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, et al. (2008) Widespread changes in protein synthesis induced by microRNAs. Nature.
19. Zhang L, Volinia S, Bonome T, Calin GA, Greshock J, et al. (2008) Genomic and epigenetic alterations deregulate microRNA expression in human epithelial ovarian cancer. Proc Natl Acad Sci U S A 105: 7004–7009.
20. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, et al. (2008) Ensembl 2008. Nucleic Acids Res 36: D707–714.
21. Griffiths-Jones S (2006) miRBase: the microRNA sequence database. Methods Mol Biol 342: 129–138.
22. Karolchik D, Hinrichs AS, Kent WJ (2007) The UCSC Genome Browser. Curr Protoc Bioinformatics Chapter 1: Unit 1 4.

6.4. DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association

In the following publication we present an important update of the DIANA Web server which we performed in order to further enhance its scientific significance. The Web server has been updated to support predictions for two additional to *H. sapiens* and *M. musculus* widely studied species: *D melanogaster* and *C elegans*. Most importantly, in the updated version, we have associated miRNAs to diseases through bibliographic analysis and therefore provide insights for the potential involvement of miRNAs in biological processes. Also, we have analyzed the nomenclature used to describe mature miRNAs along different miRBase (Griffiths-Jones 2006) versions, and have extracted the naming history of each miRNA. This enables the identification of miRNA publications regardless of possible nomenclature changes. The work was published in Maragkakis, Vergoulis *et al* (Maragkakis, Vergoulis et al. 2011).

Nucleic Acids Research Advance Access published May 6, 2011

Nucleic Acids Research, 2011, **1–4**
doi:10.1093/nar/gkr294

DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association

Manolis Maragkakis[1,2], Thanasis Vergoulis[3], Panagiotis Alexiou[1,4], Martin Reczko[1], Kyriaki Plomaritou[5], Mixail Gousis[5], Kornilios Kourtis[6], Nectarios Koziris[6], Theodore Dalamagas[3,]* and Artemis G. Hatzigeorgiou[1,]*

[1]Institute of Molecular Oncology, Biomedical Sciences Research Center 'Alexander Fleming', 16672, Vari, Greece, [2]Institute of Computer Science, Martin Luther University Halle-Wittenberg, 06120, Halle, Germany, [3]IMIS Institute, 'Athena' Research Center, 11524, Athens, [4]School of Biology, Aristotle University of Thessaloniki, 54124, Thessaloniki, [5]Department of Computer and Communication Engineering, University of Thessaly, 38221, Bolos and [6]Department of Computer Science, School of Electrical and Computer Engineering, National Technical University of Athens, 15773, Zografou, Greece

Received March 4, 2011; Revised April 10, 2011; Accepted April 13, 2011

ABSTRACT

microRNAs (miRNAs) are small endogenous RNA molecules that are implicated in many biological processes through post-transcriptional regulation of gene expression. The DIANA-microT Web server provides a user-friendly interface for comprehensive computational analysis of miRNA targets in human and mouse. The server has now been extended to support predictions for two widely studied species: *Drosophila melanogaster* and *Caenorhabditis elegans*. In the updated version, the Web server enables the association of miRNAs to diseases through bibliographic analysis and provides insights for the potential involvement of miRNAs in biological processes. The nomenclature used to describe mature miRNAs along different miRBase versions has been extensively analyzed, and the naming history of each miRNA has been extracted. This enables the identification of miRNA publications regardless of possible nomenclature changes. User interaction has been further refined allowing users to save results that they wish to analyze further. A connection to the UCSC genome browser is now provided, enabling users to easily preview predicted binding sites in comparison to a wide array of genomic tracks, such as single nucleotide polymorphisms. The Web server is publicly accessible in www.microrna.gr/microT-v4.

INTRODUCTION

microRNAs are small endogenous RNA molecules that affect many biological processes by regulating gene expression in a post-transcriptional way. They are ∼21–22 nt in length whose primary role is to regulate gene expression through translational repression and/or mRNA degradation (1). The first miRNA molecules and miRNA targets were identified in 1993 via classical genetic techniques in *Caenorhabditis elegans* (2). Since then, there has been a dramatic increase in the number of miRNAs registered in miRBase (3). In parallel, the development of the first computational target prediction programs (4–7) led to the experimental identification of dozens of miRNA targets (8), and emphasized the need to provide miRNA target predictions in an efficient way to assist biologists in experimental design.

The previous version of the DIANA-microT Web server (9) presented extensive information for predicted miRNA target gene interactions in a user-friendly interface. It offers links to nomenclature, sequence and protein databases, information for experimentally verified targets through TarBase (8) and targets predicted by PicTar (6) or

*To whom correspondence should be addressed. Tel: +30 210 9656310 (ext. 190); Fax: +30 210 9653934; Email: artemis@fleming.gr
Correspondence may also be addressed to Theodore Dalamagas. Tel: +30 210 6990522 (ext. 102); Fax: +30 210 6990552; Email: dalamag@dblab.ece.ntua.gr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

TargetScan (10). Also, users are facilitated by being able to search for targeted genes using different names or functional features.

Here, we describe an extensive update of this Web server with several important improvements: (i) an advanced bibliographic analysis which correlates miRNAs to diseases, (ii) support for two additional species, (iii) a graphical display with all relevant functional information from the UCSC genome browser, (iv) tracking of changes in miRNA nomenclature and (v) user personalized sessions allowing personal query history and bookmarks.

METHODS AND RESULTS

Relation of miRNAs to functional features, diseases and medical descriptors

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 (Figure 1), 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. For example, in Figure 2 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 (11).

The feature that allowed the user to filter the targets of a miRNA based on KEGG pathways (12) is now integrated in the initial input query form. The user has now the option to enter a miRNA together with names of pathways of interest and provided immediately with only the miRNA target genes found in the pathways of interest. For example, a query for hsa-miR-221 would return 113 predicted targets at a score threshold of 0.5 while the combined search of the miRNA with the term ‘MAPK-signaling-pathway’ would filter the results to 21 relevant targets.

In addition, the positions of the binding sites on the mRNA of the target gene are graphically presented through the UCSC genome browser. This automatic upload can be used to provide information in comparison to other tracks of interest such as single nucleotide polymorphisms (SNPs), repeat elements and alternative 3′-UTR splice forms.

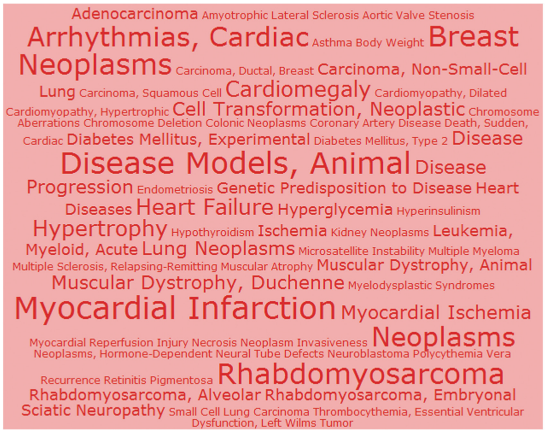
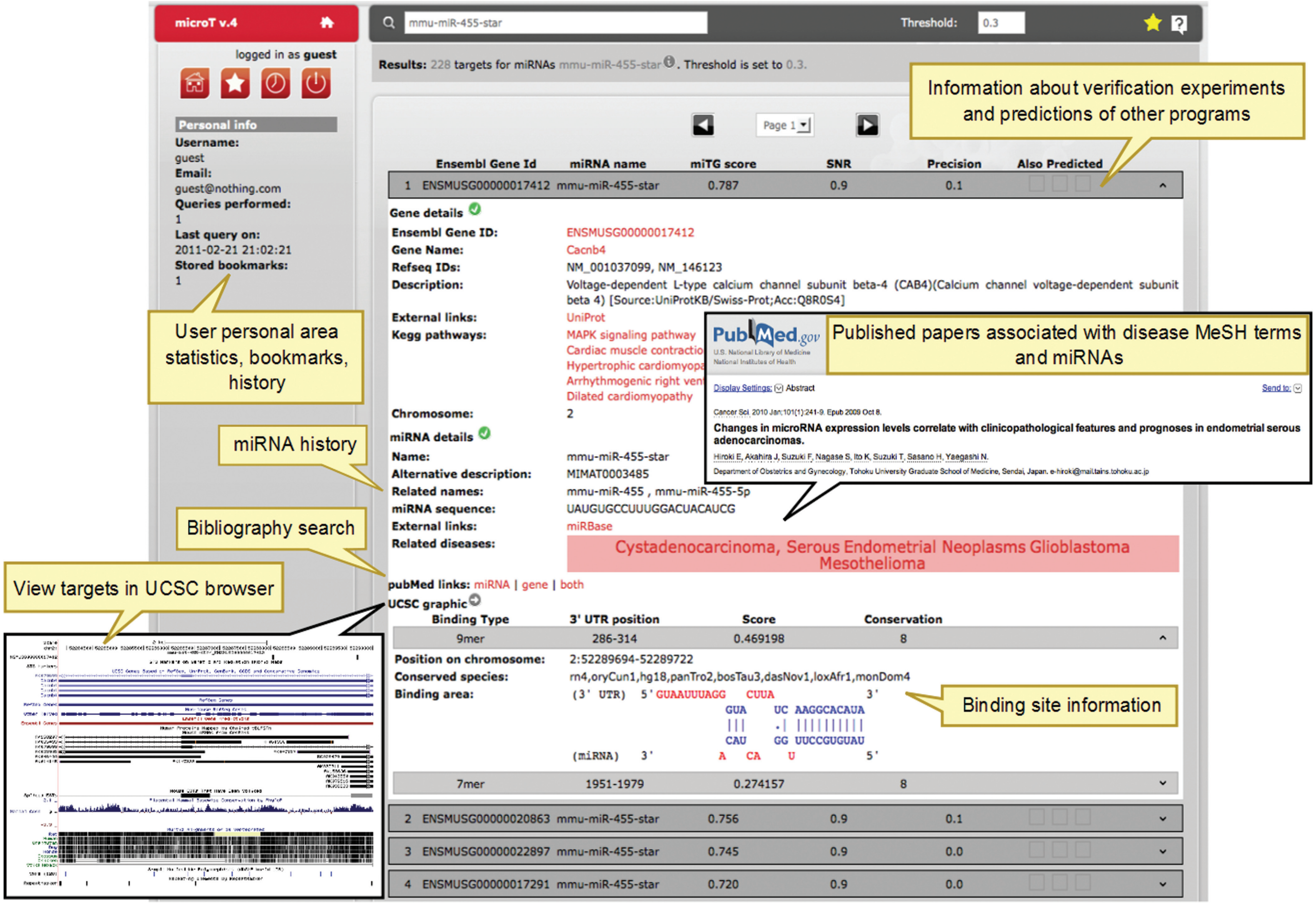


Figure 1. An example of a tag cloud for hsa-miR-1 showing relevant disease associated MeSH terms.

Extraction of miRNA history from miRBase

The relation of miRNAs to function and diseases described above is based on miRNA nomenclature. Since miRNA biology is still a field in flux, it can occur that a miRNA may change name between two successive versions of miRBase. Due to such changes, researchers may lose track of a miRNA full history and related literature searches will remain incomplete. To address this issue, an extended analysis on 13 versions of miRBase (versions 7.1. to 14) is performed, and the nomenclature history of each miRNA is extracted. The analysis uses version 13 of miRBase as the reference database. This version includes 1884 mature miRNAs for the four species supported in the Web server. Each miRNA is assigned a unique identification number denoted as ‘MIMAT id’ and one associated miRNA specific name. Among versions, changes are found in 77 MIMAT ids (38 in human, 37 in mouse and 2 in Drosophila) and 151 miRNA names (76 in human, 71 in mouse and 4 in Drosophila). This indicates that name changes are more frequent than changes in MIMAT ids. To keep track of these changes, a history index is integrated in the Web server. This index information is made available to the user through a specific feature called ‘miRNA history’ which is also used for miRNA related bibliography searches. For example, mmu-miR-455 first appeared in miRBase v8.1. Its name was later changed to mmu-miR-455-5 p in version 8.2 and later appeared as mmu-miR-455* in version 10.0 (Figure 3). This information can be retrieved from the miRNA history, but is also used to extend the association of mmu-mir-455* with endometrial serous adenocarcinoma in the publication where it is referred to as miR-455-5 p (Figure 2).



microT v.4

mmu-miR-455-star

Threshold: 0.3

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.

Page 1

	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 cardiomyopa
Arrhythmogenic right ven
Dilated cardiomyopathy
Chromosome: 2

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Display Settings: Abstract   Send to:

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. e-hiroki@mail.tains.tohoku.ac.jp

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'GUAAUUUAGG    CUUA          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	□ □ □

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

Figure 2. 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.

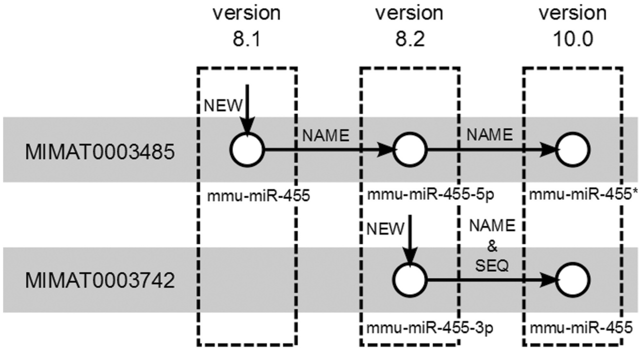


Figure 3. Graphic presentation of the changes involved in the history of miRNA mmu-miR-455. Initially, MIMAT0003485 was presented in version 8.1 as mmu-miR-455 but its name changed consecutively to mmu-miR-455-5 p in version 8.2 and mmu-miR-455-star (mmu-miR-455*) in version 10.0. Similarly, MIMAT0003742 was first presented in version 8.2 as mmu-miR-455-3 p, while in version 10.0, its sequence changed and it was renamed to mmu-miR-455.

Target prediction and supported miRNAs

The first version of DIANA-microT Web server was designed to support the functional analysis of human and mouse miRNAs. Now the server has been updated with predictions for two additional species and newer versions of miRBase (miRBase 13) and Ensembl (Ensembl 54) (13). In total predictions for 723 new miRNAs have been added, out of which 147 correspond to *Drosophila melanogaster*, 154 to *C. elegans* and the rest being new *Homo sapiens* and *Mus musculus* miRNAs. This results in an approximately doubled number of predicted targets in comparison to the previous version, totaling to more than six million predicted target genes. The Web server can support different prediction algorithms and currently provides the targets of an updated version of the previously used algorithm (14). While microT-v3.0 is based on features separating real and mock (shuffled) miRNAs the current version, microT-v4.0, uses high throughput experimental data for the same purpose (15).

Personalized user sessions

To allow users to take full advantage of the Web server's functions, several functional improvements have been implemented (Figure 2). The most important is an integrated personal user space in which users can easily save important searches and results that they wish to keep for future analysis. In particular, the system keeps the most recent user searches, giving the opportunity to repeat searches. A bookmarking mechanism provides the opportunity to save interesting results along with user comments. The personal space provides usage statistics regarding the most recent searches, thus, enabling them to keep track of their latest findings. It is noted that researchers may use any feature of the Web server irrespectively of the personal user space feature. Finally, special attention has been given to the Web Server documentation introducing hovering help notes for important fields.

CONCLUSION

In recent years it has become apparent that Web applications are an essential tool for researchers to decipher complex biological processes. As one of these, the DIANA microT Web server has been updated to integrate different additional resources in order to offer insight for putative miRNA involvement in biological processes, and to allow researchers to supplement their knowledge with already published scientific material. Performing an extended analysis on the versioning pattern of miRBase the history of each miRNA has been extracted, offering a unique feature in computational miRNA analysis. We believe that the current Web server update provides important tools for biological analysis and improves interaction with the complicated interconnections regarding miRNA functionality.

ACKNOWLEDGEMENT

We thank the anonymous reviewers for their helpful comments.

FUNDING

The Hellenic Republic, General Secretariat for Research and Technology (09SYN-13-1055). Funding for open access charge: 09SYN-13-1055.

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.
2. Lee,R.C., Feinbaum,R.L. and Ambros,V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
3. Griffiths-Jones,S. (2010) miRBase: microRNA sequences and annotation. *Current Protocols in Bioinformatics*, **Chapter 12**, pp. 11–10.
4. Kiriakidou,M., Nelson,P.T., Kouranov,A., Fitziev,P., Bouyioukos,C., Mourelatos,Z. and Hatzigeorgiou,A. (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes Dev*., **18**, 1165–1178.
5. Alexiou,P., Maragkakis,M., Papadopoulos,G.L., Reczko,M. and Hatzigeorgiou,A.G. (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, **25**, 3049–3055.
6. Lall,S., Grun,D., Krek,A., Chen,K., Wang,Y.L., Dewey,C.N., Sood,P., Colombo,T., Bray,N., Macmenamin,P. *et al*. (2006) A genome-wide map of conserved microRNA targets in C. elegans. *Curr. Biol*., **16**, 460–471.
7. Friedman,R.C., Farh,K.K., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*., **19**, 92–105.
8. Papadopoulos,G.L., Reczko,M., Simossis,V.A., Sethupathy,P. and Hatzigeorgiou,A.G. (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res*., **37**, D155–D158.
9. Maragkakis,M., Reczko,M., Simossis,V.A., Alexiou,P., Papadopoulos,G.L., Dalamagas,T., Giannopoulos,G., Goumas,G., Koukis,E., Kourtis,K. *et al*. (2009) DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res*., **37**, W273–W276.
10. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
11. Hiroki,E., Akahira,J., Suzuki,F., Nagase,S., Ito,K., Suzuki,T., Sasano,H. and Yaegashi,N. (2010) Changes in microRNA expression levels correlate with clinicopathological features and prognoses in endometrial serous adenocarcinomas. *Cancer Sci*., **101**, 241–249.
12. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*., **38**, D355–D360.
13. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al*. (2011) Ensembl 2011. *Nucleic Acids Res*., **39**, D800–D806.
14. Maragkakis,M., Alexiou,P., Papadopoulos,G.L., Reczko,M., Dalamagas,T., Giannopoulos,G., Goumas,G., Koukis,E., Kourtis,K., Simossis,V.A. *et al*. (2009) Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics*, **10**, 295.
15. Selbach,M., Schwanhausser,B., Thierfelder,N., Fang,Z., Khanin,R. and Rajewsky,N. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.

7. REVIEWS

Since miRNA biology is a rather a new scientific topic, reviews are particularly important for organizing available knowledge and highlighting open issues in the field. In this chapter, I present two reviews that I have co-authored regarding miRNAs. The first one is a review for available miRNA target prediction programs while the second one discusses online tools and resources which may be used for miRNA analysis in general.

7.1. Lost in translation: an assessment and perspective for computational miRNA target identification

In the following review we discuss and evaluate available miRNA target prediction methods. In the last years more than a dozen miRNA target prediction programs have been developed but not all of them perform equally well. Therefore the evaluation of prediction performance is a critical step for choosing which of these programs are best to be used for experimental design. The review was published in Alexiou *et al* (Alexiou, Maragkakis et al. 2009).

BIOINFORMATICS REVIEW Vol. 25 no. 23 2009, pages 3049–3055
doi:10.1093/bioinformatics/btp565

Gene expression

Lost in translation: an assessment and perspective for computational microRNA target identification

Panagiotis Alexiou[1,*], Manolis Maragkakis[1], Giorgos L. Papadopoulos[1], Martin Reczko[1,2] and Artemis G. Hatzigeorgiou[1,3]

[1]Institute of Molecular Oncology, Biomedical Sciences Research Center 'Alexander Fleming', 166 72 Varkiza, Greece, [2]Synaptic Ltd., 700 13 Heraklion, Greece and [3]Computer and Information Sciences, University of Pennsylvania, 19104-6391 Philadelphia, USA

Received on June 16, 2009; revised on September 24, 2009; accepted on September 27, 2009

Advance Access publication September 29, 2009

Associate Editor: Jonathan Wren

ABSTRACT

MicroRNAs (miRNAs) are a class of short endogenously expressed RNA molecules that regulate gene expression by binding directly to the messenger RNA of protein coding genes. They have been found to confer a novel layer of genetic regulation in a wide range of biological processes. Computational miRNA target prediction remains one of the key means used to decipher the role of miRNAs in development and disease. Here we introduce the basic idea behind the experimental identification of miRNA targets and present some of the most widely used computational miRNA target identification programs. The review includes an assessment of the prediction quality of these programs and their combinations.

Contact: p.alexiou@fleming.gr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

It was only recently that the term microRNA (miRNA) was introduced to describe short RNA molecules that regulate gene expression by binding preferably to the 3′ untranslated region (3′UTR) of protein coding genes (Bartel, 2004). Although miRNAs were first identified in 1993 (Lee *et al*., 1993) via classical genetic techniques in *Caenorhabditis elegans*, in 2001 it was suggested that they are widespread and abundant in cells (Lagos-Quintana *et al*., 2001; Lau *et al*., 2001; Lee and Ambros, 2001). Each miRNA is 19–24 nucleotides in length and is processed from a longer transcript, referred to as the primary transcript (pri-miRNA), which can be up to thousands of nucleotides long. Primary transcripts are processed in the cell nucleus to short, ∼70 nucleotide long stem-loop structures known as pre-miRNAs. In animals, this processing is performed by a protein complex known as the Microprocessor complex, consisting of the nuclease Drosha and the double-stranded RNA binding protein Pasha (Denli *et al*., 2004). Pre-miRNAs are processed to mature miRNAs in the cytoplasm by interaction with the endonuclease Dicer which cleaves the pre-miRNA stem-loop into two complementary short RNA molecules. One of these molecules is integrated into the RNA-induced silencing complex (RISC) and guides it to the mRNA where it can inhibit translation or induce mRNA degradation (Fig. 1) (Liu *et al*., 2004). Generally, miRNA transcripts may be located within the introns of protein-coding genes, entirely outside of protein-coding genes ('intergenic') or more rarely in coding exons, untranslated regions (UTRs) or exons of non-coding transcripts. Frequently, pri-miRNA transcripts code for more than one miRNAs which are transcribed together and are referred to as a miRNA cluster.

Since their initial identification, miRNAs have been found to confer a novel layer of genetic regulation in a wide range of biological processes. Their involvement in cellular commitment and cell cycle regulation gives an important role to the miRNA class of regulatory modules in animal development and human diseases. Specifically, miRNAs have been found to regulate various developmental stages in animals such as *C.elegans* (Lau *et al*., 2001; Lee and Ambros, 2001; Lee *et al*., 1993; Reinhart *et al*., 2000), *Danio Rerio* (Wienholds *et al*., 2005), *Drosophila melanogaster* (Aravin *et al*., 2003), *Mus musculus* (Baroukh *et al*., 2007), *Homo sapiens* (Chen *et al*., 2004; Lu *et al*., 2007; Yi *et al*., 2006) and in plants (Kidner and Martienssen, 2005). miRNA-mediated regulation of pathways involved in human disease is currently a very active field and miRNAs have been linked to several human pathologies such

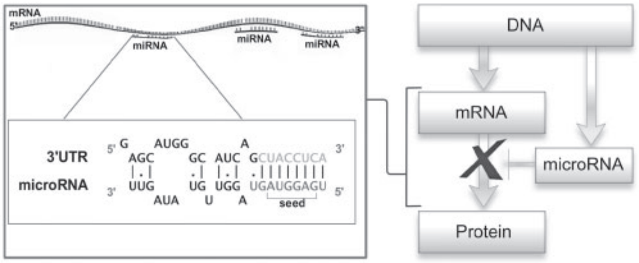


Fig. 1. The binding of a miRNA to a miTG. Multiple miRNAs may bind on the 3′UTR of a miTG. The seed sequence corresponds to six nucleotides at positions 2–7 of the miRNA sequence. The position where a miRNA binds to a miTG is called the MRE. miRNAs are transcribed mostly through Pol II from DNA. Protein coding genes are transcribed into mRNA molecules which then are translated to proteins. miRNAs integrate into the RISC complex and by binding to mRNA molecules they inhibit translation or induce mRNA degradation.

*To whom correspondence should be addressed.

as cardiovascular and neurodegenerative diseases (Hebert and De Strooper, 2007; Hebert *et al*., 2008; Zhang, 2008) as well as in human malignancies (Croce and Calin, 2005; Esquela-Kerscher and Slack, 2006; Fabbri *et al*., 2007; Gartel and Kandel, 2008; Garzon *et al*., 2006; Slack and Weidhaas, 2006). In particular, miRNAs are believed to be involved in many stages of cancer progression by both promoting and/or suppressing oncogenesis (He *et al*., 2005; Ivanovska *et al*., 2008; Lee and Dutta, 2007; Tagawa *et al*., 2007), tumor growth (Johnson *et al*., 2007; Si *et al*., 2007), invasion and metastasis (Asangani *et al*., 2008; Huang *et al*., 2008; Ma *et al*., 2007; Tavazoie *et al*., 2008; Zhu *et al*., 2008).

For many years, researchers have been analyzing microarray expression data of protein coding genes in different cancer types in order to identify specific expression signatures. The limited number of miRNAs, makes them an ideal candidate for this type of analysis. Currently, there are ∼700 human miRNAs registered in miRBase (Griffiths-Jones *et al*., 2008), and according to estimates their number may reach 1000 (Fig. 2). Analyzing their expression, several miRNA signatures have already been successfully associated with human cancers (Calin and Croce, 2006) such as leukemias (Calin and Croce, 2007; Landais *et al*., 2007), thyroid carcinomas (He *et al*., 2005), breast (Iorio *et al*., 2005), lung (Yanaihara *et al*., 2006) and pancreatic cancer (Lee *et al*., 2007).

2 EXPERIMENTAL IDENTIFICATION OF miRNA TARGETS

In order to analyze miRNA function, a large number of studies have been published that attempt to validate miRNA:mRNA interactions, using direct and indirect experimental methods. Direct methods allow the validation of specific miRNA:mRNA interactions, while indirect methods, based on high-throughput experiments such as microarrays and protein quantification experiments, provide an overview of changes in a larger number of gene products.

Direct validation of miRNA target genes is often based on the quantification of a reporter construct [e.g. Luciferase or Green Fluorescent Protein (GFP)] carrying the 3′UTR of the putative target gene after the introduction of a miRNA to the cell (Kiriakidou *et al*., 2004). Alternatively, quantitative RT–PCR can be used to monitor changes in mRNA levels after a miRNA has been introduced in a cell. Even though such methods can validate the miRNA:mRNA interaction, they fail to identify the specific miRNA recognition elements (MREs) responsible for the interaction. Such MREs can be identified using an integration of the reporter gene assay with site directed mutagenesis and/or by restoring the complementarity by mutating the miRNA sequence.

High-throughput techniques can provide information about global miRNA effects in cells and are based on measuring differential gene expression in the presence or absence of a miRNA in the cell. For the overexpression of a miRNA (Lim *et al*., 2005), expression constructs can be engineered using the mature miRNA, the precursor (hairpin) miRNA, or the pre-miRNA sequence for transfection *in vitro* or *in vivo*. Silencing of a miRNA can be accomplished by introducing chemically modified oligonucleotides perfectly complementary to the mature miRNA (Krutzfeldt *et al*., 2005) or by knocking down a miRNA gene. Until recently such gene expression levels changes have been monitored through gene expression microarrays (Krutzfeldt *et al*., 2005; Lim *et al*., 2005). These methods give significant information for miRNA targets where gene expression repression is caused by mRNA degradation (see also Supplementary Material), but is missing the targets where expression repression is caused by translation repression. Such targets were only recently identified using high-throughput proteomics methods (Baek *et al*., 2008; Selbach *et al*., 2008). In these studies, stable isotope labeling with amino acids in cell culture (SILAC) was applied and the protein expression levels for thousands of genes were measured. It should be noted that both methods provide indirect validation of targets. Recently, immunoprecipitation of RISC components has been used to identify mRNAs targeted by miRNAs (Beitzinger *et al*., 2007; Easow *et al*., 2007; Zhang *et al*., 2007). Moreover, high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation (HITS-CLIP) has been used (Chi *et al*., 2009) in order to identify and sequence specific miRNA binding sites on targeted mRNAs. Methods based on the measurement of differential expression of genes (microarrays, pSILAC), may contain many secondary and nonspecific effects and therefore the identified group of target genes does not constitute a comprehensive list of miRNA targets. Such results should be rather treated as enriched in direct miRNA targets of a specific miRNA. HITS-CLIP on the other hand might also identify non-functional binding sites of RISC. Summarizing, high-throughput methods can provide a broad set of miRNA targets in a cell that are hard to identify using direct verification methods but are not as specific as direct validation methods.

The rapid development in the methods of the experimental validation of miRNA targets and the increased interest of many labs for the function of miRNAs has caused a dramatic increase of miRNA target genes (miTGs) with experimental evidence (Fig. 2). An up to date collection of such targets including information for both the validated interaction and the methods used can be found in TarBase (Papadopoulos *et al*., 2009), a manually curated database with currently more than 1300 miRNA:mRNA interactions in several species.

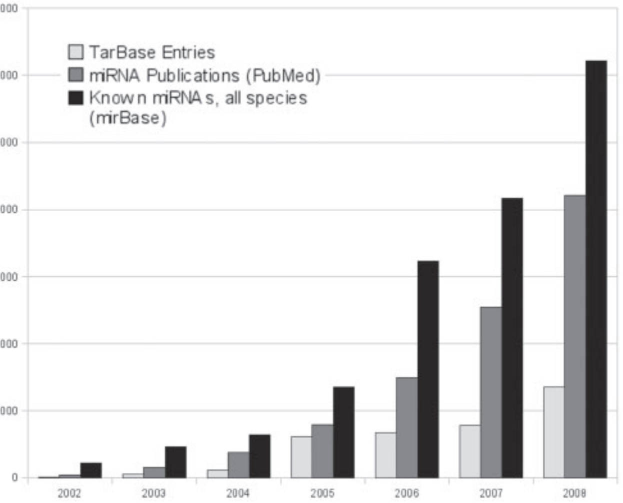


Fig. 2. The growth of known miRNA genes in miRBase database (black bars), the growth of miRNA related publications in PubMed (dark-gray bars) and the growth of the human experimentally determined miRNA target interactions in TarBase (light-gray bars).

3 OVERVIEW OF miRNA TARGET PREDICTION PROGRAMS

Despite the significant increase of experimentally validated miTGs the majority of miRNA targeted genes still remains unknown and computational target prediction programs remain the only source for a rapid identification of a putative miRNA target. Therefore, the development of computational target prediction programs goes hand in hand with the understanding of miRNA function. The first programs were developed back in 2003 shortly after it became evident that miRNAs are abundant in cells. Although a typical miRNA is ∼22 nucleotides (nt) long, several groups (Doench and Sharp, 2004; Kiriakidou *et al*., 2004) have shown experimentally that the nucleotides close to the 5′end of the miRNA are the most crucial for recognizing and binding to a target sequence. Additionally, a statistical analysis by Lewis *et al*. (2005) revealed that motifs in the 3′UTR of protein coding genes corresponding to nucleotides 2–7 of the miRNA are preferentially conserved in several species. These six nucleotides have been denoted as the 'seed' sequence of the miRNA (Fig. 1). However, later Krek *et al*. (2005) used seven nucleotides starting at position 1 or 2 of a miRNA to locate potential targets on the 3′UTR.

In the last years, several miRNA target prediction programs have been published (Sethupathy *et al*., 2006). The main prediction feature used in most of these programs is the sequence alignment of the miRNA seed to the 3′UTR of candidate target genes. Their specificity is usually increased by exploiting the evolutionary conservation of binding sites or by using additional features such as structural accessibility (Kertesz *et al*., 2007; Long *et al*., 2007), nucleotide composition (Grimson *et al*., 2007) or location of the binding sites within the 3′UTR (Baek *et al*., 2008; Gaidatzis *et al*., 2007; Grimson *et al*., 2007).

Here we summarize, in alphabetical order, eight of the most commonly used algorithms for miRNA target prediction for the human and mouse genome.

3.1 DIANA-microT 3.0

The DIANA-microT 3.0 (Maragkakis *et al*., 2009) algorithm is based on parameters calculated individually for each miRNA and each MRE depending on binding and conservation features. The prediction score of a miTG interaction is the weighted sum of the scores of conserved and non-conserved MREs on a gene. A signal to noise ratio (SNR) and a precision score are calculated for each interaction to provide an estimate of the false positive rate of each predicted miTG. Prediction data is available at http://microrna.gr/microT.

3.2 ElMMo

ElMMo (Gaidatzis *et al*., 2007) uses a general Bayesian method that scores the conservation of miRNA binding sites according to an evolutionary model that utilizes the assumed phylogenetic relationship among several species. Flat files of ElMMo target prediction data (v2, January 2008) are downloaded from http://www.mirz.unibas.ch/Computational_prediction_of_microRNA_targets_BULK.shtml. As suggested by the authors, a score threshold of 0.8 is used for high confidence in the comparisons.

3.3 miRanda

miRanda (John *et al*., 2004) uses a two-step approach for the identification of miRNA targets. First, the whole length of the miRNA is aligned against the 3′UTR sequence. Alignments that contain G:U wobble pairs are down-weighted accordingly. Second, for the highest scoring alignments, the thermodynamic stability of the complex is calculated and reported. Flat files of miRanda target prediction data are downloaded (January 2008) from: http://www.microrna.org/microrna/getDownloads.do.

3.4 miRBase

miRBase (Griffiths-Jones *et al*., 2008) uses the miRanda algorithm to identify potential binding sites for a given miRNA. Dynamic programming alignment is used to identify highly complementary sites. Strict complementarity at the 5′ seed region is demanded. Thermodynamic stability is estimated for each target site. For inclusion in the database, conservation of the target site at the exact same position in at least two species is needed. miRBase target prediction data is downloaded from http://microrna.sanger.ac.uk/cgi-bin/targets/v4/download.pl.

3.5 Pictar

Pictar (Lall *et al*., 2006) identifies two types of miRNA:target interactions: (i) those with perfect complementarity between the seed region of the miRNA (7 nt starting at position 1 or 2 of the miRNA's 5′end) and the 3′UTR target site and (ii) those for which the perfect complementarity is interrupted by at most one nucleotide bulge, mismatch, or G:U wobble. In both instances, the algorithm requires that the binding stability of the putative miRNA:target interaction, as measured by thermodynamic binding energy, exceeds a specified threshold. Once individual miRNA:target interactions are identified, the algorithm labels highly conserved (among 4 or 5 species) target sites as 'anchors' and filters out those 3′UTRs that do not harbor a specified number of anchors. A hidden Markov model is then used to score the likelihood of a 3′UTR being targeted by miRNAs in a combinatorial manner. These scores are computed for a set of species and combined to compute the final score. Since the bulk download files for Pictar on the UCSC Genome Browser are outdated, the target results are downloaded from the Pictar web page (http://pictar.org/) following the link for 'Predictions in vertebrates, flies and nematodes' (Lall *et al*., 2006). The four species conservation is used.

3.6 PITA

PITA (Kertesz *et al*., 2007) considers the effect of target site accessibility on the strength of miRNA repression. Essentially, for each target site, an energy-based measure that represents the difference between the free energy gained by the binding of the miRNA to the target and the free energy lost by unpairing the nucleotides within the target site itself is calculated. The energy used to unpair additional nucleotides flanking the target sites is also taken into account. A flat file with target prediction data is downloaded from http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html. The 'no 3_15' option in the PITA Targets Catalog version 5 (November 20, 2007) is used with the top targets identified as those with a score lower than −5.

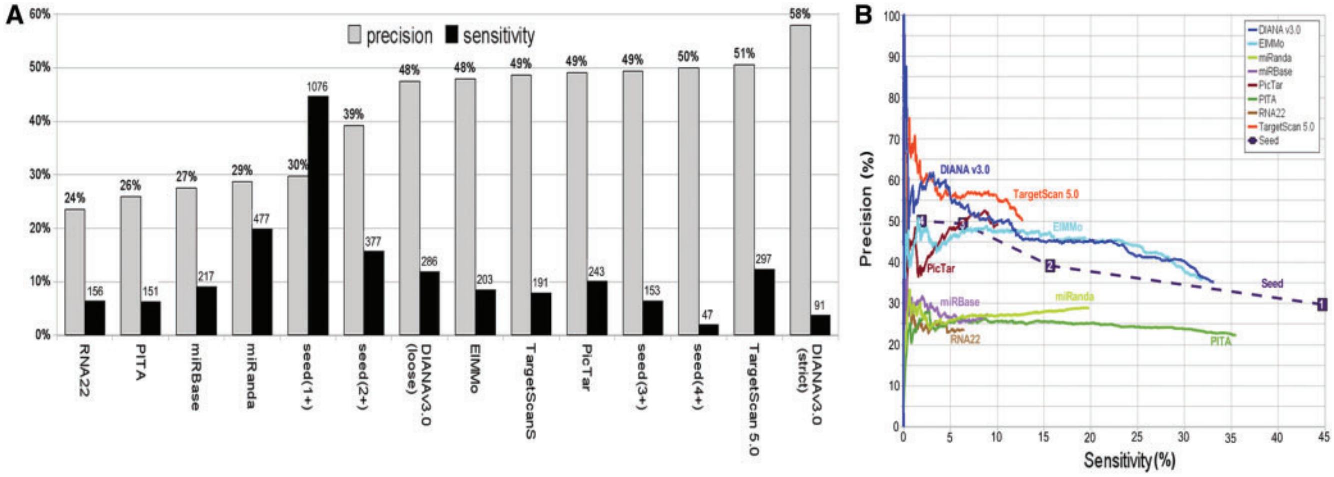


Fig. 3. Comparison of nine miRNA target prediction programs and the seed measure on the results provided by Selbach *et al.* (http://psilac.mdc-berlin.de). (**A**) The gray columns indicate precision (correctly predicted/total predicted) while the black columns show sensitivity (correctly predicted/total correct). The graph shows all targets above the score threshold of each program. A scatterplot of the same results is available in the Supplementary Materials. (**B**) A precision-receiver operating characteristic pROC (curve) showing the precision against the sensitivity of the miRNA target prediction programs. The seed measure has distinct values denoted as purple squares connected by a dotted line, where the numbers on the squares denote the minimum number of seeds per gene at each threshold. We annotate the four points having one to four seed matches.

3.7 RNA22

RNA22 (Miranda *et al.*, 2006) is a miRNA target prediction program that incorporates identifying redundant patterns in mature miRNA sequences. A second-order Markov chain is implemented to estimate the statistical significance of the identified patterns. The reverse complement of all miRNA patterns are then identified within 3′UTR sequences. A ‘Target Island’ is an area where many such reverse complement hits accumulate. miRNAs are paired to target islands and the strength of the pairing is calculated based on the free energy and the number of nucleotides involved. The target prediction data is downloaded from http://cbcsrv.watson.ibm.com/rna22_download_content.html. The date of the precompiled predictions is November 11, 2006.

3.8 TargetScan 5.0

TargetScan (Friedman *et al.*, 2009) predicts miRNA targets based on the identification of aligned seed matches and their conservation in several species. The overall scoring of a miRNA target site depends on the level of conservation, whether it binds to the miRNA on position 8 and/or whether it has an A at position 1, the distance of the target from the 3′UTR end and the AU composition of the flanking area. Data was downloaded from http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_50.

3.9 Simple seed measure

In this approach, genes are identified and sorted according to the number of occurrences of the hexamer complementary to the seed (nucleotides 2–7) of the miRNA in the 3′UTR sequence. Unless stated otherwise, all genes containing at least one instance of the seed were used in comparisons. When multiple annotated 3′UTR sequences were available for a gene, the longest one was used.

The user interfaces of the miRNA target prediction programs described above offer a variety of options to the user and are summarized in the Supplementary Material. We would like to mention here that only a few programs (DIANA-microT 3.0, TargetScan 5.0) offer the option to predict targets for user defined novel miRNAs, and some programs offer the option of a meta analysis through information regarding miRNA and mRNA expression or/and Gene Ontology (ElMMo, miRBase). At this point, we would like to point out that programs are not always up-to-date regarding the number of miRNAs and genes used. This number ranges currently from 178 to 675. A table with the number of miRNAs for which each program gives predictions can be found in the Supplementary Materials.

4 COMPARISON OF miRNA TARGET PREDICTION PROGRAMS

In the two recently published works (Baek *et al.*, 2008; Selbach *et al.*, 2008) that measured changes of protein levels after overexpression or underexpression of a miRNA, several miRNA target prediction programs are evaluated. Similarly, we tested here all miRNA target prediction programs mentioned above against genes proposed as targeted in Selbach *et al.* (Material and Methods section in Supplementary Material) In Figure 3, the results for 5 miRNAs are summarized. Nearly half of the down-regulated genes contain at least one occurrence of a miRNA specific seed sequence (Fig. 3A). We notice that a group of five programs (DIANA-microT 3.0, TargetScan 5.0, TargetScanS, Pictar and ElMMo) has a precision of ~50% with a sensitivity that ranges from 6 to 12%. All these programs rely heavily on the evolutionary conservation of the seed region or some small extensions of this region, and combine this information with other features that characterize miTGs.

Such features are detailed phylogenetic models to assess conservation, a miRNA specific SNR or a hidden Markov model to combine different MRE scores into a total miTG score (Fig. 1). Other programs include promising features like accessibility of the

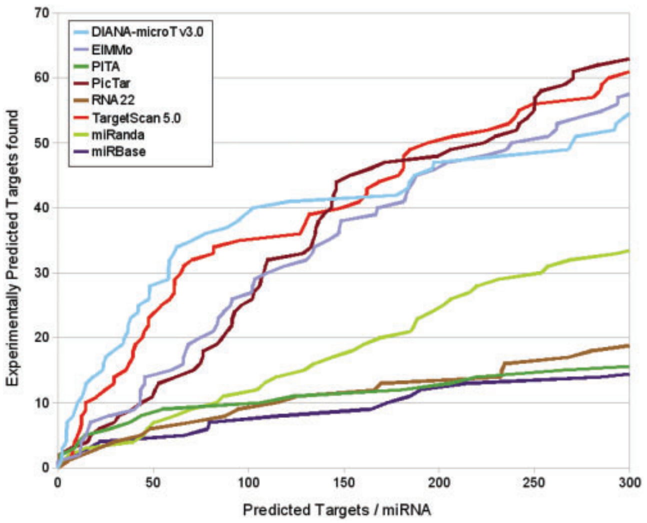


Fig. 4. Comparison of the miRNA target prediction programs on an experimentally supported miRNA target dataset. The number of correctly predicted targets is shown by different scores for increasing numbers of predicted targets per miRNA.

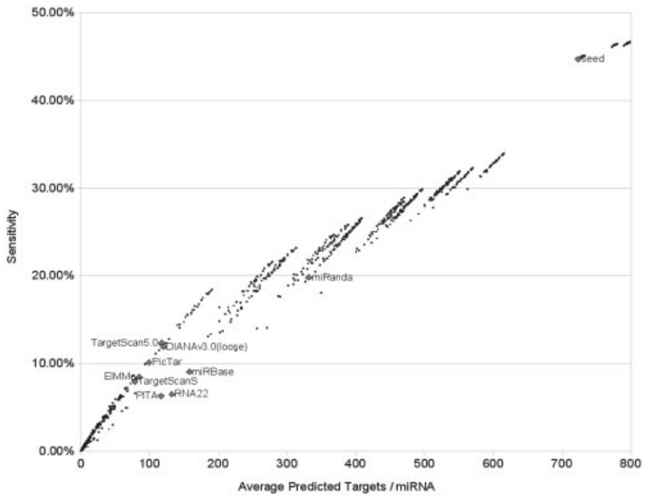


Fig. 5. Comparison of the combinations of several miRNA target prediction programs on the results provided by Selbach *et al.* The sensitivity of the prediction versus the number of predicted targets per miRNA is plotted. A larger version of this figure and an excel file with all sensitivity and specificity numbers can be found in the Supplementary Material.

binding site region, local concentration of redundant patterns of miRNA sequences, or thermodynamic stability but at the current stage they show lower predictive power. It has to be explored if these features in combination with other predictive methods can enhance target prediction.

We also investigate the very simple measure of counting the number of seed regions per gene. Nearly half of the down-regulated genes contain at least one occurrence of a miRNA specific seed sequence (Fig. 3A). Comparing the more sensitive prediction methods it can be noticed that the simple seed measure [Seed(1+) and Seed(2+)] outperforms other more complex computational methods, but fail when higher specificity is required [Seed(3+) and Seed(4+)]. Figure 3B presents the sensitivity and precision using different score cutoffs for all programs and the simple seed measure (see Supplementary Methods). The performance of the seed measure divides consistently the programs in two groups.

Further, we test the same programs with results obtained from overepxression of 2 miRNAs (hsa-mir-1 and hsa-mir-124) in HeLa cells and the subsequent measurement of mRNA levels using microarrays (Lim *et al.*, 2005) (see Supplementary Figs S3 and S4). For these data we compute the sensitivity measure at different levels for all programs. The results give a similar picture as discussed above (Supplementary Figs S3a and b, S4a and b).

A different test was performed for the same programs on a dataset of experimentally supported targets derived from TarBase (Papadopoulos *et al.*, 2009). This set includes 150 targets of 61 different miRNAs that were verified with direct experimental methods (available as Supplementary Material). The ranking of the prediction power of the tested programs shows the same order (Fig. 4).

To the non-expert, the choice of miRNA targets based on predictions by algorithms may seem like a daunting task. A natural inclination of a researcher is to assume that targets predicted by more than one algorithm are more accurate than other targets and thus leading to higher prediction precision. In a similar fashion, the union of different programs might improve the sensitivity. We test this by calculating all possible union and intersection combinations of the programs mentioned above (Fig. 5) for the high-throughput data provided for five miRNAs by Selbach *et al*. It can be observed that in most cases an accurate algorithm is better than a combination of predictions. Many of the combinations perform worse than the prediction of a single algorithm. The reason is that better specificity of a combination is achieved by a higher price for the sensitivity. Similar results are obtained using pairwise combinations of programs on the expression array data set (see Supplementary Fig. S3c).

5 FUTURE CHALLENGES OF miRNA TARGET IDENTIFICATION

The arrival of high-throughput proteomics analysis allows researchers to obtain a wider view of miRNA function in cells. Such data may help in the identification of new rules that govern miRNA function and also serve as training sets for applications based on machine learning approaches. As expression data is becoming increasingly available, it will be soon possible to train adaptive algorithms that will highlight additional rules for miRNA interactions with targeted genes. This notion is in line with the results provided in a recent publication that describes a miRNA target prediction method in *C.elegans*, mirWIP (Hammell *et al.*, 2008), which uses experimental data to define miTG prediction rules. Specifically, data from an immunoprecipitation experiment which identifies mRNAs targeted by the RISC were used and filters based on the structural accessibility of the target site, total energy of the miRNA-target hybridization as well as base pairing of the driver sequence were combined for the prediction of miTGs.

Another interesting field opening in miRNA target prediction, is the elucidation of the combinatorial effect of miRNAs. It is widely accepted that several miRNAs are co-regulated in miRNA gene clusters and are transcribed together. Additionally, levels of several miRNAs may be correlated as markers for disease, indicating a co-regulation by more than one miRNAs. Therefore, two main questions may be asked: how do multiple miRNAs affect a single gene, and how do multiple miRNAs regulate a biological pathway or disease. High-throughput experiments involving the knock-out or overexpression of several miRNAs simultaneously as well as independently, could produce the data needed in order to tackle the first question. The second question requires more complex computational approaches that will precisely identify and predict miRNA regulatory networks and will model the interplay between miRNAs (Ivanovska and Cleary, 2008).

Traditionally, the 3′UTR has been thought of as the main region of miRNA binding. However, from as early as 2004 (Kloosterman *et al*., 2004), there have been reports that miRNA-binding sites could be functional even when artificially placed inside coding regions. In an important article laying basic rules for miRNA binding (Lewis *et al*., 2005), miRNA targeting was also detected in open reading frames of protein coding genes. More recently, the effect of introducing miRNA target sites into the 5′UTR of luciferase reporter mRNAs was extensively studied (Lytle *et al*., 2007) and naturally occurring miRNA targets in the amino acid coding sequence of mouse genes were experimentally identified (Tay *et al*., 2008). These findings indicate that miRNAs could target mRNAs by binding to positions outside the 3′UTR but it is still believed that these binding sites are scarce (Baek *et al*., 2008). However, it is possible that miRNAs act in these regions by different mechanisms and/or binding rules and therefore are hard to identify. Specifically, miRNA target prediction in coding regions would pose the difficulty of high background conservation and biased nucleotide composition.

6 CONCLUSION

Results produced by recently developed high throughput experimental techniques suggest that miRNAs have a broad impact on cellular processes. Moreover, the availability of such data allows for extensive benchmarking of existing target prediction algorithms. These benchmarks reveal that even the most sensitive programs fail to identify a large part of the targeted genes.

We believe that the dramatic progress in high throughput experimental methods will soon lead to significant qualitative and quantitative improvements in the characterization of miRNA regulation.

This will allow the development of more powerful algorithms from the statistical or machine learning field trained on such high throughput data. These methods will likely identify novel prediction rules and optimize those currently used, to create more accurate models of the underlying biological phenomena.

Closing we would like to apologize to the large number of groups working in this field whose work is not included in this review due to size limitations.

Conflict of Interest: none declared.

REFERENCES

Aravin,A.A. *et al*. (2003) The small RNA profile during Drosophila melanogaster development. *Dev. Cell*, **5**, 337–350.

Asangani,I.A. *et al*. (2008) MicroRNA-21 (miR-21) post-transcriptionally downregulates tumor suppressor Pdcd4 and stimulates invasion, intravasation and metastasis in colorectal cancer. *Oncogene*, **27**, 2128–2136.

Baek,D. *et al*. (2008) The impact of microRNAs on protein output. *Nature*, **455**, 64–71.

Baroukh,N. *et al*. (2007) MicroRNA-124a regulates Foxa2 expression and intracellular signaling in pancreatic beta-cell lines. *J. Biol. Chem.*, **282**, 19575–19588.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Beitzinger,M. *et al*. (2007) Identification of human microRNA targets from isolated argonaute protein complexes. *RNA Biol.*, **4**, 76–84.

Calin,G.A. and Croce,C.M. (2006) MicroRNA signatures in human cancers. *Nat. Rev. Cancer*, **6**, 857–866.

Calin,G.A. and Croce,C.M. (2007) Investigation of microRNA alterations in leukemias and lymphomas. *Methods Enzymol.*, **427**, 193–213.

Chen,C.Z. *et al*. (2004) MicroRNAs modulate hematopoietic lineage differentiation. *Science*, **303**, 83–86.

Chi,S.W. *et al*. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, **460**, 479–486.

Croce,C.M. and Calin,G.A. (2005) miRNAs, cancer, and stem cell division. *Cell*, **122**, 6–7.

Denli,A.M. *et al*. (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature*, **432**, 231–235.

Doench,J.G. and Sharp,P.A. (2004) Specificity of microRNA target selection in translational repression. *Genes Dev.*, **18**, 504–511.

Easow,G. *et al*. (2007) Isolation of microRNA targets by miRNP immunopurification. *RNA*, **13**, 1198–1204.

Esquela-Kerscher,A. and Slack,F.J. (2006) Oncomirs – microRNAs with a role in cancer. *Nat. Rev. Cancer*, **6**, 259–269.

Fabbri,M. *et al*. (2007) Regulatory mechanisms of microRNAs involvement in cancer. *Expert. Opin. Biol. Ther.*, **7**, 1009–1019.

Friedman,R.C. *et al*. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.

Gaidatzis,D. *et al*. (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, **8**, 69.

Gartel,A.L. and Kandel,E.S. (2008) miRNAs: little known mediators of oncogenesis. *Semin. Cancer Biol.*, **18**, 103–110.

Garzon,R. *et al*. (2006) MicroRNA expression and function in cancer. *Trends Mol. Med.*, **12**, 580–587.

Griffiths-Jones,S. *et al*. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.

Grimson,A. *et al*. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.

Hammell,M. *et al*. (2008) mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *Nat. Methods*.

He,H. *et al*. (2005) The role of microRNA genes in papillary thyroid carcinoma. *Proc. Natl Acad. Sci. USA*, **102**, 19075–19080.

He,L. *et al*. (2005) A microRNA polycistron as a potential human oncogene. *Nature*, **435**, 828–833.

Hebert,S.S. and De Strooper,B. (2007) Molecular biology. miRNAs in neurodegeneration. *Science*, **317**, 1179–1180.

Hebert,S.S. *et al*. (2008) Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/beta-secretase expression. *Proc. Natl Acad. Sci. USA*, **105**, 6415–6420.

Huang,Q. *et al*. (2008) The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis. *Nat. Cell Biol.*, **10**, 202–210.

Iorio,M.V. *et al*. (2005) MicroRNA gene expression deregulation in human breast cancer. *Cancer Res.*, **65**, 7065–7070.

Ivanovska,I. and Cleary,M.A. (2008) Combinatorial microRNAs: working together to make a difference. *Cell Cycle*, **7**, 3137–3142.

Ivanovska,I. *et al*. (2008) MicroRNAs in the miR-106b family regulate p21/CDKN1A and promote cell cycle progression. *Mol. Cell Biol.*, **28**, 2167–2174.

John,B. *et al*. (2004) Human microRNA targets. *PLoS Biol.*, **2**, e363.

Johnson,C.D. *et al*. (2007) The let-7 microRNA represses cell proliferation pathways in human cells. *Cancer Res.*, **67**, 7713–7722.

Kertesz,M. *et al*. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.

Kidner,C.A. and Martienssen,R.A. (2005) The developmental role of microRNA in plants. *Curr. Opin. Plant Biol.*, **8**, 38–44.

Kiriakidou,M. *et al*. (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes Dev.*, **18**, 1165–1178.

Krutzfeldt,J. *et al*. (2005) Silencing of microRNAs in vivo with 'antagomirs'. *Nature*, **438**, 685–689.

Lagos-Quintana,M. *et al*. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.

Lall,S. *et al*. (2006) A genome-wide map of conserved microRNA targets in C.elegans. *Curr. Biol.*, **16**, 460–471.

Landais,S. *et al*. (2007) Oncogenic potential of the miR-106-363 cluster and its implication in human T-cell leukemia. *Cancer Res.*, **67**, 5699–5707.

Lau,N.C. *et al*. (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, **294**, 858–862.

Lee,E.J. *et al*. (2007) Expression profiling identifies microRNA signature in pancreatic cancer. *Int. J. Cancer*, **120**, 1046–1054.

Lee,R.C. and Ambros,V. (2001) An extensive class of small RNAs in Caenorhabditis elegans. *Science*, **294**, 862–864.

Lee,R.C. *et al*. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.

Lee,Y.S. and Dutta,A. (2007) The tumor suppressor microRNA let-7 represses the HMGA2 oncogene. *Genes Dev.*, **21**, 1025–1030.

Lewis,B.P. *et al*. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.

Lim,L.P. *et al*. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, **433**, 769–773.

Liu,J. *et al*. (2004) Argonaute2 is the catalytic engine of mammalian RNAi. *Science*, **305**, 1437–1441.

Long,D. *et al*. (2007) Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol.*, **14**, 287–294.

Lu,Y. *et al*. (2007) Transgenic over-expression of the microRNA miR-17-92 cluster promotes proliferation and inhibits differentiation of lung epithelial progenitor cells. *Dev. Biol.*, **310**, 442–453.

Lytle,J.R. *et al*. (2007) Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. *Proc. Natl Acad Sci. USA*, **104**, 9667–9672.

Ma,L. *et al*. (2007) Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. *Nature*, **449**, 682–688.

Maragkakis,M. *et al*. (2009) Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics*, **10**, 295.

Miranda,K.C. *et al*. (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203–1217.

Papadopoulos,G.L. *et al*. (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.*, **37**, D155–D158.

Reinhart,B.J. *et al*. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, **403**, 901–906.

Selbach,M. *et al*. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.

Sethupathy,P. *et al*. (2006) A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat. Methods*, **3**, 881–886.

Si,M.L. *et al*. (2007) miR-21-mediated tumor growth. *Oncogene*, **26**, 2799–2803.

Slack,F.J. and Weidhaas,J.B. (2006) MicroRNAs as a potential magic bullet in cancer. *Future Oncol.*, **2**, 73–82.

Tagawa,H. *et al*. (2007) Synergistic action of the microRNA-17 polycistron and Myc in aggressive cancer development. *Cancer Sci.*, **98**, 1482–1490.

Tavazoie,S.F. *et al*. (2008) Endogenous human microRNAs that suppress breast cancer metastasis. *Nature*, **451**, 147–152.

Tay,Y. *et al*. (2008) MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, **455**, 1124–1128.

Wienholds,E. *et al*. (2005) MicroRNA expression in zebrafish embryonic development. *Science*, **309**, 310–311.

Yanaihara,N. *et al*. (2006) Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell*, **9**, 189–198.

Yi,R. *et al*. (2006) Morphogenesis in skin is governed by discrete sets of differentially expressed microRNAs. *Nat. Genet.*, **38**, 356–362.

Zhang,C. (2008) MicroRNAs: role in cardiovascular biology and disease. *Clin. Sci.*, **114**, 699–706.

Zhang,L. *et al*. (2007) Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. *Mol. Cell*, **28**, 598–613.

Zhu,S. *et al*. (2008) MicroRNA-21 targets tumor suppressor genes in invasion and metastasis. *Cell Res.*, **18**, 350–359.

7.2. Online resources for miRNA analysis

In the following review we discuss available online resources for miRNA genomics, gene finding, target prediction and functional analysis in an attempt to assist researchers who wish to acquaint themselves with miRNA biology and the available online tools specifically designed for miRNA analysis. The review was published in Alexiou *et al.* (Alexiou, Maragkakis et al. 2011)

Online resources for microRNA analysis

Panagiotis Alexiou, Manolis Maragkakis, Artemis G. Hatzigeorgiou

Biomedical Science Research Center Alexander Fleming, Vari, Greece

Abstract

The use of online tools for bioinformatics analyses is becoming increasingly widespread. Resources specific to the field of microRNAs are available, varying in scope and usability. Online tools are the most useful for casual as well as power users since they need no installation, are hardware independent and are used mostly through graphic user interfaces and links to external sources. Here, we present an overview of useful online resources that have to do with microRNA genomics, gene finding, target prediction and functional analysis.

Introduction

microRNAs are post-transcriptional regulatory molecules which belong in a recently identified group of short, 20-25 nucleotides long sequences of single-stranded, non-coding RNAs. microRNAs are produced by longer RNA precursors (pre-microRNA) whose length reaches approximately 100 nt. These precursors usually form an imperfect stem-loop structure (hairpin) and are in turn derived from longer primary RNA transcripts which can be thousands of nucleotides long and can contain several hairpins in transcriptional clusters.[1]

Generally, microRNA functionality derives from their base pairing on expressed mRNA molecules, usually on the 3'UTR but also on the coding sequence. This pairing, for animal microRNA in contrast to plant microRNAs, is rarely complete along the full length of the microRNA and can lead to degradation of the corresponding mRNA or to its translational repression.[2]

The discovery of microRNAs in the early 90s[3] and their subsequent connection with a wide array of developmental programs and disease, has come in a time when bioinformatic techniques are becoming widespread, and the web all pervasive. Resources used by microRNA researchers on the web are numerous and continuously in flux. Here we present some of the most commonly used online resources in four categories sorted in alphabetical order per category (Figure 1, Table 1).

Genomics

This category contains resources concerning genomic locations of microRNA primary transcripts, microRNA transcriptional clusters and genomic features associated with microRNAs such as transcription start sites and transcription factor binding sites near microRNA transcripts.

MiRBase 16.0

MiRBase 16.0[4] is a repository where newly discovered microRNAs are deposited and unique identification numbers are given. The basic unit of the database is the microRNA hairpin, with genomic location, sequence, references provided for hairpins in several species. The location and sequence of mature microRNAs on each hairpin is also provided. The database is searchable via an online interface, or can be downloaded as flat files and accessed offline.

miRGen 2.0

miRGen 2.0[5] is a database that provides information on the genomic position and near-by features of human and mouse microRNA transcripts and cotranscribed microRNA clusters. Experimentally predicted transcription start sites and nearby predicted transcription factor binding sites are provided. Additionally, expression profiles of microRNAs in several tissues and cell lines, single nucleotide polymorphism locations, microRNA target prediction on protein coding genes and mapping of microRNA targets of co-regulated microRNAs on biological pathways are also integrated into the database and user interface.

microRNA Genes

The identification of novel microRNA genes,[6] generally starts from the discovery of the distinctive hairpin structures that pre-microRNAs produce. With the onset of high-throughput experimental methods for the discovery of microRNA genes, the rate of identification of putative hairpin structures is ever increasing (Figure 2). There is a variety of online and offline tools for the prediction of the location of pre-microRNA hairpins in given sequences or genomic locations.

Among the on-line tools: miRNASVM[7] is a machine learning classifier that predicts the processing sites for Drosha, the Class 2 RNase III enzyme that processes pre-microRNAs. The classifier attempts to find 5′ Drosha processing sites in hairpins that are candidate microRNAs thus attempting to separate true from false microRNA hairpin predictions.

Correspondence: Panagiotis Alexiou,
B.S.R.C. Alexander Fleming 34 Fleming Street,
16672, Vari, Greece.
E-mail: pan.alexiou@fleming.gr

Key words: microRNA, bioinformatics, online resources, computational, small RNA.

Conflict of interest: the authors report no conflicts of interest.

Received for publication: 15 November 2010.
Revision received: 14 February 2011.
Accepted for publication: 14 February 2011.

Journal of Nucleic Acids Investigation 2011; 2:e4
doi:10.4081/jnai.2011.e4

ProMiR II[8] is a web-server that identifies microRNA hairpin structures in given sequences. It consists of three distinct programs. One searches for novel microRNA hairpins near known microRNAs, one predicts hairpins near a candidate sequence, and the last one is more general, using a moving window approach to scan larger sequences. Several parameters and thresholds can be set by the user.

Targeting

Resources for validated microRNA targets

Experimental validation of microRNA targets has been progressing in bounds in the past few years. Besides the more direct methods of target validation employing luciferase constructs and other traditional molecular biology methods, a great increase in high-throughput validation methods has been evident in the past few years.

miRecords

miRecords[9] is an integrated resource for animal microRNA-target interactions. The Validated Targets component of this resource hosts a manually curated database of experimentally validated microRNA-target interactions with systematic documentation of experimental support for each interaction. The current release of this database includes 1135 records of validated microRNA-target interactions between 301 microRNAs and 902 target genes in seven animal species. The Predicted Targets component of miRecords stores predicted microRNA targets produced by 11 established microRNA target prediction programs.

Tarbase

Tarbase[10] is a database which houses a manually curated collection of experimentally supported microRNA targets in several species. The current version includes more than 1300 experimentally supported targets. Each target site is described by the microRNA that binds it, the gene in which it occurs, the nature of the experiments that were conducted to test it and other factors. The whole database can be accessed online or downloaded.

Resources for microRNA target prediction

Although it is becoming increasingly easier to experimentally validate microRNA targets of interest, the computational prediction of microRNA targets is still relevant. The use of novel high-throughput experimental methods allows researchers to obtain a wider range of known microRNA targets. Such data will probably help in the identification of new rules that govern microRNA function and also serve as training sets for applications based on machine learning approaches. As expression data is becoming increasingly available, it will be soon possible to train adaptive algorithms that will highlight additional rules for miRNA interactions with targeted genes. However, to date, most microRNA target prediction programs are based on fixed rules. Since the field of microRNA target prediction is very fast changing and competitive with large differ-

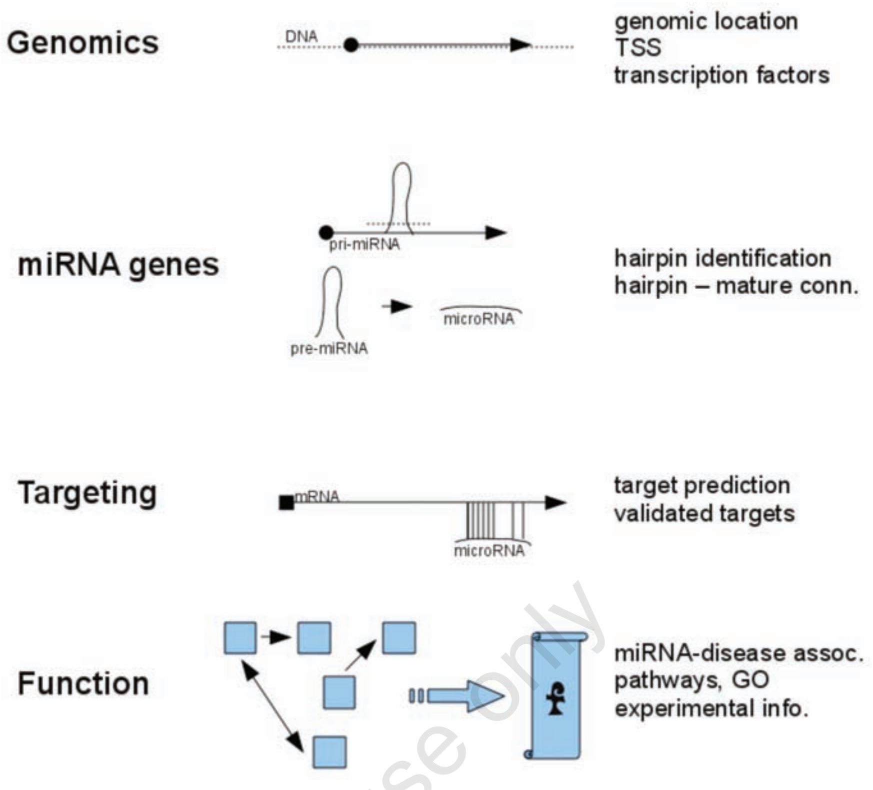


Figure 1. Online resources for microRNA analysis can be roughly divided in four categories. Genomic resources have to do with the genomic location and transcriptional interplay of microRNA genes. microRNA gene resources predict the hairpin structures associated with microRNAs. Targeting resources store experimentally validated or computationally predicted targets. Function resources show association of microRNAs with disease or function in general and of experimental results with microRNA deregulation.

Table 1. The web address of each of the online resources discussed here.

Genomics		
miRBase	www.mirbase.org	Griffiths-Jones, 2006
MiRGen 2.0	www.microrna.gr/mirgen	Alexiou *et al.*, 2010
Pre-miRNA Prediction		
ProMiR II	cbit.snu.ac.kr/~ProMiR2	Nam *et al.*, 2006
miRNASVM	demo1.interagon.com/miRNA/cgi-bin/MiRNASVM.cgi	Helvik *et al.*, 2006
Targeting (validated)		
Tarbase	www.microrna.gr/tarbase	Papadopoulos *et al.*, 2009
miRecords	miRecords.umn.edu/miRecords	Xiao *et al.*, 2009
Targeting (predicted)		
DIANA-microT 3.0	www.microrna.gr/microT	Maragkakis *et al.*, 2009
miRanda-mirSVR	www.microrna.org/microrna	Betel *et al.*, 2010
MicroCosm	www.ebi.ac.uk/enright-srv/microcosm	Griffiths-Jones *et al.*, 2008
Pictar	pictar.mdc-berlin.de	Lall *et al.*, 2006
PITA	genie.weizmann.ac.il/pubs/mir07	Kertesz *et al.*, 2008
TargetScan 5	www.targetscan.org	Friedman *et al.*, 2009
Function (miRNA process)		
miR2Disease	www.mir2disease.org	Jiang *et al.*, 2009
DIANA-mirPath	microrna.gr/mirpath	Papadopoulos *et al.*, 2009
miReg	www.iioab.webs.com/mireg.htm	Bahr *et al.*, 2010
Function (genelist miRNA)		
DIANA-mirExTra	www.microrna.gr/mirextra	Alexiou *et al.*, 2010
MiRonTop	www.microarray.fr:8080/miRonTop/index	Le Brigand *et al.*, 2010
Sylarray	www.ebi.ac.uk/enright/sylarray	Bartonicek *et al.*, 2010

ences in performance among programs. An overview of the performance of microRNA target prediction programs on high-throughput experimental data shows great discrepancies in the predictive strengths of each method.[11] Here we provide a brief overview of the most accurate and widely used microRNA target prediction programs.

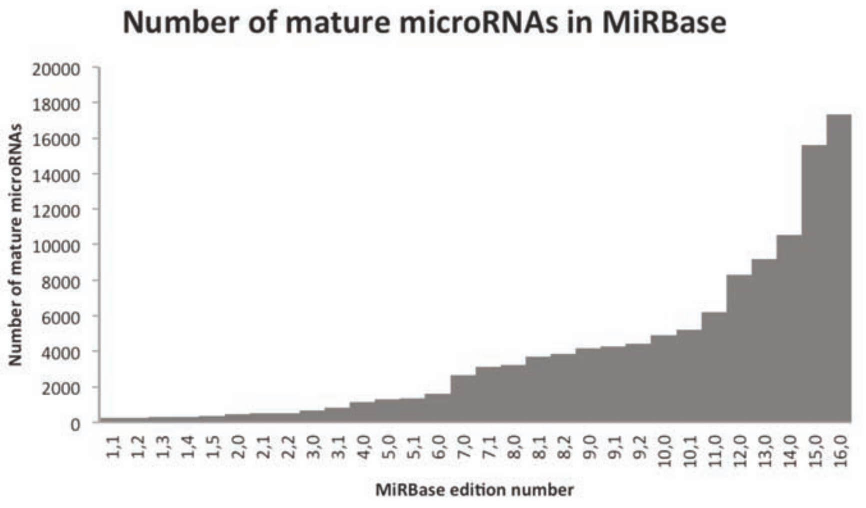


Figure 2. The growth of the number of microRNA sequences deposited in miRBase in the past decade.

DIANA-microT

DIANA-microT 3.0[12] is an algorithm based on several parameters calculated individually for each microRNA and it combines conserved and non-conserved microRNA recognition elements into a final prediction score. The program reports a signal to noise ratio and a precision score which help in the evaluation of the significance of the predicted results. The web server provides extensive information for predicted microRNA:target gene interactions providing extensive connectivity to online biological resources. Target gene and microRNA functions may be elucidated through automated bibliographic searches and functional information is accessible through KEGG pathways. The web server offers links to nomenclature, sequence and protein databases and users are facilitated by being able to search for targeted genes using different nomenclatures or functional features, such as the genes possible involvement in biological pathways.

MicroCosm

MicroCosm[13] (formely known as miRBase Targets) uses the miRanda algorithm to initially identify potential binding sites for a given microRNA. Dynamic programming alignment is used to identify highly complementary sites. Strict complementarity at the microRNA seed region is demanded. Thermodynamic stability is estimated for each target site. For inclusion in the database, conservation of the target site at the exact same position in at least two species is required.

miRanda - mirSVR

miRanda - mirSVR[14] mirSVR is a new machine learning method for ranking microRNA target sites by a down-regulation score. The algorithm trains a regression model on sequence and contextual features extracted from miRanda-predicted target sites. In a large-scale evaluation, miRanda-mirSVR is competitive with other target prediction methods in identifying target genes and predicting the extent of their downregulation at the mRNA or protein levels. Importantly, the method identifies a significant number of experimentally determined non-canonical and non-conserved sites.

PicTar

PicTar[15] identifies microRNA targets with perfect or imperfect complementarity in a 7nt seed region. Conservation is taken into account, and an HMM approach provides the final score by combining the multiple microRNA targets identified on the same gene. Although PicTar is still relatively accurate[11] when compared to other microRNA prediction algorithms, it has not been updated to the latest identified microRNAs since its initial release, thus missing hundreds of newly identified microRNAs.

PITA

PITA[16] incorporates binding site structural accessibility as a feature and does not take into account the evolutionary conservation of the binding site. Although, it is not among the best performing programs[11,17] it remains an interesting approach that shows high potential of being used along with other prediction programs that are more dependent to the evolutionary conservation of binding sites.

TargetScan 5.1

TargetScan 5.1[18] is one of the most widely used microRNA target prediction programs. In TargetScan, microRNA binding sites are predicted through the identification of seed matches on the 3'UTR of mRNAs and the assessment of their evolutionary conservation across several species. The overall scoring of a microRNA binding site denoted as *context score* depends on binding features such as whether the identified match involves binding on position 8 and/or whether it has an A at position 1, the localization of the binding site within the 3'UTR and the AU content of the area flanking the binding site. The final prediction score indicating whether a microRNA target a particular gene is calculated by summing the context scores of all corresponding binding sites identified on that gene 3'UTR.

Function

Association of microRNA with processes

A field of interest for many researchers is the function of microRNAs. When a list of microRNAs or a list of known or putative microRNA targets is given, a researcher would be interested to find out whether they are associated with any diseases or physiological processes.

DIANA-mirPath

DIANA-mirPath[19] is a web-based computational tool developed to identify molecular pathways potentially altered by the expression of single or multiple microRNAs. The software performs an enrichment analysis of multiple microRNA target genes comparing each set of microRNA targets to all known KEGG pathways. The combinatorial effect of co-expressed microRNAs in the modulation of a given pathway is taken into account by the simultaneous analysis of multiple microRNAs. The graphical output of the program provides an overview of the parts of the pathway modulated by microRNAs, facilitating the interpretation and presentation of the analysis results.

miR2Disease

miR2Disease[20] is a manually curated database which aims at providing a comprehensive resource of microRNA deregulation in various human diseases mined from published data. Users can also suggest associations based on publications.

miReg

miReg[21] is a manually curated microRNA

Regulation Resource that represents regulatory relationships between TFs, microRNAs and other regulators. The information is based on published resources.

Association of gene lists with microRNAs

Although microRNA expression levels may not be routinely measured in high-throughput experiments, a possible involvement of microRNAs in the deregulation of gene expression can be computationally predicted. Especially with the increasing use of high-throughput expression arrays and sequencing to measure deregulation in mRNA and even protein levels, these techniques for the computational prediction of the possible involvement of microRNAs are becoming more relevant.

DIANA-mirExTra

DIANA-mirExTra[22] allows the comparison of frequencies of microRNA associated motifs between sets of genes that can lead to the identification of microRNAs responsible for the deregulation of large numbers of genes.

MiRonTop

MiRonTop[23] is an online java web tool that integrates DNA microarrays or high-throughput sequencing data to identify the potential implication of microRNAs on a specific biological system. It also provides useful representations of the enrichment scores according to the position of the target site along the 3'-UTR, where the contribution of the sites located in the vicinity of the stop codon and of the polyA tail can be clearly highlighted. It provides different graphs of microRNA enrichment associated with up- or down-regulated transcripts and different summary tables about selections of mRNA targets and their functional annotations by Gene Ontology.

SylArray

SylArray[24] is a web-based analysis resource designed to examine influence of small RNAs on expression profiles. It can be used to find significant enrichment or depletion of microRNA or siRNA seed sequences from microarray expression data.

Conclusions

As an increasing number of resources in the fields related to microRNAs are becoming available it is of the greatest importance for users to know which resources are available in order to be able to choose which one to use for a specific task. The onset of the sequencing era in genomics brings great expectations for all the sub-fields of microRNA analysis. Databases will need to scale accordingly to increasing data loads and user requests, machine learning approaches will be used more often and in wider scopes and possibly user generated content could start being used. In closing, we would like to apologize to the large number of groups that work in this field whose work was impossible to be included in this review.

References

1. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004; 116:281-97.
2. Filipowicz W, Jaskiewicz L, Kolb FA, Pillai RS. Post-transcriptional gene silencing by siRNAs and miRNAs. Curr Opin Struct Biol 2005;15:331-41.
3. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993;75:843-54.
4. Griffiths-Jones S. miRBase: the microRNA sequence database. Methods Mol Biol 2006;342:129-38.
5. Alexiou P, Vergoulis T, Gleditzsch M, et al. miRGen 2.0: a database of microRNA genomic information and regulation. Nucleic Acids Res 2010;38:D137-41.
6. Oulas A, Reczko M, Poirazi P. MicroRNAs and Cancer inverted question mark The Search Begins! IEEE Trans Inf Technol Biomed 2008 Aug 15.
7. Helvik SA, Snove O, Jr, Saetrom P. Reliable prediction of Drosha processing sites improves microRNA gene prediction. Bioinformatics 2007;23:142-9.
8. Nam JW, Kim J, Kim SK, Zhang BT. ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. Nucleic Acids Res 2006;34:W455-8.
9. Xiao F, Zuo Z, Cai G, et al. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 2009;37:-D105-10.
10. Papadopoulos GL, Reczko M, Simossis VA, et al. The database of experimentally supported targets: a functional update of TarBase. Nucleic Acids Res 2009;37:D155-8.
11. Alexiou P, Maragkakis M, Papadopoulos GL, et al. Lost in translation: an assessment and perspective for computational microRNA target identification. Bioinformatics 2009;25:3049-55.
12. Maragkakis M, Alexiou P, Papadopoulos GL, et al. Accurate microRNA target prediction correlates with protein repression levels. BMC Bioinformatics 2009;10:295.
13. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res 2008;36: D154-8.
14. Betel D, Koppal A, Agius P, et al. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 2010;11:R90.
15. Lall S, Grun D, Krek A, et al. A genome-wide map of conserved microRNA targets in C. elegans. Curr Biol 2006;16:460-71.
16. Kertesz M, Iovino N, Unnerstall U, et al. The role of site accessibility in microRNA target recognition. Nat Genet 2007;39: 1278-84.
17. Selbach M, Schwanhausser B, Thierfelder N, et al. Widespread changes in protein synthesis induced by microRNAs. Nature 2008;455:58-63.
18. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 2009;19:92-105.
19. Papadopoulos GL, Alexiou P, Maragkakis M, et al. DIANA-mirPath: Integrating human and mouse microRNAs in pathways. Bioinformatics 2009;25:1991-3.
20. Jiang Q, Wang Y, Hao Y, et al. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res 2009;37:D98-104.
21. Barh D, Bhat D, Viero C. miReg: a resource for microRNA regulation. J Integr Bioinform 2010;7.
22. Alexiou P, Maragkakis M, Papadopoulos GL, et al. The DIANA-mirExTra web server: from gene expression data to microRNA function. PLoS One 2010;5:e9171.
23. Le Brigand K, Robbe-Sermesant K, Mari B, Barbry P. MiRonTop: mining microRNAs targets across large scale gene expression studies. Bioinformatics 2010;26:3131-2.
24. Bartonicek N, Enright AJ. SylArray: a web server for automated detection of miRNA effects from expression data. Bioinformatics 2010;26:2900-1.

8. GENERAL DISCUSSION AND CONCLUSION

In this thesis the first objective regarding miRNA target prediction has been addressed by two major releases of the microT program. Overall the second release denoted as microT-CDS introduces several important modifications which have resulted in an improved prediction performance over microT-v3.0. In addition, I participated in the development of a novel alignment algorithm for identification of putative miRNA binding sites. The algorithm is called CoProHMM and has been shown to perform more accurately than other alignment methods. Importantly, since it is a data driven approach its performance is expected to improve as new biological data become available. Finally, using target prediction for viral miRNAs it has become possible to identify that ebv-miR-BART6-5p regulates Dicer through multiple target sites in its 3'UTR. The predictions were experimentally verified and it has been suggested that mutation and A-to-I editing of viral miRNAs appear to be adaptive mechanisms that antagonize ebv-miR-BART6 activities and consequently affect viral latency. This work shows how computational tools can directly help in experimental design and thus provide valuable biological conclusions.

The second objective regarding miRNA functional analysis has been addressed by developing a Web server which offers an interface between bioinformatics tools and researchers. Also it offers unique information regarding miRNA function and provides extensive information and wide connectivity to online biological resources in a user friendly interface. Later, following user requests I updated the server to support predictions for two additional widely studied species: *D. melanogaster* and *C. elegans*. Also, through bibliographic analysis we associated miRNAs to diseases providing important insights for the potential involvement of miRNAs in biological processes. Additionally, I participated in the extensive analysis of the nomenclature used to describe mature miRNAs along different miRBase versions and the extraction of the naming history of each miRNA. Using this information within the bibliographic searches it is possible to identify related publications regardless of possible nomenclature changes.

Also, in terms of the second objective, I contributed in the development of two additional programs. The first one, DIANA-mirExTra, aims in the identification of miRNAs involved in the differential expression of genes. The second one, DIANA-mirPath, aims in the assessment of miRNA involvement in biological pathways. Both of these tools are widely used by researchers. As an example, an early version of DIANA-mirExTra was used in a publication by Zhang *et al.* (Zhang, Volinia et al. 2008) where hsa-miR-495 was successfully identified as an important regulator in human epithelial ovarian cancer.

Overall the publications contained in this thesis have been cited 92 times since the first publication in 2009. This citation rate results in an h-index of 4. Additionally, the DIANA Web server which is accessible at www.microrna.gr has received more than 697,931 page views by more than 110,016 users from more than 60 countries (Figure 3) and it currently receives more than 30,000 page views by more than 5500 users per month.

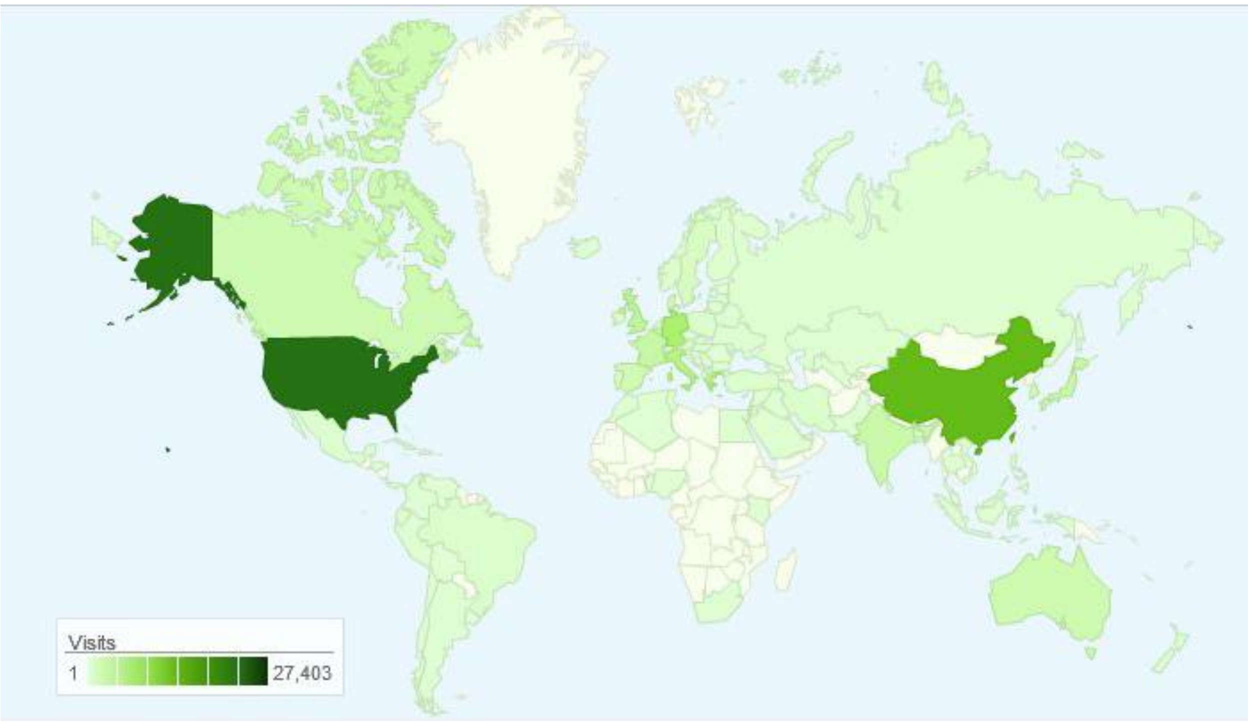


Figure 3: Graphical representation of the DIANA Web server usage statistics. Color darkness corresponds to the number of visits received from a particular region.

9. REFERENCES

Abdelmohsen, K., S. Srikantan, et al. (2008). "miR-519 reduces cell proliferation by lowering RNA-binding protein HuR levels." Proc Natl Acad Sci U S A **105**(51): 20297-20302.

Alexiou, P., M. Maragkakis, et al. (2011). "Online resources for microRNA analysis." Journal of Nucleic Acids Investigation **2**(1): 2-5.

Alexiou, P., M. Maragkakis, et al. (2009). "Lost in translation: an assessment and perspective for computational microRNA target identification." Bioinformatics **25**(23): 3049-3055.

Alexiou, P., M. Maragkakis, et al. (2010). "The DIANA-mirExTra web server: from gene expression data to microRNA function." PLoS One **5**(2): e9171.

Alexiou, P., T. Vergoulis, et al. (2010). "miRGen 2.0: a database of microRNA genomic information and regulation." Nucleic Acids Res **38**(Database issue): D137-141.

Almqvist, J., J. Zou, et al. (2005). "Functional interaction of Oct transcription factors with the family of repeats in Epstein-Barr virus oriP." J Gen Virol **86**(Pt 5): 1261-1267.

Ambros, V. (2001). "microRNAs: tiny regulators with great potential." Cell **107**(7): 823-826.

Aravin, A. A., M. Lagos-Quintana, et al. (2003). "The small RNA profile during Drosophila melanogaster development." Dev Cell **5**(2): 337-350.

Asangani, I. A., S. A. Rasheed, et al. (2008). "MicroRNA-21 (miR-21) post-transcriptionally downregulates tumor suppressor Pdcd4 and stimulates invasion, intravasation and metastasis in colorectal cancer." Oncogene **27**(15): 2128-2136.

Athanasiadis, A., A. Rich, et al. (2004). "Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome." PLoS Biol **2**(12): e391.

Baek, D., J. Villen, et al. (2008). "The impact of microRNAs on protein output." Nature **455**(7209): 64-71.

Bailey, T. L., N. Williams, et al. (2006). "MEME: discovering and analyzing DNA and protein sequence motifs." Nucleic Acids Res **34**(Web Server issue): W369-373.

Baldi, P. and S. r. Brunak (2001). Bioinformatics the machine learning approach. Adaptive computation and machine learning. Cambridge, Mass., MIT Press.

Barh, D., D. Bhat, et al. (2010). "miReg: a resource for microRNA regulation." J Integr Bioinform **7**(1).

Baroukh, N., M. A. Ravier, et al. (2007). "MicroRNA-124a regulates Foxa2 expression and intracellular signaling in pancreatic beta-cell lines." J Biol Chem **282**(27): 19575-19588.

Bartel, B. and D. P. Bartel (2003). "MicroRNAs: at the root of plant development?" Plant Physiol **132**(2): 709-717.

Bartel, D. P. (2004). "MicroRNAs: genomics, biogenesis, mechanism, and function." Cell **116**(2): 281-297.

Bartel, D. P. (2009). "MicroRNAs: target recognition and regulatory functions." Cell **136**(2): 215-233.

Barth, S., T. Pfuhl, et al. (2008). "Epstein-Barr virus-encoded microRNA miR-BART2 down-regulates the viral DNA polymerase BALF5." Nucleic Acids Res **36**(2): 666-675.

Bartonicek, N. and A. J. Enright (2010). "SylArray: a web server for automated detection of miRNA effects from expression data." Bioinformatics **26**(22): 2900-2901.

Basilio, C., A. J. Wahba, et al. (1962). "Synthetic polynucleotides and the amino acid code. V." Proc Natl Acad Sci U S A **48**: 613-616.

Bass, B. L. (2002). "RNA editing by adenosine deaminases that act on RNA." Annu Rev Biochem **71**: 817-846.

Beitzinger, M., L. Peters, et al. (2007). "Identification of human microRNA targets from isolated argonaute protein complexes." RNA Biol **4**(2): 76-84.

Berezikov, E. and R. H. Plasterk (2005). "Camels and zebrafish, viruses and cancer: a microRNA update." Hum Mol Genet **14 Spec No. 2**: R183-190.

Betel, D., A. Koppal, et al. (2010). "Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites." Genome Biol **11**(8): R90.

Bi, Y., G. Liu, et al. (2009). "MicroRNAs: novel regulators during the immune response." J Cell Physiol **218**(3): 467-472.

Blow, M., P. A. Futreal, et al. (2004). "A survey of RNA editing in human brain." Genome Res **14**(12): 2379-2387.

Blow, M. J., R. J. Grocock, et al. (2006). "RNA editing of human microRNAs." Genome Biol **7**(4): R27.

Borchert, G. M., W. Lanier, et al. (2006). "RNA polymerase III transcribes human microRNAs." Nat Struct Mol Biol **13**(12): 1097-1101.

Brennecke, J., A. Stark, et al. (2005). "Principles of microRNA-target recognition." PLoS Biol **3**(3): e85.

Bussemaker, H. J., H. Li, et al. (2001). "Regulatory element detection using correlation with expression." Nat Genet **27**(2): 167-171.

Cai, X., A. Schafer, et al. (2006). "Epstein-Barr virus microRNAs are evolutionarily conserved and differentially expressed." PLoS Pathog **2**(3): e23.

Calin, G. A. and C. M. Croce (2006). "MicroRNA signatures in human cancers." Nat Rev Cancer **6**(11): 857-866.

Calin, G. A. and C. M. Croce (2007). "Investigation of microRNA alterations in leukemias and lymphomas." Methods Enzymol **427**: 193-213.

Cerquides, J. and R. d. M. a. Lopez (2005). Robust Bayesian Linear Classifier Ensembles. 16th European Conference on Machine Learning.

Chen, C. Z., L. Li, et al. (2004). "MicroRNAs modulate hematopoietic lineage differentiation." Science **303**(5654): 83-86.

Chendrimada, T. P., R. I. Gregory, et al. (2005). "TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing." Nature **436**(7051): 740-744.

Cheng, A. M., M. W. Byrom, et al. (2005). "Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis." Nucleic Acids Res **33**(4): 1290-1297.

Chi, S. W., J. B. Zang, et al. (2009). "Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps." Nature **460**(7254): 479-486.

Choy, E. Y., K. L. Siu, et al. (2008). "An Epstein-Barr virus-encoded microRNA targets PUMA to promote host cell survival." J Exp Med **205**(11): 2551-2560.

Cosmopoulos, K., M. Pegtel, et al. (2009). "Comprehensive profiling of Epstein-Barr virus microRNAs in nasopharyngeal carcinoma." J Virol **83**(5): 2357-2367.

Croce, C. M. and G. A. Calin (2005). "miRNAs, cancer, and stem cell division." Cell **122**(1): 6-7.

Cullen, B. R. (2009). "Viral and cellular messenger RNA targets of viral microRNAs." Nature **457**(7228): 421-425.

Dabiri, G. A., F. Lai, et al. (1996). "Editing of the GLuR-B ion channel RNA in vitro by recombinant double-stranded RNA adenosine deaminase." EMBO J **15**(1): 34-45.

Denli, A. M., B. B. Tops, et al. (2004). "Processing of primary microRNAs by the Microprocessor complex." Nature **432**(7014): 231-235.

Ding, Y., C. Y. Chan, et al. (2004). "Sfold web server for statistical folding and rational design of nucleic acids." Nucleic Acids Res **32**(Web Server issue): W135-141.

Doench, J. G. and P. A. Sharp (2004). "Specificity of microRNA target selection in translational repression." Genes Dev **18**(5): 504-511.

Duursma, A. M., M. Kedde, et al. (2008). "miR-148 targets human DNMT3b protein coding region." RNA **14**(5): 872-877.

Easow, G., A. A. Teleman, et al. (2007). "Isolation of microRNA targets by miRNP immunopurification." RNA **13**(8): 1198-1204.

Edwards, R. H., A. R. Marquitz, et al. (2008). "Epstein-Barr virus BART microRNAs are produced from a large intron prior to splicing." J Virol **82**(18): 9094-9106.

Elcheva, I., S. Goswami, et al. (2009). "CRD-BP protects the coding region of betaTrCP1 mRNA from miR-183-mediated degradation." Mol Cell **35**(2): 240-246.

Elkon, R. and R. Agami (2008). "Removal of AU bias from microarray mRNA expression data enhances computational identification of active microRNAs." PLoS Comput Biol **4**(10): e1000189.

Enright, A. J., B. John, et al. (2003). "MicroRNA targets in Drosophila." Genome Biol **5**(1): R1.

Esquela-Kerscher, A. and F. J. Slack (2006). "Oncomirs - microRNAs with a role in cancer." Nat Rev Cancer **6**(4): 259-269.

Eulalio, A., E. Huntzinger, et al. (2008). "Getting to the root of miRNA-mediated gene silencing." Cell **132**(1): 9-14.

Fabbri, M., M. Ivan, et al. (2007). "Regulatory mechanisms of microRNAs involvement in cancer." Expert Opin Biol Ther **7**(7): 1009-1019.

Filipowicz, W., S. N. Bhattacharyya, et al. (2008). "Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?" Nat Rev Genet **9**(2): 102-114.

Flicek, P., B. L. Aken, et al. (2008). "Ensembl 2008." Nucleic Acids Res **36**(Database issue): D707-714.

Flicek, P., M. R. Amode, et al. (2011). "Ensembl 2011." Nucleic Acids Res **39**(Database issue): D800-806.

Forman, J. J., A. Legesse-Miller, et al. (2008). "A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence." Proc Natl Acad Sci U S A **105**(39): 14879-14884.

Forstemann, K., Y. Tomari, et al. (2005). "Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein." PLoS Biol **3**(7): e236.

Friedman, R. C., K. K. Farh, et al. (2009). "Most mammalian mRNAs are conserved targets of microRNAs." Genome Res **19**(1): 92-105.

Gaidatzis, D., E. van Nimwegen, et al. (2007). "Inference of miRNA targets using evolutionary conservation and pathway analysis." BMC Bioinformatics **8**: 69.

Gandy, S. Z., S. D. Linnstaedt, et al. (2007). "RNA editing of the human herpesvirus 8 kaposin transcript eliminates its transforming activity and is induced during lytic replication." J Virol **81**(24): 13544-13551.

Gartel, A. L. and E. S. Kandel (2008). "miRNAs: Little known mediators of oncogenesis." Semin Cancer Biol **18**(2): 103-110.

Garzon, R. and C. M. Croce (2008). "MicroRNAs in normal and malignant hematopoiesis." Curr Opin Hematol **15**(4): 352-358.

Garzon, R., M. Fabbri, et al. (2006). "MicroRNA expression and function in cancer." Trends Mol Med **12**(12): 580-587.

Gennarino, V. A., M. Sardiello, et al. (2009). "MicroRNA target prediction by expression analysis of host genes." Genome Res **19**(3): 481-490.

Godshalk, S. E., S. Bhaduri-McIntosh, et al. (2008). "Epstein-Barr virus-mediated dysregulation of human microRNA expression." Cell Cycle **7**(22): 3595-3600.

Grau, J., D. Arend, et al. (2010). Predicting miRNA targets utilizing an extended profile HMM. 25th German Conference on Bioinformatics 2010. 2010. 25th German Conference on Bioinformatics, Bone.

Grau, J., J. Keilwagen, et al. (2007). Supervised Posteriors for DNA-motif Classification. German Conference on Bioinformatics.

Gregory, R. I., T. P. Chendrimada, et al. (2005). "Human RISC couples microRNA biogenesis and posttranscriptional gene silencing." Cell **123**(4): 631-640.

Gregory, R. I., K. P. Yan, et al. (2004). "The Microprocessor complex mediates the genesis of microRNAs." Nature **432**(7014): 235-240.

Griffiths-Jones, S. (2006). "miRBase: the microRNA sequence database." Methods Mol Biol **342**: 129-138.

Griffiths-Jones, S. (2010). "miRBase: microRNA sequences and annotation." Curr Protoc Bioinformatics **Chapter 12**: Unit 12 19 11-10.

Griffiths-Jones, S., H. K. Saini, et al. (2008). "miRBase: tools for microRNA genomics." Nucleic Acids Res **36**(Database issue): D154-158.

Grimson, A., K. K. Farh, et al. (2007). "MicroRNA targeting specificity in mammals: determinants beyond seed pairing." Mol Cell **27**(1): 91-105.

Grundhoff, A., C. S. Sullivan, et al. (2006). "A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses." Rna **12**(5): 733-750.

Guo, H., N. T. Ingolia, et al. (2010). "Mammalian microRNAs predominantly act to decrease target mRNA levels." Nature **466**(7308): 835-840.

Hafner, M., M. Landthaler, et al. (2010). "Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP." Cell **141**(1): 129-141.

Han, J., Y. Lee, et al. (2004). "The Drosha-DGCR8 complex in primary microRNA processing." Genes Dev **18**(24): 3016-3027.

He, H., K. Jazdzewski, et al. (2005). "The role of microRNA genes in papillary thyroid carcinoma." Proc Natl Acad Sci U S A **102**(52): 19075-19080.

He, L., J. M. Thomson, et al. (2005). "A microRNA polycistron as a potential human oncogene." Nature **435**(7043): 828-833.

Hebert, S. S. and B. De Strooper (2007). "Molecular biology. miRNAs in neurodegeneration." Science **317**(5842): 1179-1180.

Hebert, S. S., K. Horre, et al. (2008). "Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/beta-secretase expression." Proc Natl Acad Sci U S A **105**(17): 6415-6420.

Heckerman, D., D. Geiger, et al. (1995). "Learning Bayesian networks:
The combination of knowledge and statistical data." In Machine Learning: 197–243.

Helvik, S. A., O. Snove, Jr., et al. (2007). "Reliable prediction of Drosha processing sites improves microRNA gene prediction." Bioinformatics **23**(2): 142-149.

Hiroki, E., J. Akahira, et al. (2010). "Changes in microRNA expression levels correlate with clinicopathological features and prognoses in endometrial serous adenocarcinomas." Cancer Sci **101**(1): 241-249.

Hislop, A. D., G. S. Taylor, et al. (2007). "Cellular responses to viral infection in humans: lessons from Epstein-Barr virus." Annu Rev Immunol **25**: 587-617.

Hsu, S. D., C. H. Chu, et al. (2008). "miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes." Nucleic Acids Res **36**(Database issue): D165-169.

Huang, Q., K. Gumireddy, et al. (2008). "The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis." Nat Cell Biol **10**(2): 202-210.

Hutvagner, G., J. McLachlan, et al. (2001). "A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA." Science **293**(5531): 834-838.

Iizasa, H., B. E. Wulff, et al. (2010). "Editing of Epstein-Barr virus-encoded BART6 microRNAs controls their dicer targeting and consequently affects viral latency." J Biol Chem **285**(43): 33358-33370.

Iorio, M. V., M. Ferracin, et al. (2005). "MicroRNA gene expression deregulation in human breast cancer." Cancer Res **65**(16): 7065-7070.

Ivanovska, I., A. S. Ball, et al. (2008). "MicroRNAs in the miR-106b family regulate p21/CDKN1A and promote cell cycle progression." Mol Cell Biol **28**(7): 2167-2174.

Ivanovska, I. and M. A. Cleary (2008). "Combinatorial microRNAs: working together to make a difference." Cell Cycle **7**(20): 3137-3142.

Jepson, J. E. and R. A. Reenan (2008). "RNA editing in regulating gene expression in the brain." Biochim Biophys Acta **1779**(8): 459-470.

Joels, C. S., B. D. Matthews, et al. (2005). "Evaluation of adhesion formation, mesh fixation strength, and hydroxyproline content after intraabdominal placement of polytetrafluoroethylene mesh secured using titanium spiral tacks, nitinol anchors, and polypropylene suture or polyglactin 910 suture." Surg Endosc **19**(6): 780-785.

John, B., A. J. Enright, et al. (2004). "Human MicroRNA targets." PLoS Biol **2**(11): e363.

Johnson, C. D., A. Esquela-Kerscher, et al. (2007). "The let-7 microRNA represses cell proliferation pathways in human cells." Cancer Res **67**(16): 7713-7722.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Res **28**(1): 27-30.

Kanehisa, M., S. Goto, et al. (2010). "KEGG for representation and analysis of molecular networks involving diseases and drugs." Nucleic Acids Res **38**(Database issue): D355-360.

Kanehisa, M., S. Goto, et al. (2004). "The KEGG resource for deciphering the genome." Nucleic Acids Res **32**(Database issue): D277-280.

Karolchik, D., A. S. Hinrichs, et al. (2007). "The UCSC Genome Browser." Curr Protoc Bioinformatics **Chapter 1**: Unit 1 4.

Kawahara, Y., M. Megraw, et al. (2008). "Frequency and fate of microRNA editing in human brain." Nucleic Acids Res **36**(16): 5270-5280.

Kawahara, Y., B. Zinshteyn, et al. (2007). "RNA editing of microRNA-151 blocks cleavage by the Dicer-TRBP complex." EMBO Reports **8**: 763-769.

Kawahara, Y., B. Zinshteyn, et al. (2007). "Redirection of silencing targets by adenosine-to-inosine editing of miRNAs." Science **315**(5815): 1137-1140.

Kawamata, T., H. Seitz, et al. (2009). "Structural determinants of miRNAs for RISC loading and slicer-independent unwinding." Nat Struct Mol Biol.

Kelly, G. L., A. E. Milner, et al. (2005). "Epstein-Barr virus nuclear antigen 2 (EBNA2) gene deletion is consistently linked with EBNA3A, -3B, and -3C expression in Burkitt's lymphoma cells and with increased resistance to apoptosis." J Virol **79**(16): 10709-10717.

Kertesz, M., N. Iovino, et al. (2007). "The role of site accessibility in microRNA target recognition." Nat Genet **39**(10): 1278-1284.

Khvorova, A., A. Reynolds, et al. (2003). "Functional siRNAs and miRNAs exhibit strand bias." Cell **115**(2): 209-216.

Kidner, C. A. and R. A. Martienssen (2005). "The developmental role of microRNA in plants." Curr Opin Plant Biol **8**(1): 38-44.

Kim, D. D., T. T. Kim, et al. (2004). "Widespread RNA editing of embedded alu elements in the human transcriptome." Genome Res **14**(9): 1719-1725.

Kim, V. N., J. Han, et al. (2009). "Biogenesis of small RNAs in animals." Nat Rev Mol Cell Biol **10**(2): 126-139.

Kiriakidou, M., P. T. Nelson, et al. (2004). "A combined computational-experimental approach predicts human microRNA targets." Genes Dev **18**(10): 1165-1178.

Krek, A., D. Grun, et al. (2005). "Combinatorial microRNA target predictions." Nat Genet **37**(5): 495-500.

Krogh, A., M. Brown, et al. (1994). "Hidden Markov models in computational biology. Applications to protein modeling." J Mol Biol **235**(5): 1501-1531.

Krutzfeldt, J., N. Rajewsky, et al. (2005). "Silencing of microRNAs in vivo with 'antagomirs'." Nature **438**(7068): 685-689.

Kumar, M. S., J. Lu, et al. (2007). "Impaired microRNA processing enhances cellular transformation and tumorigenesis." Nat Genet **39**(5): 673-677.

Lagos-Quintana, M., R. Rauhut, et al. (2001). "Identification of novel genes coding for small expressed RNAs." Science **294**(5543): 853-858.

Lall, S., D. Grun, et al. (2006). "A genome-wide map of conserved microRNA targets in C. elegans." Curr Biol **16**(5): 460-471.

Landais, S., S. Landry, et al. (2007). "Oncogenic potential of the miR-106-363 cluster and its implication in human T-cell leukemia." Cancer Res **67**(12): 5699-5707.

Landthaler, M., A. Yalcin, et al. (2004). "The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis." Curr Biol **14**(23): 2162-2167.

Lau, N. C., L. P. Lim, et al. (2001). "An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans." Science **294**(5543): 858-862.

Le Brigand, K., K. Robbe-Sermesant, et al. (2010). "MiRonTop: mining microRNAs targets across large scale gene expression studies." Bioinformatics **26**(24): 3131-3132.

Lee, E. J., Y. Gusev, et al. (2007). "Expression profiling identifies microRNA signature in pancreatic cancer." Int J Cancer **120**(5): 1046-1054.

Lee, R. C. and V. Ambros (2001). "An extensive class of small RNAs in Caenorhabditis elegans." Science **294**(5543): 862-864.

Lee, R. C., R. L. Feinbaum, et al. (1993). "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14." Cell **75**(5): 843-854.

Lee, Y., C. Ahn, et al. (2003). "The nuclear RNase III Drosha initiates microRNA processing." Nature **425**(6956): 415-419.

Lee, Y., K. Jeon, et al. (2002). "MicroRNA maturation: stepwise processing and subcellular localization." EMBO J **21**(17): 4663-4670.

Lee, Y., M. Kim, et al. (2004). "MicroRNA genes are transcribed by RNA polymerase II." EMBO J **23**(20): 4051-4060.

Lee, Y. S. and A. Dutta (2007). "The tumor suppressor microRNA let-7 represses the HMGA2 oncogene." Genes Dev **21**(9): 1025-1030.

Levanon, E. Y., E. Eisenberg, et al. (2004). "Systematic identification of abundant A-to-I editing sites in the human transcriptome." Nat Biotechnol **22**(8): 1001-1005.

Lewis, B. P., C. B. Burge, et al. (2005). "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." Cell **120**(1): 15-20.

Lewis, B. P., I. H. Shih, et al. (2003). "Prediction of mammalian microRNA targets." Cell **115**(7): 787-798.

Li, Q. and R. I. Gregory (2008). "MicroRNA regulation of stem cell fate." Cell Stem Cell **2**(3): 195-196.

Lim, L. P., N. C. Lau, et al. (2005). "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs." Nature **433**(7027): 769-773.

Linsley, P. S., J. Schelter, et al. (2007). "Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression." Mol Cell Biol **27**(6): 2240-2252.

Liu, J., M. A. Carmell, et al. (2004). "Argonaute2 is the catalytic engine of mammalian RNAi." Science **305**(5689): 1437-1441.

Lo, A. K., K. F. To, et al. (2007). "Modulation of LMP1 protein expression by EBV-encoded microRNAs." Proc Natl Acad Sci U S A **104**(41): 16164-16169.

Long, D., R. Lee, et al. (2007). "Potent effect of target structure on microRNA function." Nat Struct Mol Biol **14**(4): 287-294.

Lu, Y., J. M. Thomson, et al. (2007). "Transgenic over-expression of the microRNA miR-17-92 cluster promotes proliferation and inhibits differentiation of lung epithelial progenitor cells." Dev Biol **310**(2): 442-453.

Luciano, D. J., H. Mirsky, et al. (2004). "RNA editing of a miRNA precursor." Rna **10**(8): 1174-1177.

Lujambio, A., G. A. Calin, et al. (2008). "A microRNA DNA methylation signature for human cancer metastasis." Proc Natl Acad Sci U S A **105**(36): 13556-13561.

Lund, E., S. Guttinger, et al. (2004). "Nuclear export of microRNA precursors." Science **303**(5654): 95-98.

Luo, B., Y. Wang, et al. (2005). "Expression of Epstein-Barr virus genes in EBV-associated gastric carcinomas." World J Gastroenterol **11**(5): 629-633.

Ma, L., J. Teruya-Feldstein, et al. (2007). "Tumour invasion and metastasis initiated by microRNA-10b in breast cancer." Nature **449**(7163): 682-688.

MacKay, D. J. C. (1998). " Choice of Basis for Laplace Approximation." Machine Learning **33**(1): 77–86.

Maniataki, E. and Z. Mourelatos (2005). "A human, ATP-independent, RISC assembly machine fueled by pre-miRNA." Genes Dev **19**(24): 2979-2990.

Maragkakis, M., P. Alexiou, et al. (2009). "Accurate microRNA target prediction correlates with protein repression levels." BMC Bioinformatics **10**: 295.

Maragkakis, M., M. Reczko, et al. (2009). "DIANA-microT web server: elucidating microRNA functions through target prediction." Nucleic Acids Res **37**(Web Server issue): W273-276.

Maragkakis, M., T. Vergoulis, et al. (2011). "DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association." Nucleic Acids Res.

Melcher, T., S. Maas, et al. (1996). "A mammalian RNA editing enzyme." Nature **379**(6564): 460-464.

Miranda, K. C., T. Huynh, et al. (2006). "A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes." Cell **126**(6): 1203-1217.

Nam, J. W., J. Kim, et al. (2006). "ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs." Nucleic Acids Res **34**(Web Server issue): W455-458.

Nam, S., B. Kim, et al. (2008). "miRGator: an integrated system for functional annotation of microRNAs." Nucleic Acids Res **36**(Database issue): D159-164.

Nishikura, K. (2006). "Editor meets silencer: crosstalk between RNA editing and RNA interference." Nat Rev Mol Cell Biol: in press.

Ohman, M. (2007). "A-to-I editing challenger or ally to the microRNA process." Biochimie **89**(10): 1171-1176.

Okuda, S., T. Yamada, et al. (2008). "KEGG Atlas mapping for global analysis of metabolic pathways." Nucleic Acids Res **36**(Web Server issue): W423-426.

Oulas, A., M. Reczko, et al. (2008). "MicroRNAs and Cancer inverted question mark The Search Begins!" IEEE Trans Inf Technol Biomed.

Pagano, J. S., M. Blaser, et al. (2004). "Infectious agents and cancer: criteria for a causal relation." Semin Cancer Biol **14**(6): 453-471.

Papadopoulos, G. L., P. Alexiou, et al. (2009). "DIANA-mirPath: Integrating human and mouse microRNAs in pathways." Bioinformatics **25**(15): 1991-1993.

Papadopoulos, G. L., M. Reczko, et al. (2009). "The database of experimentally supported targets: a functional update of TarBase." Nucleic Acids Res **37**(Database issue): D155-158.

Pearson, G. R., J. Luka, et al. (1987). "Identification of an Epstein-Barr virus early gene encoding a second component of the restricted early antigen complex." Virology **160**(1): 151-161.

Petersen, C. P., M. E. Bordeleau, et al. (2006). "Short RNAs repress translation after initiation in mammalian cells." Mol Cell **21**(4): 533-542.

Pfeffer, S., A. Sewer, et al. (2005). "Identification of microRNAs of the herpesvirus family." Nat Methods **2**(4): 269-276.

Pfeffer, S., M. Zavolan, et al. (2004). "Identification of virus-encoded microRNAs." Science **304**(5671): 734-736.

Pillai, R. S. (2005). "MicroRNA function: multiple mechanisms for a tiny RNA?" RNA **11**(12): 1753-1761.

Rehmsmeier, M., P. Steffen, et al. (2004). "Fast and effective prediction of microRNA/target duplexes." RNA **10**(10): 1507-1517.

Reinhart, B. J., F. J. Slack, et al. (2000). "The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans." Nature **403**(6772): 901-906.

Samuel, C. E. (2001). "Antiviral actions of interferons." Clin Microbiol Rev **14**(4): 778-809, table of contents.

Schaefer, B. C., J. L. Strominger, et al. (1995). "Redefining the Epstein-Barr virus-encoded nuclear antigen EBNA-1 gene promoter and transcription initiation site in group I Burkitt lymphoma cell lines." Proc Natl Acad Sci U S A **92**(23): 10565-10569.

Schwarz, D. S., G. Hutvagner, et al. (2003). "Asymmetry in the assembly of the RNAi enzyme complex." Cell **115**(2): 199-208.

Selbach, M., B. Schwanhausser, et al. (2008). "Widespread changes in protein synthesis induced by microRNAs." Nature **455**(7209): 58-63.

Sethupathy, P., B. Corda, et al. (2006). "TarBase: A comprehensive database of experimentally supported animal microRNA targets." RNA **12**(2): 192-197.

Sethupathy, P., M. Megraw, et al. (2006). "A guide through present computational approaches for the identification of mammalian microRNA targets." Nat Methods **3**(11): 881-886.

Shin, C., J. W. Nam, et al. (2010). "Expanding the microRNA targeting code: functional sites with centered pairing." Mol Cell **38**(6): 789-802.

Si, M. L., S. Zhu, et al. (2007). "miR-21-mediated tumor growth." Oncogene **26**(19): 2799-2803.

Slack, F. J. and J. B. Weidhaas (2006). "MicroRNAs as a potential magic bullet in cancer." Future Oncol **2**(1): 73-82.

Smith, P. R. and B. E. Griffin (1992). "Transcription of the Epstein-Barr virus gene EBNA-1 from different promoters in nasopharyngeal carcinoma and B-lymphoblastoid cells." J Virol **66**(2): 706-714.

Sood, P., A. Krek, et al. (2006). "Cell-type-specific signatures of microRNAs on target mRNA expression." Proc Natl Acad Sci U S A **103**(8): 2746-2751.

Stefani, G. and F. J. Slack (2008). "Small non-coding RNAs in animal development." Nat Rev Mol Cell Biol **9**(3): 219-230.

Strobel, S. A., T. R. Cech, et al. (1994). "The 2,6-diaminopurine riboside.5-methylisocytidine wobble base pair: an isoenergetic substitution for the study of G.U pairs in RNA." Biochemistry **33**(46): 13824-13835.

Sun, G., J. Yan, et al. (2009). "SNPs in human miRNA genes affect biogenesis and function." RNA **15**(9): 1640-1651.

Tagawa, H., K. Karube, et al. (2007). "Synergistic action of the microRNA-17 polycistron and Myc in aggressive cancer development." Cancer Sci **98**(9): 1482-1490.

Takagi, S., M. Nakajima, et al. (2010). "MicroRNAs regulate human hepatocyte nuclear factor 4alpha, modulating the expression of metabolic enzymes and cell cycle." J Biol Chem **285**(7): 4415-4422.

Tavazoie, S. F., C. Alarcon, et al. (2008). "Endogenous human microRNAs that suppress breast cancer metastasis." Nature **451**(7175): 147-152.

Tay, Y., J. Zhang, et al. (2008). "MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation." Nature **455**(7216): 1124-1128.

Tokumaru, S., M. Suzuki, et al. (2008). "let-7 regulates Dicer expression and constitutes a negative feedback loop." Carcinogenesis **29**(11): 2073-2077.

van Dongen, S., C. Abreu-Goodger, et al. (2008). "Detecting microRNA binding and siRNA off-target effects from expression data." Nat Methods **5**(12): 1023-1025.

Venables, W. and B. Ripley (2002). Modern Applied Statistics with S., Springer.

Wang, X. (2006). "Systematic identification of microRNA functions by combining target prediction and expression profiling." Nucleic Acids Res **34**(5): 1646-1652.

Wang, X. (2008). "miRDB: a microRNA target prediction and functional annotation database with a wiki interface." RNA **14**(6): 1012-1017.

Wienholds, E., W. P. Kloosterman, et al. (2005). "MicroRNA expression in zebrafish embryonic development." Science **309**(5732): 310-311.

Winter, J., S. Jung, et al. (2009). "Many roads to maturity: microRNA biogenesis pathways and their regulation." Nat Cell Biol **11**(3): 228-234.

Xia, T., A. O'Hara, et al. (2008). "EBV microRNAs in primary lymphomas and targeting of CXCL-11 by ebv-mir-BHRF1-3." Cancer Res **68**(5): 1436-1442.

Xiao, F., Z. Zuo, et al. (2009). "miRecords: an integrated resource for microRNA-target interactions." Nucleic Acids Res **37**(Database issue): D105-110.

Yanaihara, N., N. Caplen, et al. (2006). "Unique microRNA molecular profiles in lung cancer diagnosis and prognosis." Cancer Cell **9**(3): 189-198.

Yang, W., T. P. Chendrimada, et al. (2006). "Modulation of microRNA processing and expression through RNA editing by ADAR deaminases." Nat Struct Mol Biol **13**(1): 13-21.

Yi, R., D. O'Carroll, et al. (2006). "Morphogenesis in skin is governed by discrete sets of differentially expressed microRNAs." Nat Genet **38**(3): 356-362.

Zhang, C. (2008). "MicroRNAs: role in cardiovascular biology and disease." Clin Sci (Lond) **114**(12): 699-706.

Zhang, L., L. Ding, et al. (2007). "Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2." Mol Cell **28**(4): 598-613.

Zhang, L., S. Volinia, et al. (2008). "Genomic and epigenetic alterations deregulate microRNA expression in human epithelial ovarian cancer." Proc Natl Acad Sci U S A **105**(19): 7004-7009.

Zhang, R. and B. Su (2009). "Small but influential: the role of microRNAs on gene regulatory network and 3'UTR evolution." J Genet Genomics **36**(1): 1-6.

Zhu, S., H. Wu, et al. (2008). "MicroRNA-21 targets tumor suppressor genes in invasion and metastasis." Cell Res **18**(3): 350-359.

10.DECLARATION

“I declare that I have completed this dissertation single-handedly without the unauthorized help of a second party and only with the assistance acknowledged therein. I have appropriately acknowledged and referenced all text passages that are derived literally from or based on the content of published or unpublished work of others, and all information that relates to verbal communications”

Emmanouil Maragkakis

Halle/Saale 20 May 2011

11.PERSONAL INFORMATION

Name: Emmanouil Maragkakis
Date of birth: 31 May 1983
Place of birth: Greece
Gender: Male
Residence: Greece
Address: B.S.R.C. Alexander Fleming, 34 Fleming Street, 16672, Athens, Greece
Nationality: Greek

Education

Period	October 2001 – September 2002
Organization	National Kapodistrian University of Athens
Department	Biology

Period	October 2002 – October 2007
Organization	National Kapodistrian University of Athens
Department	Physics
Specialization	Electronics - Computers - Telecommunications - Automation
Thesis	Computational simulation of signal transmission through biological neurons in the form of action potentials.

Period	October 2007 – Today
Organization	> Martin Luther Universität Halle Wittenberg, Germany > Biomedical Science Research Center “Alexander Fleming”, Greece
Title	PhD student
Funding	Scholar of Biomedical Science Research Center “Alexander Fleming”
Specialization	Bioinformatics

Existing academic degree

Physics diploma

Field of PhD studies

Computer Science

Spoken languages

English: Proficiency in English
French: Delf I

List of publications

The published work that has resulted in the 4-year period from September 2007 to July 2011 is listed in chronological order.

1. Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, Vergoulis T, Koziris N, Sellis T, Tsanakas P, Hatzigeorgiou AG. DIANA-microT web server: elucidating miRNA functions through target prediction. **Nucleic Acids Res**. 2009 Jul 1;37(Web Server issue):W273-6. Epub 2009 Apr 30.

2. Papadopoulos GL, Alexiou P, Maragkakis M, Reczko M, Hatzigeorgiou AG. DIANA-mirPath: Integrating human and mouse miRNAs in pathways. **Bioinformatics**. 2009 Aug 1;25(15):1991-3. Epub 2009 May 12.

3. Maragkakis M, Alexiou P, Papadopoulos GL, Reczko M, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, Simossis VA, Sethupathy P, Vergoulis T, Koziris N, Sellis T, Tsanakas P, Hatzigeorgiou AG. Accurate miRNA target prediction correlates with protein repression levels. **BMC Bioinformatics**. 2009 Sep 18;10:295.

4. Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG. Lost in translation: an assessment and perspective for computational miRNA target identification. **Bioinformatics**. 2009 Dec 1;25(23):3049-55. Epub 2009 Sep 29. Review.

5. Alexiou P, Maragkakis M, Papadopoulos GL, Simmosis VA, Zhang L, Hatzigeorgiou AG. The DIANA-mirExTra web server: from gene expression data to miRNA function. **PLoS One**. 2010 Feb 11;5(2):e9171.

6. Dalamagas, T., Farmakakis, T., Maragkakis, M., Hatzigeorgiou, A. FreePub: Collecting and Organizing Scientific Material Using Mindmaps. **3rd International Workshop on Semantic Web Applications and Tools for the Life Sciences**. 2010.

7. Grau, J., Arend, D., Grosse, I., Hatzigeorgiou, A. G., Keilwagen, J., Maragkakis, M., Weinholdt, C., Posch, S. Predicting miRNA targets utilizing an extended profile HMM. **25th German Conference on Bioinformatics 2010**. 2010

8. Iizasa H, Wulff BE, Alla NR, Maragkakis M, Megraw M, Hatzigeorgiou A, Iwakiri D, Takada K, Wiedmer A, Showe L, Lieberman P, Nishikura K. Editing of Epstein-Barr virus-encoded

BART6 miRNAs controls their dicer targeting and consequently affects viral latency. **J Biol Chem**. 2010 Oct 22;285(43):33358-70. Epub 2010 Aug 17.

9. Alexiou, P., Maragkakis, M., Hatzigeorgiou, A. G. Online resources for miRNA analysis. **Journal of Nucleic Acids Investigation**. 2011;*2*(1), 2-5. Review.

10. Maragkakis, M., Vergoulis, T., Alexiou, P., Reczko, M., Plomaritou, K., Gousis, M., Kourtis, K., Koziris, N., Dalamagas, T., Hatzigeorgiou, A. G. DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association. **Nucleic Acids Res**. 2011. [Epub ahead of print]

11. Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I., Hatzigeorgiou, A. G. microRNA targeting in coding regions: a computational and experimental study of functionality. (under submission)

Emmanouil Maragkakis

Halle/Saale 20 May 2011

Acknowledgements

Foremost, I wish to express my gratitude to my advisor Dr. Artemis Hatzigeorgiou for giving me the opportunity to work in her lab on an exciting and challenging topic, and for her continuous help in moving things forward, always in a positive and motivating manner.

I would equally like to acknowledge and express my sincere gratitude to my supervisor Prof. Dr. Ivo Grosse for making his advice and support available in a number of ways.

Also, I would like to thank all my present and past fellow lab members, especially Dr. Martin Reczko and Panagiotis Alexiou as well as Giorgio Papadopoulos, Dr. Victor Simossis and George Georgakilas for all their help and for our fruitful collaboration.

I would also like to thank Dr. Theodore Dalamagas and Thanasis Vergoulis for our very productive collaboration and our stimulating discussions.

I would like to thank all the members of Prof. Dr. Ivo Grosse lab and especially Dr. Jan Grau for our wonderful collaboration. Also I am grateful to Prof. Dr. Aris Moustakas for pointing me to the right direction in the beginning of my research career.

Most importantly, a great thanks to my parents, Penny and John, and to my brother, Gregory, for making things easy for me and supporting me through my entire life. Finally, I wish to say a big thanks to Dimitra for keeping me always happy and for being there whenever I need her.