University of Thessaly

Department of Electrical and Computer Engineering

Diploma Thesis

Papadopoulou Chrysa

Papadopoulou Iliana

Field of Study:

Bioinformatics

Title:

"Evaluation of microRNA target prediction with experimentally supported targets"

"Αξιολόγηση πρόβλεψης στόχων microRNA με πειραματικά υποστηριζόμενο σύνολο δειγμάτων"

Supervisor Professors: Dr. Hatzigeorgiou Arthemis

Dr. Potamianos Gerasimos

June 2018, Volos, Greece

Abstract

MicroRNAs (miRNAs) are endogenous non-coding RNAs that regulate gene expression at the posttranscriptional level by binding to the complementary sites in the 3' untranslated region (UTR) of the target mRNAs and repress its translation or initiate its degradation. The deregulations of genes, controlled by miRNAs, along with the modified miRNA expression are associated with many diseases such as cancer, cardiovascular, metabolic and neurodegenerative disorders. Hence, it is of the utmost importance to forecast accurately candidate miRNA targets which might participate in these diseases. However, finding a functional miRNA target constitutes an arduous task, due to the vast number of miRNAs and potential targets as well as the exorbitant experimental prediction design. Consequently, sophisticated computational methodologies for miRNA target identification comprise cornerstones in miRNA research.

There is a plethora of available tools for miRNA target prediction, which encompass a range of diverse computational approaches, from the modeling of physical interactions to the integration of machine learning. In this study, the origins, the biogenesis and the fundamental functionality of miRNAs are presented along with the compendium of in vivo experimental techniques that validate predicted interactions. Furthermore, thirteen (13) currently available and frequently used target prediction algorithms are examined and their characteristics are analyzed. These characteristics include training data, test data, machine learning models, binding regions and features.

Subsequently, the performance of the tools is evaluated by conducting an in-depth analysis of miRNA-gene and miRNA-site interactions. The algorithms are tested on two (2) sets of experimentally validated true miRNA-target pairs, obtained from DIANA-TarBase repository in order the shared interactions to be revealed. It is corroborated that TargetScan outperforms the other miRNA target prediction tools in terms of gene and site level. Finally, insights on future directions are discussed.

Keywords: miRNA, target prediction, binding site, experimental techniques, machine learning models, features, interaction, performance.

Περίληψη

Τα microRNAs (miRNAs) είναι ενδογενή μη κωδικά RNAs που ρυθμίζουν την γονιδιακή έκφραση στο μετα-μεταγραφικό επίπεδο, προσδένονταςστις συμπληρωματικές θέσεις της 3 'αμετάφραστης περιοχής (UTR) των mRNA στόχων και καταστέλλουν τη μετάφρασή ή ξεκινούν την αποικοδόμηση. Η αππορύθμιση των γονιδίων που ελέγχονται από miRNAs μαζί με την τροποποιημένη έκφραση του miRNA, σχετίζονται με πολλές ασθένειες όπως ο καρκίνος, οι καρδιαγγειακές, οι μεταβολικές και οι νευροεκφυλιστικές διαταραχές. Ως εκ τούτου, είναι εξαιρετικά σημαντικό να προβλεφθούν με ακρίβεια υποψήφιοι στόχοι miRNA που θα μπορούσαν να συμμετέχουν σε αυτές τις ασθένειες. Ωστόσο, η εύρεση ενός λειτουργικού στόχου miRNA αποτελεί ένα δύσκολο έργο λόγω του μεγάλου αριθμού miRNAs και των πιθανών στόχων καθώς και του υπερβολικά κοστοβόρου σχεδιασμού πειραματικής πρόβλεψης. Συνεπώς, οι εξελιγμένες υπολογιστικές μεθοδολογίες για τον προσδιορισμό των στόχων miRNA αποτελούν ακρογωνιαίο λίθο στην έρευνα των miRNA.

Υπάρχει μια πληθώρα διαθέσιμων εργαλείων για την πρόβλεψη στόχων miRNA, τα οποία περιλαμβάνουν μια σειρά διαφορετικών υπολογιστικών προσεγγίσεων, από τη μοντελοποίηση των φυσικών αλληλεπιδράσεων έως την ενσωμάτωση της μηχανικής μάθησης. Σε αυτή τη μελέτη παρουσιάζονται η προέλευση, η βιογένεση και η θεμελιώδης λειτουργικότητα των miRNAs μαζί με τη σύνοψη των in vivo πειραματικών τεχνικών που επικυρώνουν τις προβλεπόμενες αλληλεπιδράσεις. Επιπλέον, εξετάζονται δεκατρείς (13) διαθέσιμοι και συχνά χρησιμοποιούμενοι αλγόριθμοι πρόβλεψης στόχων και αναλύονται τα χαρακτηριστικά τους. Αυτά τα χαρακτηριστικά περιλαμβάνουν δεδομένα εκπαίδευσης, δεδομένα δοκιμών, μοντέλα μηχανικής μάθησης, περιοχές πρόσδεσης και χαρακτηριστικά.

Στη συνέχεια, η απόδοση των εργαλείων αξιολογείται διεξάγοντας μία εις βάθος ανάλυση των αλληλεπιδράσεων μεταξύ miRNAsκαι γονιδίων και μεταξύ miRNAs και περιοχής πρόσδεσης. Οι αλγόριθμοι δοκιμάζονται σε δύο (2) σύνολα πειραματικά επικυρωμένων αληθινών ζευγών miRNA-στόχου, που λαμβάνονται από τη βάση δεδομένων DIANA-TarBase με σκοπό να αποκαλυφθούν οι κοινές αλληλεπιδράσεις. Επιβεβαιώνεται ότι ο TargetScan ξεπερνά τα υπόλοιπα εργαλεία πρόβλεψης στόχων miRNA σε επίπεδο γονιδίων και περιοχής πρόσδεσης. Τέλος, συζητούνται ιδέες για τις μελλοντικές κατευθύνσεις.

Λέξεις-κλειδιά: miRNA, πρόβλεψη στόχου, περιοχή πρόσδεσης, πειραματικές τεχνικές, μοντέλα μηχανικής μάθησης, χαρακτηριστικά, αλληλεπίδραση, απόδοση.

Acknowledgement

We would like to express our sincere gratitude to

- Dr. Artemis Hatzigeorgiou, Professor at the University of Thessaly, Department of Electrical and Computer Engineering, for allowing us to undertake this work and providing us with invertible suggestion throughout our research.
- Our supervisor, Associate Professor Dr. Gerasimos Potamianos, Department of Electrical and Computer Engineering, for his continuous guidance advice effort.
- DIANA-Lab and especially our supervisor, Dimitra Karagkouni, PhD student at University of Thessaly, for providing us the logistic support and her valuable suggestion to carry out our research successfully.
- Dr. Elias Xoustis, Emeritus Professor at the University of Thessaly, Department of Electrical and Computer Engineering, who believed in us, consulted and encouraged us to proceed to our graduate studies.
- Dr. Nikolaos Bellas, Associate Professor at the University of Thessaly, Department of Electrical and Computer Engineering for his ultimate support throughout our university years and his profound motivation to achieve our goals.
- Dr. Chistos Sotiriou, Associate Professor of the same Department, for offering us the summer internship opportunities in his group and leading us working on diverse exciting projects.
- Dr. Lefteris Tsoukalas, President at the University of Thessaly, Department of Electrical and Computer Engineering for his confidence and support.
- Dr. Panagiota Tsompanopoulou, our professor advisor, for her precious advice during our university life.
- And last but not least, our parents for providing us with unfailing support and continuous encouragement not only throughout the process of researching and writing this thesis, but throughout our whole life. This accomplishment would not have been possible without them.

We thank them all.

Contents

List of Tables

List of Figures

Chapter I

1. Introduction

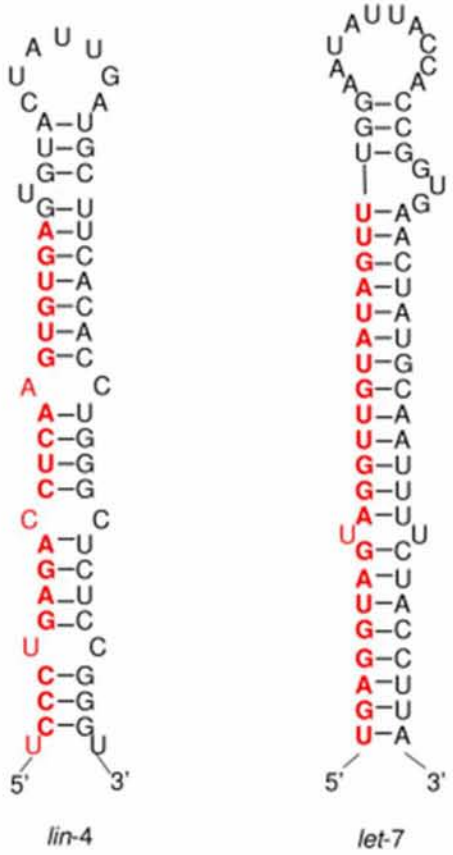
The central dogma of molecular biology explains the flow of genetic information within a biological system. In particular, the process, during which the coded genetic information into DNA is transcribed into messenger RNA (mRNA), which is then translated into proteins. In general, the classic view of central dogma of biology reflects how molecular biological data are organized within databases (e.g. by molecule type such as genomic DNA, mRNA and protein). However, many challenges to this dogma emerge due to the in-depth study of the genome in recent years. For instance, some segments of the DNA, transcribed into the mRNA precursor (pre-mRNA), even though they do not necessarily encode proteins, are known to regulate the expression of various types of functional RNAs. As a result, non-coding RNAs are generated and are abundant in the human genome (nearly 95%).

MicroRNAs (miRNAs) are a class of single-stranded, evolutionary conserved, endogenous non-coding RNA molecules ~22 nucleotides (nt) in length with 5'-phosphate and 3'-hydroxyl ends. They play a significant role in the regulation of gene expression in many eukaryotes as they bind to the 3' untranslated regions (UTRs) of target mRNAs, inhibiting their translation [1].

1.1 miRNA Discovery

Victor Ambros' laboratory first discovered the lin-4 RNA in an attempt to identify the timing of stem-cell division and differentiation in Caenorhabditis elegans in 1993 [2]. Simultaneously, Gary Ruvkun's laboratory identified lin-4's first microRNA protein-coding target gene, lin-14 [3]. From the above experiments, it was found that there are multiple sites in the 3'UTR (untranslated region) of lin-14 that are complementary to lin-4 and cause inhibition of lin-14 protein expression [2, 3]. Almost a decade later, a second miRNA, the let-7, was identified in various species from C. elegans to humans [4, 5]. Indeed, let-7 is considered the first known human miRNA that regulates the late larval development by inhibiting lin-41 expression in some tissues [5, 6]. The homology search, with the use of let-7 sequence, revealed that let-7 and its family members are highly conserved across many organisms during the larval stage. The same applies in the case of lin-41. The misregulation of let-7 leads to the increase of cell-based diseases such as cancer.

Figure 1. The predicted stem loops for lin-4 (left) and let-7 (right) in C. elegans [2, 5]. The sequences of mature miRNAs are shown in red [7].



The novel detection of miRNAs lin-4 and let-7 contributed to the subsequent large numbers of miRNAs that were discovered by low-throughput and high-throughput experimental methods along with computational approaches in multiple biological procedures. In the recent years, it is evident that miRNAs are a profuse class of gene regulators, appearing daily in various databases.

miRBase is the major online repository for compiling published miRNA sequences and associated annotations [8]. Each entry represents a predicted hairpin portion of a miRNA transcript (termed mir in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). According to the latest release of miRBase (miRBase 21), in total, 4,196 new hairpin sequences and 5,441 new mature products have been added compared to the previous version. Consequently, 28,645 entries, which represent hairpin precursor miRNAs, expressing 35,828 mature miRNA products, are now present in 223 species. Indeed, until now, 2,585 human and 1,899 mouse miRNAs have already been identified and this number is rapidly growing.

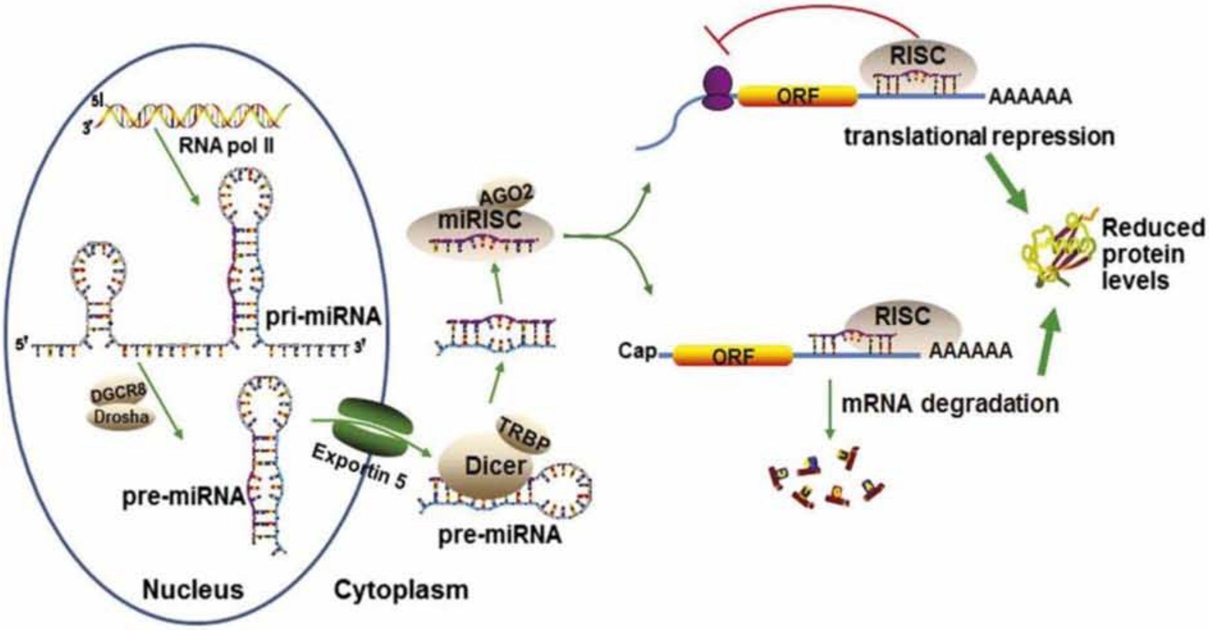
1.2 Biogenesis of miRNA: Gene Transcription and Maturation

MicroRNAs are derived from the double-stranded region of a 60-70 nt RNA hairpin precursor [9]. MiRNA precursors are commonly found within intergenic regions and introns of protein-coding genes [10]. The identification of intergenic miRNA and protein-coding intronic miRNA revealed that miRNAs are transcribed by RNA polymerase II (Pol II), which is responsible for the transcription of DNA into mRNA and polymerase III (Pol III) which contributes to the regulation of cell growth and the cell cycle, generating precursors that undergo a series of cleavage events to form mature microRNA.

The conventional biogenesis pathway consists of two cleavage events, one nuclear and one cytoplasmic and depends on both DiGeorge syndrome critical region 8 (Drosha-DGCR8) and Dicer-TRBP (transactivation-response RNA-binding protein) complexes. In animals, the nuclear RNase III Drosha and DGCR8/Pasha cleave long primary transcripts 500–3,000 nt in length (pri-miRNAs) to release double-stranded hairpin-shaped pre-miRNAs, approximately 60-100 nt in length [11, 12, 13, 14, 15]. Pre-miRNAs are then exported from the nucleus by Exportin-5 and RAN GTPase [16] and are cut by the cytoplasmic RNase III Dicer to produce double-stranded miRNA duplexes. Therefore, a mature miRNA sequence, approximately 20 nt in length and its short-lived complementary sequence, which is denoted miR, are generated. In plants, this two-step processing of pri-miRNA into mature miRNA occurs completely in the nucleus and is carried out by a single RNase III enzyme, DCL1 (Dicer-like 1).

Once incorporated into the Argonaute 2 protein, containing miRNA-induced silencing complexes (miRISCs) in the cytoplasm, single-stranded miRNAs function as post-transcriptional regulators. The thermodynamic stability of the 5'end of the miRNA duplex determines which strand will be incorporated into miRISCs. Indeed, the guide strand with relatively unstable 5'end is selected and is integrated into RISC, while the passenger strand is discarded. Then, the RNA-induced silencing complex (RISC) is formed. Even though the composition of RISC remains to be meticulously examined, its most important protein is the AGO2. RISC guides the miRNA to the target mRNA based on a 2-8 nucleotide sequence, namely the seed, at the 3'UTR of the target mRNA. Subsequently, the miRNA impedes the translation of the mRNA by two possible pathways. When the guide strand binds with perfect complementarity at the 3'UTR of the target mRNA, RISC cleaves the target mRNA [17, 18]. In the case of imperfect complementarity to target mRNAs, reduced translation and/or stability of target mRNAs is provoked. In any case, inhibition of gene expression is carried out. **Figure 2** illustrates the canonical pathway of miRNA biogenesis and its mechanisms, controlled by post-transcriptional gene regulation.

Figure 2. MicroRNA biogenesis and function in animal cells. miRNAs are transcribed as long primary transcripts (pri-miRNAs) in the nucleus. The pri-miRNAs are processed by the RNase III-type Drosha, yielding pre-miRNAs of ~70 nucleotides (nt). The pre-miRNAs are exported to the cytoplasm by exportin-5 and are further cleaved into ~21 to 22 nucleotide miRNA duplex by another RNase III enzyme Dicer. Subsequently, the mature single strand is incorporated into RISC. Finally, once the miRNA bounds to the binding site within the 3'UTR of the targeted transcript, mRNA decay and translation repression are exerted, eliminating protein levels [19].



1.3 miRNA-mRNA interactions

Even though the rules of the binding of miRNAs in the target mRNAs are not yet fully discovered, existing research proposes a few specific features about miRNA base pairing. miRNAs are predominantly found in 3' untranslated regions (3'UTRs) of target genes but in some cases they exist in coding regions, 5'UTRs and open reading frames (ORFs) [20]. Indeed, miRNA target sites in plants are within ORFs of target genes and nearly perfect complementarity is obliged between miRNAs and their target transcripts [21]. In animals, target sites are not located within 3'UTR but they rather tend to concentrate at both ends of 3'UTR [22]. Moreover, the 3' UTR of the target mRNA can include multiple sites for the same miRNA accelerating the possibility of binding [23].

The perfect pairing is more frequent in a region called the "seed", often defined as the 2nd-8th nt from the 5' end of the miRNA and characterized by a strict or almost strict Watson-Crick pairing between miRNA and its target site. It is crucial to mention that rigid complementarity between the miRNA and the target site is a rare phenomenon. MiRNAs typically have imperfect Watson-Crick base pairing with the corresponding miRNA response elements (MREs) which are usually present within the 3′-untranslated regions (3′UTRs) of target mRNAs [24]. Nevertheless, perfect pairing is not necessary as far as miRNA-target interactions are concerned due to the fact that imperfect pairing in the seed region can be compensated by the complimentary sites at the 3' end of the miRNA.

1.4 SiRNAs

Small interfering RNAs (siRNAs) are another class of small noncoding RNAs that regulate gene expression in a similar manner as miRNAs. However, their main difference lies in their origins [1]. SiRNAs are generated from long, double-stranded RNAs or long hairpins, usually of exogenous origin and in most cases, target sequences at the same locus or another location in the genome, contributing to gene silencing [25]. This phenomenon is called RNAi. On the other hand, miRNAs are endogenous, as they are encoded within the genome and are derived from endogenous short hairpin precursors, targeting sequences at other loci.

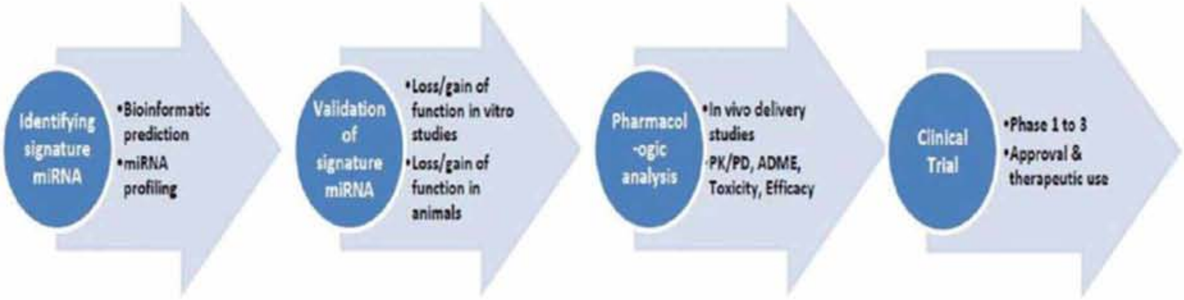
1.5 Clinical applications of miRNAs

Advancements in the miRNA field are increasing rapidly. One of the first clinical applications of miRNAs constitutes their embryonic ability as biomarkers for diagnosis, subtyping and estimation of disease progression [26, 27]. In addition, miRNAs are utilized as markers of drug response [28, 29]. However, their direct use for the development of therapeutic strategies is the most breathtaking one. Recently, miRNA research has been accelerated by technological innovation in RNA-based therapies. The misregulation of several miRNAs is chained with the enhancement of certain diseases in humans and other organisms. It is substantiated that the rejuvenation of misregulated miRNAs to their normal levels, can reduce or even eradicate diseases including tumors in animal models. Due to the existence of miRNAs as naturally occurring molecules, their application as therapeutic agents presents certain tangible benefits. Worldwide, the theory of "miRNA replacement therapy" is authenticated and includes the insertion of synthetic miRNAs (S-miRNA) or miRNA mimetics into diseased tissues in an effort to rehabilitate normal proliferation, apoptosis, cell cycle and other cellular functions that have been affected by the misregulation of one or more miRNAs.

On the other hand, miRNA inhibitors such as antagomirs, LNA-modified oligonucleotides and miRNA erasers (sponges) have been used with the aim of enhancing the endogenous levels of therapeutic proteins. Consequently, theoretically, impediment of a specific miRNA, connected to a given disease, can remove the block of expression of a therapeutic protein. However, the administration of a miRNA mimetic can enlarge the endogenous miRNA population, therefore stifling a deleterious gene. In many cases, the reactivation or inhibition of these miRNA-regulated pathways generate significant therapeutic responses.

Innovative teams of specialized pharmaceutical companies have initiated studies on devising feasible therapeutic candidates with miRNA inhibitors and miRNA mimetics in diverse fields such as cancer, cardiovascular diseases, neurodegenerative diseases as well as metabolic disorders. The process of building miRNA therapeutics is similar to drug discovery and development. In the discovery and development of miRNA therapeutics, the steps shown in **Figure 3** are engaged.

Figure 3. Process of microRNA discovery and development [30].


Identifying signature miRNA
•Bioinformatic prediction
•miRNA profiling
Validation of signature miRNA
•Loss/gain of function in vitro studies
•Loss/gain of function in animals
Pharmacol -ogic analysis
•In vivo delivery studies
·PK/PD, ADME, Toxicity, Efficacy
Clinical Trial
•Phase 1 to 3
•Approval & therapeutic use

- Identification of signature miRNA (performed by miRNA profiling in disease)
- Validation of signature miRNA (loss/gain of function studies in vitro and in animal models)
- Pharmacological analysis (in vivo miRNA delivery studies, pharmacokinetics/pharmacodynamics, (absorption, distribution, metabolism, excretion and toxicity studies)
- Clinical trials (studies on the evaluation of efficacy and safety)

Hence, miRNA expression analysis has a crucial diagnostic value and normalizing miRNA levels constitutes one of the most promising therapeutic approaches in creating a new generation of drugs. [31] **Table 1, Figure 4** summaries notable therapeutics miRNA which traverse their development phase and embody a bright future in the combat again diseases.

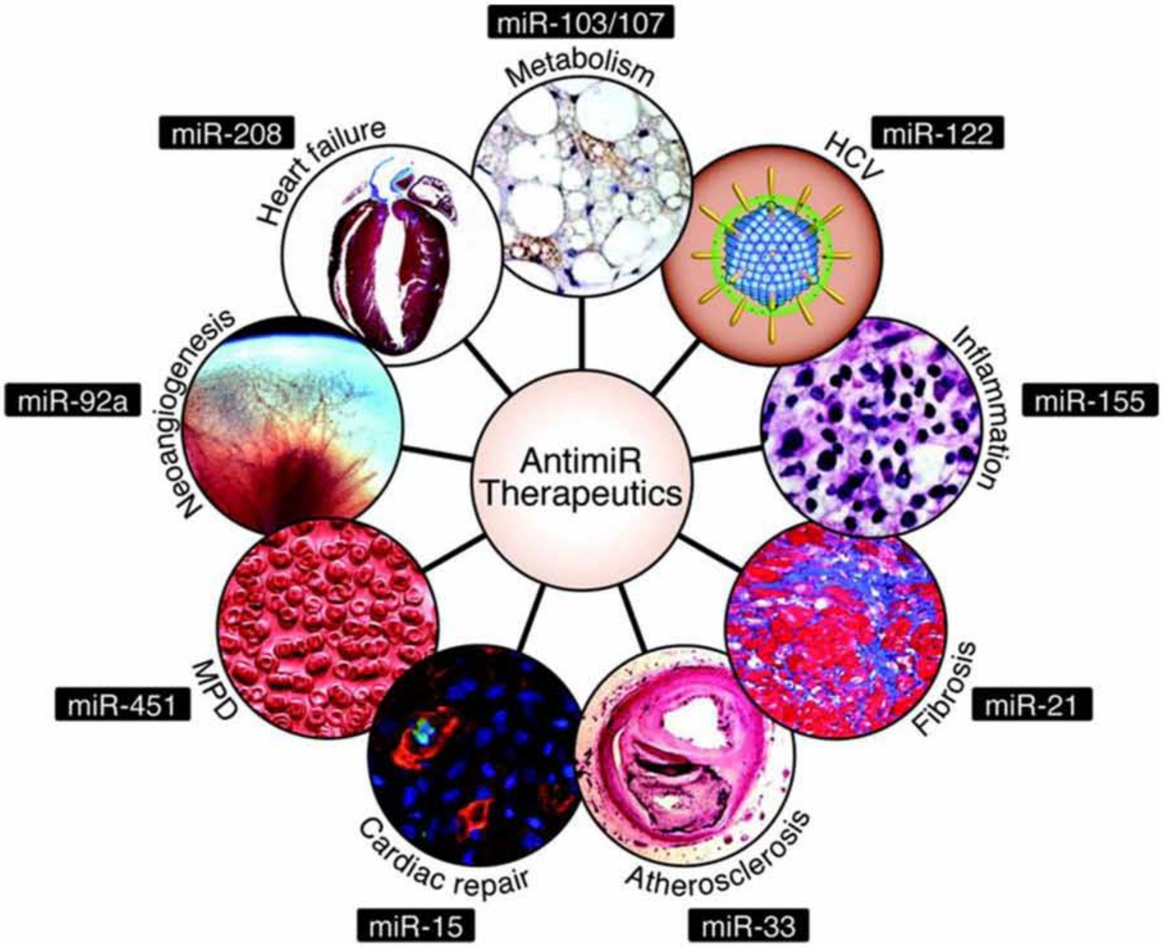
Table 1. Significant therapeutics miRNA that are in the Development Phase, their indication and their biopharmaceutical Company [32].

Therapeutic miRNAs	**Indication**	**Biopharmaceutical Company**	**Remarks**
Miravirsen	hepatitis C virus (HCV) infection	Santaris Pharma	phase IIa clinical trial

MRX34	treatment of a variety of cancers such as colon cancer, non-small-cell lung cancer (NSCLC), hepatocellular carcinoma, cervical cancer, ovarian cancer, etc.	Mirna Therapeutics	phase 1 clinical trial halted because of immune responses
RG-101	treatment of HCV	Regulus Therapeutics	an owned GalNAc-conjugated anti-miR
RG-012	treatment of Alport syndrome	Regulus Therapeutics	in the pipeline to initiate clinical trial phase II
MGN-1374	treatment of post-myocardial infarction remodeling	miRagen Therapeutics	targets miR-15 and miR-195; it is in the preclinical stage
MGN-2677	treatment of vascular disease	miRagen Therapeutics	targets miR-143/145; it is in the pipeline
MGN-4220	treatment of cardiac fibrosis	miRagen Therapeutics	targets miR-29; it is in the pipeline
MGN-4893	treatment of disorders like abnormal red blood cell production such as polycythemia vera	miRagen Therapeutics	targets miR-451; it is in the pipeline
MGN-5804	treatment of cardiometabolic disease	miRagen Therapeutics	targets miR-378; it is in the pipeline
MGN-6114	treatment of peripheral arterial disease	miRagen Therapeutics	targets miR-92; it is in the pipeline

MGN-9103	treatment of chronic heart failure	miRagen Therapeutics	targets miR-208; it is in the pipeline

Figure 4. Specific miRNAs that are currently being pursued as clinical candidates [32a].



1.6 miRNA Target Features

Several in silico tools are available for identifying putative miRNA targets. The basic features used by these tools can be gathered and divided into three categories: duplex features, local context features and global context features [33]. On the one hand, duplex features include seed match, 3′ contribution, seed pairing stability (SPS) [33], heteroduplex free energy and p-value [34]. The aforementioned parameters assess the hybridization of the miRNA to its target gene. Seed match evaluates the number of nts that can bind to the mRNA target in the seed region. The 3′ contribution calculates the possibility of binding at the 3′ position of the miRNA [35]. The types of nts which compose the seed region are encompassed in seed pairing stability (SPS) [36]. In addition, heteroduplex free energy evaluates whether the minimum free energy between the miRNA and its target is sufficient to establish hybridization, while the p-value measures whether the probability of a selected interaction has been forecasted by chance.

On the other hand, mRNA sequence properties that straightly influence target identification such as site accessibility (SA), presence of flanking AU and target-site abundance, are included in the local context features. Site accessibility (SA) evaluates the capacity of the mRNA to unfold into a potential secondary structure in the region containing the miRNA cognate sequence, which is known as the miRNA recognition element (MRE) [37]. The flanking AU refers to the number of A and U nts flanking the MRE region. High concentrations of flanking A and U nts enhance miRNA regulation [38]. What is more, target-site abundance is a measure of the number of target sites, which occur in a 3′ UTR [36].

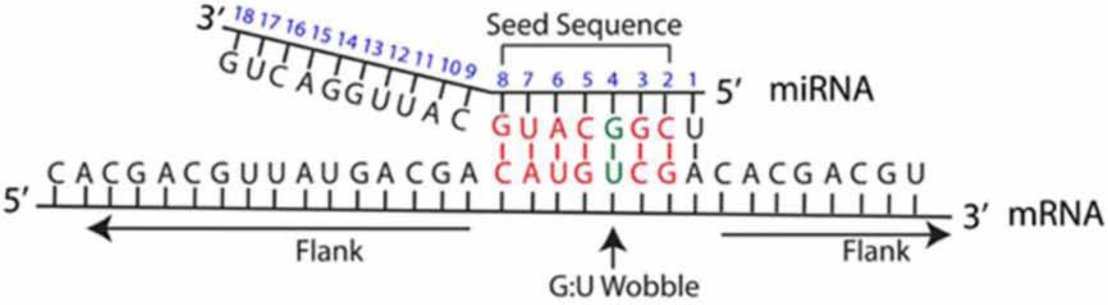
Global context features incorporate mRNA sequence properties with indirect influence on target recognition, namely transcript length, 3′UTR length, transcriptome abundance, pairing position at the 3′UTR and sequence conservation. Sequence length evaluates the total length of the string analyzed, since the chances of false prediction burgeon with target length [34]. Moreover, the 3′UTR length finds the length of the 3′UTR of the potential miRNA targets, since larger 3′UTRs are regulated more stringently than shorter ones [39]. Transcriptome abundance reveals the number of MREs of a miRNA within the transcriptome. Pairing position assesses the position of the MRE within the 3′UTR, due to the fact that MREs near the ends of the 3′UTR contain stronger regulatory potential [38]. Finally, sequence conservation evaluates the extent of conservation of the MREs among species. Together, all these binding metrics decisively regulate the determination of potential miRNA-target pairs.

Among the aforementioned features, the most commonly used in target prediction tools include seed match, sequence conservation, free energy and site accessibility. Thus, they will be described in the following sections with details.

a) miRNA:mRNA pairing: Seed match

miRNAs regulate the gene expression by binding to the corresponding mRNA. The seed sequence is critical for the binding of the miRNA to the mRNA. The seed sequence or seed region is a conserved sequence which is located at positions 2-8 from the 5′ end of the miRNA toward the 3′ end [40] and has perfect Watson–Crick complementarity to the 5′ part of miRNAs. A Watson-Crick (WC) match between a miRNA and mRNA occurs when adenosine (A) pairs with uracil (U) and guanine (G) pairs with cytosine (C) (**Figure 5**). A perfect seed match between the miRNA and the mRNA target has no gaps in alignment within the WC matching.

Figure 5. miRNA:mRNA pairing. Graphical analysis of a miRNA interaction with its mRNA target. MiRNA position number is depicted with blue. The seed match refers to nucleotides in miRNA position number 2–8. Flank refers to the mRNA sequence on either side of the region, corresponding to the miRNA seed sequence. WC matches in the seed region are demonstrated in red. An example of G-U wobble in the seed region is shown in green [41].



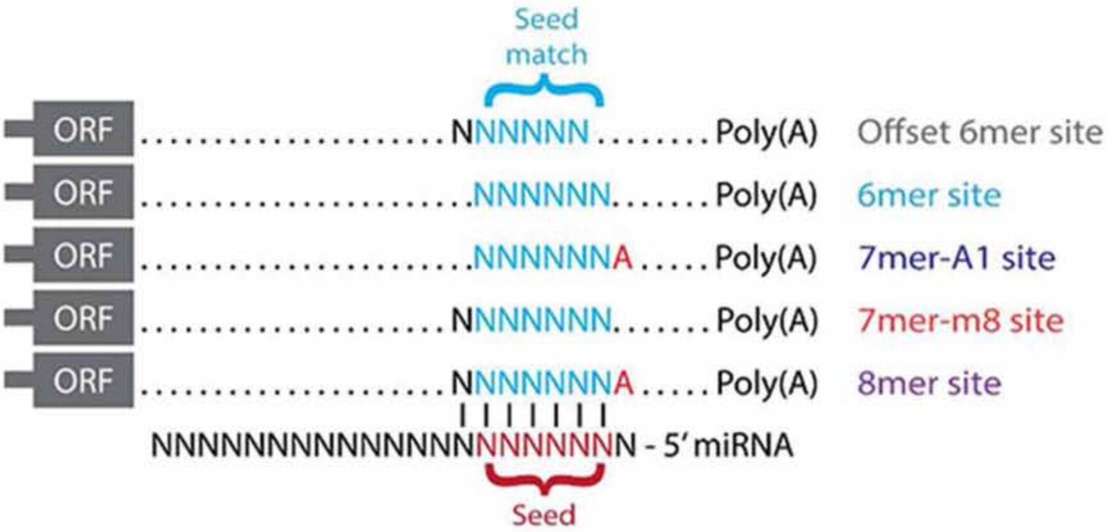
Seed matching can be categorized in four different types depending on the combination of the nucleotides in position 1 and 8 (**Figure 6**):

1. 6-mer: Perfect seed matching (WC) for six nucleotides between the miRNA seed sequence and the mRNA. Often, miRNAs within this site type downregulate target mRNAs.
2. 7mer-m8: Perfect seed match (WC) between 2-8 nucleotides of miRNA seed sequence. It is the most abundant type of sites, taking into account only those sites targeted by highly conserved miRNAs.
3. 7mer-A1: Perfect seed match (WC) between 2-7 nucleotides of miRNA seed sequence in addition to an adenine (A) across from the miRNA nucleotide 1. The presence of A creates conservation and contributes to the increased degree of gene silencing.
4. 8-mer: Perfect seed match (WC) between nucleotides 2-8 of miRNA seed sequence in addition to an A across from the miRNA nucleotide 1. The presence of adenine (A) creates conservation.
5. Offset 6mer: Shifted 6mer at positions 3-8 of miRNA seed sequence

Additional site types were discovered later:

1. 3’ supplementary: Perfect seed match (WC) between 2-7 and 13-16 nucleotides of miRNA seed sequence.
2. 3’ compensatory: Mismatch or G: U wobble in the seed region of the miRNA and perfect seed match (WC) between 13-16 nucleotides of miRNA seed sequence.

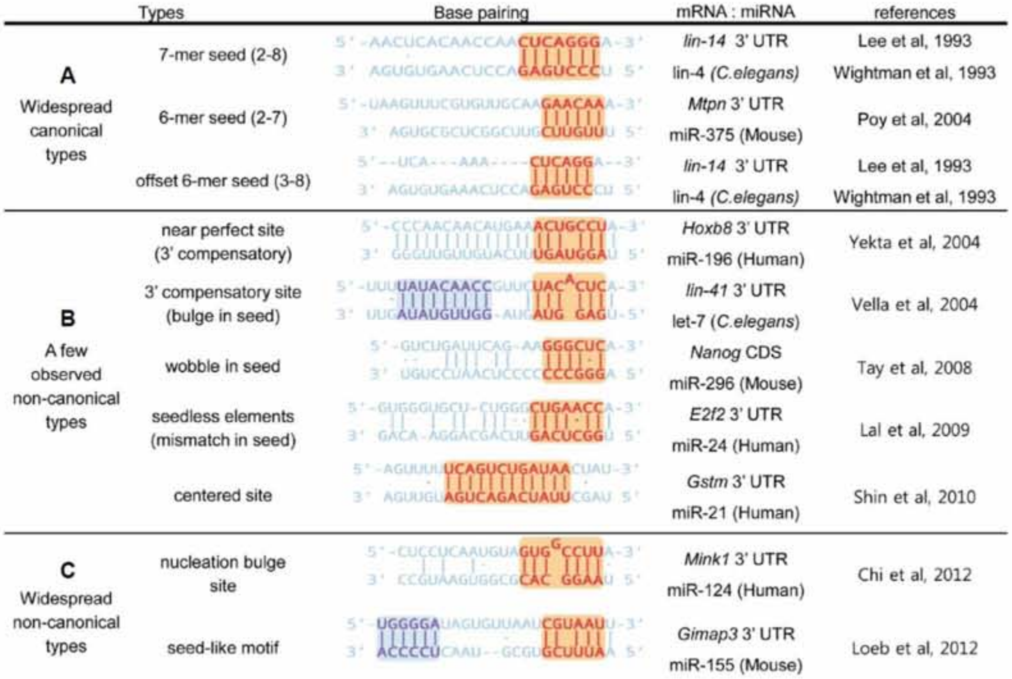
Figure 6. Canonical site types that match the seed region of every miRNA including all k-mer: 8mer, 7mer-m8, 7mer-A1, 6mer, and offset 6mer [42].


Seed match
ORF ....NNNNNN........Poly(A) Offset 6mer site
ORF ....NNNNNN.......Poly(A) 6mer site
ORF ....NNNNNNA.....Poly(A) 7mer-A1 site
ORF ....NNNNNNN.......Poly(A) 7mer-m8 site
ORF ....NNNNNNNA.....Poly(A) 8mer site
NNNNNNNNNNNNNNNNNNNNNN - 5′ miRNA
Seed

Wobble base pair constitutes a non-Watson-Crick base pair model. In RNA molecules, four main types of wobble base pairings have been discovered namely guanine-uracil (G-U), hypoxanthine-uracil (I-U), hypoxanthine-adenine (I-A) and hypoxanthine-cytosine (I-C). Due to its unique physical, dynamic and ligand binding capacity and acceptable thermodynamic stability, G-U wobble pair plays a fundamental role in various biological processes [43]. Recent studies have demonstrated that most of the target prediction algorithms, which do not take into account non-Watson-Crick seed pairing, fail to achieve good prediction accuracies.

A summary of all site types, along with their Base pairing, their mRNA: miRNA interactions and their referenced datasets are shown in **Figure 7**.

Figure 7. Canonical and non-canonical miRNA target sites. Representative examples of each type are indicated as widespread canonical (A), a few observed non-canonical (B) and widespread non-canonical types (C). Key binding regions are highlighted in red and subtly contributing regions are in purple. Solid lines indicate Watson-Crick base pairing and dots indicate G:U wobble pairs [43a].



	Types	Base pairing	mRNA : miRNA	references
A Widespread canonical types	7-mer seed (2-8)	5'-AACUCACAACCAACUCAGGGA-3' 3' AGUGUGAACUCCAGAGUCCCU 5'	*lin-14* 3' UTR lin-4 *(C.elegans)*	Lee et al, 1993 Wightman et al, 1993
	6-mer seed (2-7)	5'-UAAGUUUCGUGUUGCAAGAACAAA-3' 3' AGUGCGCUCGGCUUGCUUGUUU 5'	*Mtpn* 3' UTR miR-375 (Mouse)	Poy et al, 2004
	offset 6-mer seed (3-8)	5'--UCA---AAA----CUCAGGA--3' 3' AGUGUGAAACUCCAGAGUCCCU 5'	*lin-14* 3' UTR lin-4 *(C.elegans)*	Lee et al, 1993 Wightman et al, 1993
B A few observed non-canonical types	near perfect site (3' compensatory)	5'-CCCAACAACAUGAAACUGCCUA-3' 3' GGGUUGUUGUACUUUGAUGGAU 5'	*Hoxb8* 3' UTR miR-196 (Human)	Yekta et al, 2004
	3' compensatory site (bulge in seed)	5'-UUUUAUACAACCGUUCUAC^{A}CUCA-3' 3' UUCAUAUGUUGG-AUCAUG GAGU-5'	*lin-41* 3' UTR let-7 *(C.elegans)*	Vella et al, 2004
	wobble in seed	5'-GUCUGAUUCAG-AAGGGCUCA-3' 3' UGUCCUAACUCCCCCCCGGGA 5'	*Nanog* CDS miR-296 (Mouse)	Tay et al, 2008
	seedless elements (mismatch in seed)	5'-GUGGGUGCU-CUGGGCUGAACCA-3' 3' GACA-AGGACGACUUGACUCGGU 5'	*E2f2* 3' UTR miR-24 (Human)	Lal et al, 2009
	centered site	5'-AGUUUUUCAGUCUGAUAACUAU-3' 3' AGUUGUAGUCAGACUAUUCGAU 5'	*Gstm* 3' UTR miR-21 (Human)	Shin et al, 2010
C Widespread non-canonical types	nucleation bulge site	5'-CUCCUCAAUGUAGUG^{G}CCUUA-3' 3' CCGUAAGUGGCGCAC GGAAU 5'	*Mink1* 3' UTR miR-124 (Human)	Chi et al, 2012
	seed-like motif	5'-UGGGGAUAGUGUUAAUCGUAAUU-3' 3' ACCCCUCAAU--GCGUGCUUUAA 5'	*Gimap3* 3' UTR miR-155 (Mouse)	Loeb et al, 2012

b) Conservation

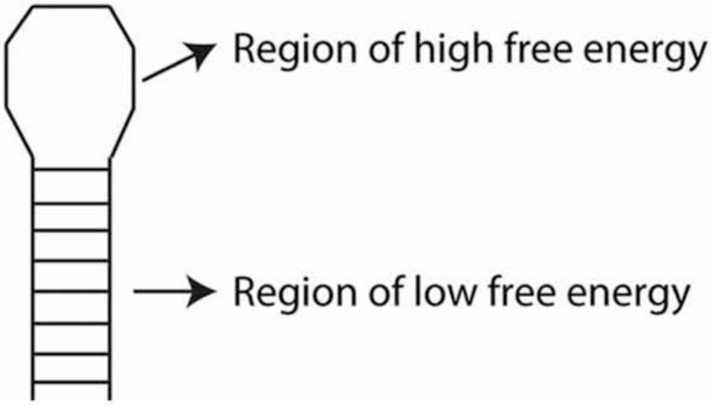
Conservation can be described as the maintenance of a sequence across species. Conservation analysis focuses on regions in the 3′ UTR, the 5′ UTR, the miRNA, or any combination of the three. In general, miRNA seed region is by far more conserved than the non-seed region [40]. In a small proportion of miRNA:mRNA target interactions, the 3′ end of the miRNA consists of conserved pairing, compensating for seed mismatches and these sites are called 3′ compensatory sites [44]. The conservation analysis is performed with the aid of phylogenetic and evolutionary distance calculations. This analysis aims to prove that a predicted miRNA target is functional due to its selection by positive natural selection. As a result, a higher degree of conservation arguably reflects a more reliable prediction. In addition, the genomic regions flanking the miRNA gene and miRNA target genes present to be the most compelling for conservation analysis. For instance, many applications of conservation analysis constitute in the promoter regions of miRNAs and their target genes [45] and in the co-localization of independently transcribed miRNAs and flanking protein coding genes [46].

In site conservation, the conservation of the binding site is counted by the number of species with the same sequence and/or by the phylogenetic distance between the species sharing the same sequence. For instance, in TargetScan a highly conserved miRNA or a conserved miRNA in PACCMIT and PACCMIT-CDS shares the same seed sequence (positions 2–8) in different species.

c) Free energy

Free energy (or Gibbs free energy) measures the stability of a biological system. If the binding of a miRNA to a candidate target mRNA is predicted to be stable, it is more prone to be a true target of the miRNA. Thus, the lower the free energy, the greater the RNA:RNA binding, increasing the probability that this interaction will actually occur. Due to the adversity in calculating free energy directly, the alteration in free energy during a reaction is considered (ΔG). Systems with increased stability are derived from reactions with a negative ΔG, which units constitute the kcal/mol and contain less energy available to react in the future. By predicting how the miRNA and its candidate target hybridize, regions of high and low free energy can be found (**Figure 8**) and the overall ΔG indicates how strongly bound these regions are [47]. To measure this energy, most miRNA:mRNA prediction tools utilize the Vienna RNA package.

Figure 8. Schematic overview of free energy (ΔG) analysis of predicted RNA hybridization structure. A hairpin loop is shown with the loop corresponding to a region of high free energy (a positive ΔG) and the stem corresponding to a region of low free energy (a negative ΔG) [47].



d) Site accessibility

Site accessibility counts the ease with which a miRNA can locate and hybridize with an mRNA target. After transcription, mRNA assumes a secondary structure [48] which can interfere with a miRNA's ability to bind to a target site. MiRNA:mRNA hybridization involves a two-step process in which a miRNA binds first to a short accessible region of the mRNA. Additionally, the mRNA secondary structure unfolds as the miRNA completes binding to a target [49]. Therefore, the evaluation of the predicted amount of energy required to render a site accessible, aids in forecasting whether a mRNA is the target of a miRNA. [41, 50, 51]

1.7 Experimental methods for the identification of miRNA: mRNA interactions

MiRNA target prediction algorithms are a cornerstone of miRNA-related research. Nevertheless, results derived even from state of the art implementations should be corroborated by molecular evidence, in order a putative miRNA to be considered valid. Experimental data is critical not only in the identification of a specific interaction, but also in the investigation of features that characterize miRNA–mRNA interactions as well as in the assessment of the accuracy of the proposed computational approaches. As a result, it is integral to introduce the experimental methodologies, which are used to detect novel miRNA-target interactions and validate predicted interactions.

Experimental techniques can be divided in two classes, depending on the type of supporting information provided: direct or indirect. In addition, the experimental data can also be classified, depending on the resultant size of dataset: individual studies or high throughput. Low-yield techniques include reporter gene-assays, qPCR (quantitative polymerase chain reaction), western blotting and ELISA (enzyme-linked immunosorbent assay). In particular, reporter gene-assays concentrate on the recognition and evaluation of specific miRNA-binding site, while qPCR, western blotting and ELISA indirectly identify events such as diminution of mRNA or protein concentration leading to mRNA degradation or translation suppression respectively.

Direct validation of miRNA target genes is based on the attachment of a reporter construct [e.g. Luciferase or Green Fluorescent Protein (GFP)] to the genes of interest and the measurement of the expression of the reporter gene before and after the insertion of the miRNA to the cell. Even though such methods can provide direct support to miRNA:mRNA interactions, they fail to detect the specific miRNA recognition elements (MREs) responsible for the understanding of the structural characteristics of the interaction. Consequently, reporter genes should bind to the original and mutated sequences of the gene of interest. Gene expression in both samples is then computed before and after miRNA transfection. Therefore, the identification of the specific site of interaction is enabled.

Due to the fact that the experimental data size that uses reporter gene assays is limited, a different experimental evaluation strategy is adopted. The measurement of expression of genes is now implemented through overexpression or knockdown of a specific gene. In the former case, a decrease in expression of target mRNAs and proteins is expected (down regulation) with increased expression of miRNAs [52], while in the latter case, an increase in expression of target mRNAs and proteins is expected (upregulation) with the miRNA expression silenced in cells [53].

On the other hand, high-throughput techniques including microarrays, proteomics as well as sequencing-based methodologies such as RNA-Seq [54], HITS-CLIP [55], PAR-CLIP [56] and Degradome-Seq [57] (**Table 2**) indirectly recognize various miRNA targets and measure differential gene expression in the presence or absence of a miRNA in the cell.

In particular, microarrays detect putative miRNA-gene interactions, as a high-throughput version of qPCR and northern blotting. Similarly, quantitative proteomic techniques such as stable isotope labelling by/with amino acids in cell culture (SILAC) [58] and pSILAC (pulsed SILAC) [59] are considered a high-yield generalization of ELISA assays and western blots, which are able to provide a high throughput dataset by using mass spectrometry (MS).

The rapid development in next-generation sequencing (NGS) has provided novel insights in the way miRNA-gene interactions are detected. These methods concentrate on NGS sequencing of mRNA sites bound by the Argonaute (AGO) protein. Indeed, RNA immunoprecipitation together with sequencing (RIP-seq) enables the detection of RNAs bound by a protein of interest. In addition, ribosome profiling sequencing (RPF-seq) is a sensitive method that enables the quantification and identification of the mRNAs at the ribosome as well as calculates the efficiency and speed of translation.

CLIP-based techniques involve RNA-protein cross-linking followed by immunoprecipitation against a protein of interest and concentrate on the transcriptome-wide recognition of RNA–protein binding regions. HITS-CLIP (high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation) was the first method that mapped miRNA binding sites in the transcriptome [55]. Nevertheless, due to the vast range of the detected regions, the site where the miRNA exactly binds is difficult to be defined. On the other hand, PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation), which incorporates 4-thiouridine (4SU) into transcripts of cultured cells, identifies the RNA binding sites by scoring for thymidine (T) to cytidine (C) transitions in the sequenced cDNA in the AGO-miRNA-RNA cross-linked regions [60]. In this technique, the detected binding locations are limited in length and edgier compared to those obtained by HITS-CLIP, whereas the precise MREs are easier to be found taking into account the T-to-C mutations near the region possessed by the RNA-induced silencing complex (RISC) [61]. However, the aforementioned methods fail to define the concrete miRNA that participates in the interaction experimentally. Recent altered versions of the above techniques, such as CLEAR-CLIP [62] and CLASH [63] protocols, contain an extra ligation step which connects miRNA molecules with their respective target-binding site in the mRNA, generating and sequencing hundreds of chimeric miRNA–mRNA fragments.

Table 2. Overview of experimental methods for miRNA-Gene interaction Identification [64, 65, 66]

Method	**Direct Technique**	**Throughput**	**Principle**	**Use in experiment**
Luciferase reporter gene [54]	✓	Low	A luciferase reporter contains the 3'UTR target sites of a miRNA, which the miRNA mimics of the miRNA of interest should target	Validation and identification of interacting miRNA-Gene regions
Northern Blotting [54]	-	Low	Investigate the level of miRNA or mRNA of interest with TaqMan or Sybr green probes	Effect of miRNA on mRNA levels
qPCR [54]	-	Low	Investigate the level of miRNA or mRNA of interest with TaqMan or Sybr green probes	Quantification of miRNA effect on mRNA levels
Western Blot [54]	-	Low	Electrophoresis to separate proteins by molecular mass, which are then identified with antibodies concrete to the target protein	Evaluation of miRNA effect on protein concentration
ELISA [54]	-	Low	ELISA detect antibodies or infectious agents in simple enzyme assays	Quantification of miRNA effect on protein concentration
5′ RLM-RACE [54]	-	Low	An RNA adapter is ligated to the free 5′ phosphate of an uncapped mRNA. The ligation product is reverse transcribed using a forward primer directed against the linker and a gene specific reverse primer which is then amplified, cloned and identified by sequencing	Identification of cleaved mRNA targets
Microarrays [54]	-	High	Hybridization of miRNAs or genes to	High-throughput evaluation of

			the complementary immobilized probes, identified via fluorescence	miRNA effect on mRNA expression
RNA-Seq [54]	-	High	Massive parallel sequencing of components of DNA from a simple sample	High-throughput evaluation of miRNA effect on mRNA expression
Quantitative Proteomics(SILAC/pSILAC) [56]	-	High	Protein abundance is estimated by mass spectrometry of samples categorized with separate isotopes	High-throughput evaluation of miRNA effects on protein concentration
AGO-IP	-	High	-	Identification of enriched transcripts (miRNAs and mRNAs) in AGO immunoprecipitates
HITS-CLIP [55]	✓	High	UV light to cross-link the RISC complex including miRNA-mRNA -AGO2, followed by immunoprecipitation and sequencing of miRNA and mRNA	Sequencing of AGO binding regions on targeted transcripts
PAR-CLIP [56]	✓	High	Incorporation of 4-thiouridine into transcripts of cultured cells, identification of the RNA binding sites by scoring for thymidine (T) to cytidine (C) transitions in the sequenced cDNA in AGO-miRNA-RNA cross-linked regions	Sequencing of AGO binding regions on targeted transcripts
CLASH [63]/ CLEAR-CLIP[62]	✓	High	-	Sequencing of AGO binding regions on targeted transcripts. Production of chimeric miRNA:mRNA reads for the detection of interactions

Biotin miRNA tagging [54]	-	High/Low	Transfection of cells with biotinylated miRNA duplexes and capture of miRNA:mRNA complexes from cell lysates using streptavidin beads	Pull-down of biotin-tagged miRNAs and approximation of bound transcript content using qPCR (Low yield), microarrays and RNA-Seq (High-throughput)
IMPACT-Seq [67]	-	High	-	High pull-down of biotin-tagged miRNAs, identification of interacting pairs and binding regions
PARE / Degradome-Seq [57]	-	High	Genome-wide identification of the miRNA-induced cleavage products, analysis of patterns of RNA degradation	Detection of cleaved mRNA targets
3Life [68]	✓	High	-	High-throughput reporter gene assay

Experimentally supported miRNA interactions remain disperse in various publications, supplementary material and raw NGS datasets. Consequently, several repositories have been created in an effort to collect, curate, analyze and deliver a centralized access to miRNA-gene experimentally corresponding interactions [64].

DIANA-TarBase's eighth version [60] is the largest currently available manually curated target database, which indexes approximately 670,000 unique experimentally supported miRNA–gene pairs, contributing to the insertion of >1 million miRNA–gene entries. This repository assembles interactions supported by >33 experimental distinct methodologies and applied to ~600 cell types/tissues under ~451 experimental conditions. Since its initial release in 2006, TarBase is constantly increasing its list of experimentally validated miRNA interactions.

The current updated version introduces ~300,000 entries since the previous version. In particular, ~419 manually curated publications and >245 high-throughput datasets, containing both direct and indirect interactions, are incorporated. Indeed, the database provides information concerning the methodology, cell type and tissue of positive and negative miRNA–gene interactions for 18 species, enabling the user to avoid a specific miRNA/gene query to the primer sequences used for cloning experiments.

In the case of direct methodologies, information of the exact miRNA-binding location as detected experimentally and in silico is listed along with the primer sequences used for cloning experiments. TarBase also combines throughput and sequencing experiments including methodologies like reporter genes, western blot, qPCR, proteomics, biotin miRNA tagging, CLIP-seq, CLEAR-CLIP, CLASH, CLIP-chimeric, IMPACT-seq, AGO-IP, RPF-seq, RIP-seq, Degradome, RNA-seq, TRAP, microarrays, HITS-CLIP and PAR-CLIP. The interactions supported by these techniques have been obtained from the analysis of other publications or from Gene Expression Omnibus (GEO) [69] and DNA Data Bank of Japan (DDBJ) [70] repositories.

miRTarBase [71] is another comprehensive database, which collects miRNA-target interactions by implementing a systematic text-mining procedure to select research articles pertinent to functional studies of miRNAs. The updated version of miRTarBase integrates 422,517 curated validated miRNA-target interactions from 4,076 miRNAs and 23,054 target genes gathered from over 8,500 articles for 18 species. The interactions are validated experimentally by low-yield and high-throughput methodologies such as reporter assays, western blot and microarray experiments with overexpression or knockdown of miRNAs. In addition, the database integrates miRNA/mRNA expression profiles obtained from the Cancer Genome Atlas (ACGA) [72].

miRecords [73] indexes a relatively small number of interactions compared to DIANA-TarBase and miRTarBase and the latest version was released in 2010. The validated targets component in this database involves a vast, high-quality manually curated database of experimentally validated miRNA–target interactions with scrupulous documentation. This current component of the database embodies 1,135 records of validated miRNA–target interactions between 301 miRNAs and 902 target genes in seven (7) animal species. The Predicted Targets component stores predicted miRNA targets generated by 11 miRNA target prediction programs. miRecords incorporates interactions from low-yield techniques mainly from reporter genes assays and the majority of data in the database supports a direct type of validation.

Apart from the aforementioned databases, there exist repositories that aim at integrating CLIP-Seq experimental outcomes from diverse RNA binding proteins, while combining them with target prediction algorithms. A pertinent example is StarBase [74], which deciphers miRNA-target interactions and gathers RNA binding proteins from 108 CLIP-Seq (HITS-CLIP, PAR-CLIP, iCLIP, CLASH) datasets.

Chapter II

2. Computational Approaches for the identification of miRNA-mRNA predicted interactions

The introduction of target prediction algorithms accelerates the need to assess the precision and authenticity of resulting scores in order to evaluate their efficiency and select the optimal ones. The existence/non-existence of an interaction between the miRNA and the corresponding target transcript constitutes the fundamental notion of target prediction programs. The performance of the programs can be measured with two (2) statistical parameters, namely sensitivity and specificity. The first one represents the percentage of correctly predicted targets out of total correct ones, whereas the latter one the percentage of correctly predicted among overall predicted ones. In particular,

Sensitivity = true positive / (true positive + false negative)

Specificity = true negative / (true negative + false positive)

Below, thirteen (13) de novo target prediction programs are analyzed and their characteristics are described in detail.

2.1 TargetScan

TargetScan [75] is the first algorithm that predicts miRNA targets in vertebrates. Since then, this tool has been updated to new versions that ameliorate accuracy in prediction. As an option, users can provide as an input the miRNA name, gene name of broadly conserved, conserved, or poorly conserved miRNA families across several species. As output, the algorithm, with the use of multiple regression, detects miRNA target genes in the 3′UTR of protein-coding transcripts by searching for the presence of 8mer, 7mer and 6m (7mer-m8 < 8mer) or within open reading frames (ORFs). Indeed, the program has the capacity to predict conserved sites as well as sites with mismatches in the seed region that are compensated by conserved 3' pairing and centered sites.

In particular, predictions are ranked based on the predicted efficacy of targeting as computed using cumulative weighted context++ score of the sites or the probability of conserved targeting (P_{CT}) [44]. Since within a 3′ UTR multiple target sites can be identified, an aggregate P_{CT} is necessary. In comparison with previous version, new determinants have been integrated in the current release of TargetScan that improve its targeting efficacy when defining the score. As a result, the context++ model considers 26 features such as 3' UTR profiles, which indicate the fraction of mRNA containing each site, updated miRNA families, 3′ compensatory pairing, local AU content, position contribution, calculation of free energy predicted duplexes, seed-pairing stability, target-site abundance, ORF target-site abundance, identification of nucleotide at position 1, 8, 9 and 10, distance from stop codon, 5′-UTR length, AU content at 5′-UTR, 3′-UTR and ORF as well as the number of 7mer-m8, 7mer-A1, 6mer and 8mer sites in the ORF.

2.1.1 Trained Features

From all these features that context++ model integrates, **Table** 3 summarizes the 14 features that are robustly selected through stepwise regression for a specific site.

Table 3. The description of 14 features that context++ model takes into account through stepwise regression [75]

	Feature	Description
miRNA		
	3′-UTR target-site abundance	Number of sites in all annotated 3′ UTRs
	Predicted seed-pairing stability	SPS Predicted thermodynamic stability of seed pairing
	sRNA position 1	Identity of nucleotide at position 1 of the sRNA
	sRNA position 8	Identity of nucleotide at position 8 of the sRNA
Site		
	Site position 8	Identity of nucleotide at position 8 of the site
	Local AU content	AU content near the site
	3′ supplementary	Supplementary pairing at the miRNA 3′ end
	Predicted structural accessibility	log10(Probability that a 14 nt segment centered on the match to sRNA positions 7 and 8 is unpaired)
	Minimum distance	log10(Minimum distance of site from stop codon or polyadenylation site)
	Probability of conserved targeting (P_{CT})	Probability of site conservation that controls dinucleotide evolution and site context
mRNA	ORF length	log10(Length of the ORF)
	3′-UTR length	log10(Length of the 3′ UTR)
	3′-UTR offset-6mer sites	Number of offset-6mer sites in the 3′ UTR
	ORF 8mer sites	Number of 8mer sites in the ORF

The characteristics of TargetScan are described below:

- 3’ pairing contribution: Reflection of consequential miRNA target complementarity outside the seed region. A more negative score demonstrates a more favorable site.
- Local AU content: Reflection of the transcript AU content 30 nt upstream and dowstream of predicted site.
- Position contribution: Reflection of the distance to the nearest end of the annotated UTR of the target gene

- Target site abundance (TA) contribution to context+ score: Reflection of the abundance of target sites of miRNA family in a set of distinct 3'UTRs. A more negative score is associated with lower target site abundance in the set of 3'UTRs.
- Seed-pairing stability contribution to context+ score: Reflection of the stability of a miRNA-target duplex, which is a function of the concentration of (A+U) in the seed region. A more negative score is associated with a weaker seed-pairing stability.

After scrutiny of the results delivered by TargetScan, the following parameters play a pivotal role in its function:

1. Cumulative weighted context++ score: this score estimates the total repression expected from multiple sites of the same miRNA, for each mRNA target predicted.

2. Branch-length score: the sum of phylogenetic branch lengths between species that contain a matching site.
   a. 8mer: Score ≥ 0.8
   b. 7mer-m8: Score ≥ 1.3
   c. 7mer-A1: Score ≥ 1

3. P_{CT} score: The higher the score, the greater the conservation and the greater mRNA destabilization expected.
   a. $0 < P_{CT}$ score<1

4. Aggregate P_{CT}: For each miRNA, this parameter includes the conserved 3′ UTR targets with multiple sites that were missed in the human 3′ UTR annotation, but were present in the mouse annotations

The datasets that were used in the analysis of this algorithm are summarized in **Table 4**.

Table 4. The set of datasets that were used in the build of the contex++ model [75]

Gene expression omnibus (GEO) ID, ArrayExpress ID, or data source	**Reference**
GSM854425, GSM854430, GSM854431, GSM854436, GSM854437, GSM854442, GSM854443	(Bazzini et al., 2012) [76]
GSM1012118, GSM1012119, GSM1012120, GSM1012121,	(Loeb et al., 2012) [77]

GSM1012122, GSM1012123	
E-TABM-232	(Rodriguez et al., 2007) [78]
GSM1122217, GSM1122218, GSM1122219, GSM1122220, GSM1122221, GSM1122222, GSM1122223, GSM1122224, GSM1122225, GSM1122226	(Helwak et al., 2013) [63]
GSM538818, GSM538819, GSM538820, GSM538821	(Hafner et al., 2010) [79]
GSM156524, GSM156532, GSM210897, GSM210898, GSM210901, GSM210903, GSM210904, GSM210907, GSM210909, GSM210911, GSM210913, GSM37599, http://psilac.mdc-berlin.de/download/ (let7b_32h, miR-30_32h, miR-155_32h, miR-16_32h)	(Lim et al., 2005 [52]; Grimson et al., 2007 [23]; Linsley et al., 2007 [80]; Selbach et al.,2008 [56])
E-MTAB-2110	(Tan et al., 2014) [81]
GSM1479572, GSM1479576, GSM1479580, GSM1479584	(Eichhorn et al., 2014) [82]
GSM210897, GSM210898, GSM210901, GSM210903, GSM210904, GSM210907, GSM210909, GSM210911, GSM210913, GSM37599, GSM37601	(Lim et al., 2005 [52]; Grimson et al., 2007 [23])
74 datasets compiled in Supplementary data 4 of Garcia et al. (2011), used as is or after normalization (Supplementary file 1); GSM119707, GSM119708, GSM119710, GSM119743, GSM119745, GSM119746, GSM119747, GSM119749, GSM119750, GSM119759, GSM119761, GSM119762, GSM119763, GSM133685, GSM133689, GSM133699, GSM133700, GSM134325, GSM134327, GSM134466, GSM134480, GSM134483, GSM134485, GSM134511, GSM134512, GSM134551, GSM210897, GSM210898, GSM210901, GSM210903, GSM210904, GSM210907, GSM210909, GSM210911, GSM210913, GSM37599, GSM37601; E-MEXP-1402 (1595297366, 1595297383, 1595297389, 1595297394, 1595297399, 1595297422, 1595297427, 1595297432, 1595297491, 1595297496, 1595297501, 1595297507, 1595297513, 1595297518, 1595297524, 1595297530, 1595297535, 1595297564, 1595297588, 1595297595, 1595297605,	(Lim et al., 2005 [52]; Birmingham et al.,2006 [83]; Schwarz et al., 2006 [84]; Jackson et al., 2006a [85]; Jackson et al., 2006b [86]; Grimson et al., 2007 [23]; Anderson et al., 2008 [87])

1595297614, 1595297621, 1595297627, 1595297644, 1595297650, 1595297662); E-MEXP-668 (16012097016666, 16012097016667, 16012097016668, 16012097016669, 16012097017938, 16012097017939, 16012097017952, 16012097017953, 16012097018568, 251209725411)	
GSM95614, GSM95615, GSM95616, GSM95617, GSM95618, GSM95619	(Giraldez et al., 2006) [88]
GSM1269344, GSM1269345, GSM1269348, GSM1269349, GSM1269350, GSM1269351, GSM1269354, GSM1269355, GSM1269356, GSM1269357, GSM1269360, GSM1269361, GSM1269362, GSM1269363	(Nam et al., 2014) [89]
http://icb.med.cornell.edu/faculty/betel/lab/betelab_v1/Data.html	(Lipchina et al., 2011) [90]
http://psilac.mdc-berlin.de/media/database/release-1.0/protein/pSILAC_all_protein_ratios_OE.txt (miR155)	(Selbach et al., 2008) [56]
GSM416753	(Mayr and Bartel, 2009) [91]
GSM156522, GSM156580, GSM156557, GSM156548, GSM156533, GSM156532, GSM156524, processed and normalized	(Linsley et al., 2007) [80]
GSM37601	(Lim et al., 2005) [52]
GSM363763, GSM363766, GSM363769, GSM363772, GSM363775, GSM363778	(Hausser et al., 2009) [92]

2.1.2 Training data

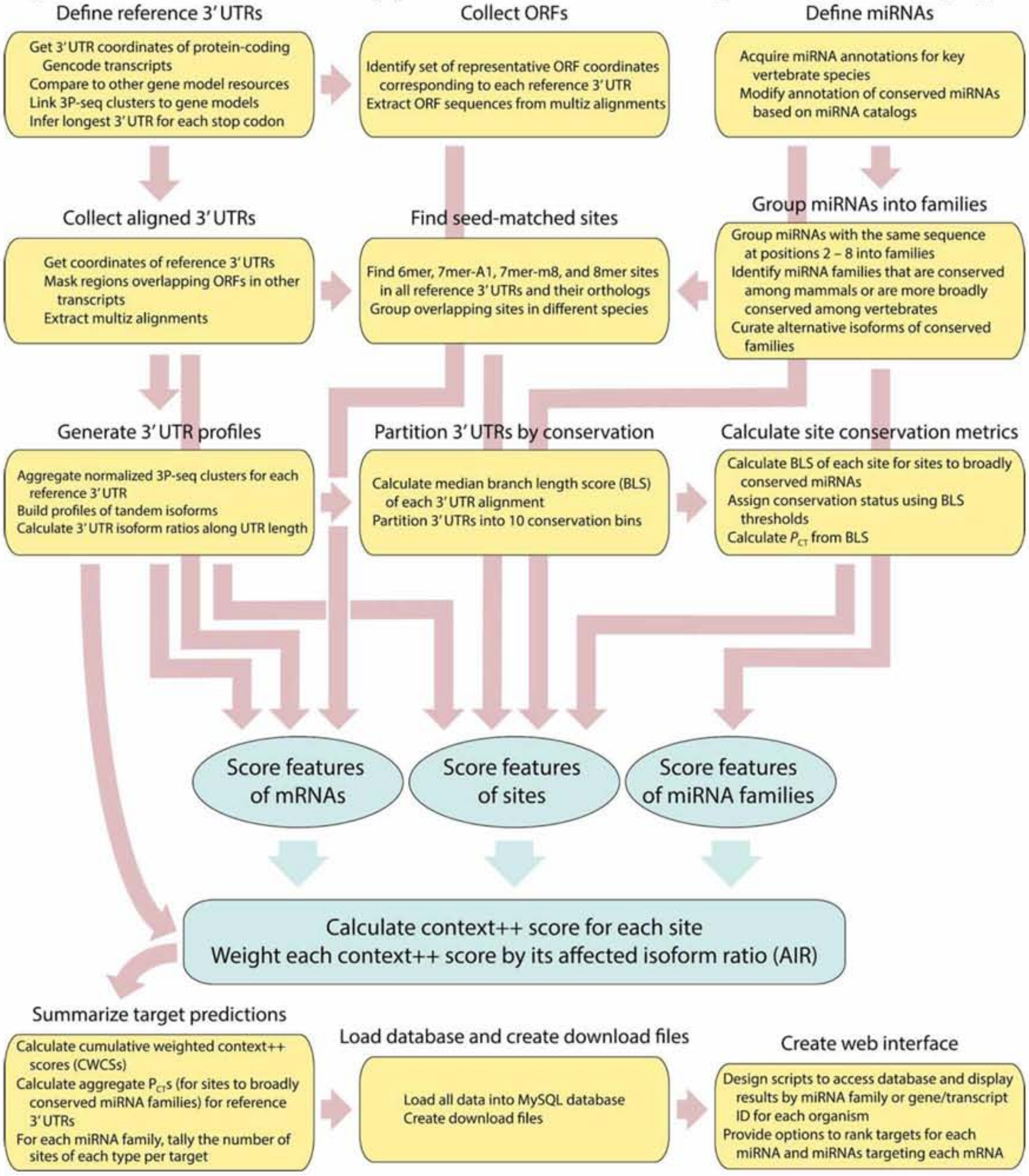
In the study, 1000 bootstrap samples were used, each including 70% of the data from each transfection experiment of the compendium of 74 filtered datasets. For each type of site, stepwise regression was implemented, creating a set of features that was chosen as those that were selected for at least 99% of the bootstrap samples of at least two site types. This set of features along with the entire compendium of 74 datasets as a training set was utilized in a multiple linear regression model for each site type. As a result, scores for 8mer, 7mer-m8, 7mer-A1 and 6mer sites were selected to be no greater than −0.03, −0.02, −0.01, and 0 respectively for each site.

2.1.3 Test data

Samples which contained the remaining 70% of the data from each transfection experiment of the compendium of 74 filtered datasets were reserved and used as a test set.

The overall procedure followed in the build, training and testing Targetscan is indicated in **Figure 9**.

Figure 9. Flow diagram of the pipeline used to build the TargetScan7 database [75].


Define reference 3′ UTRs
Get 3′ UTR coordinates of protein-coding Gencode transcripts
Compare to other gene model resources
Link 3P-seq clusters to gene models
Infer longest 3′ UTR for each stop codon
Collect ORFs
Identify set of representative ORF coordinates corresponding to each reference 3′ UTR
Extract ORF sequences from multiz alignments
Define miRNAs
Acquire miRNA annotations for key vertebrate species
Modify annotation of conserved miRNAs based on miRNA catalogs
Collect aligned 3′ UTRs
Get coordinates of reference 3′ UTRs
Mask regions overlapping ORFs in other transcripts
Extract multiz alignments
Find seed-matched sites
Find 6mer, 7mer-A1, 7mer-m8, and 8mer sites in all reference 3′ UTRs and their orthologs
Group overlapping sites in different species
Group miRNAs into families
Group miRNAs with the same sequence at positions 2 – 8 into families
Identify miRNA families that are conserved among mammals or are more broadly conserved among vertebrates
Curate alternative isoforms of conserved families
Generate 3′ UTR profiles
Aggregate normalized 3P-seq clusters for each reference 3′ UTR
Build profiles of tandem isoforms
Calculate 3′ UTR isoform ratios along UTR length
Partition 3′ UTRs by conservation
Calculate median branch length score (BLS) of each 3′ UTR alignment
Partition 3′ UTRs into 10 conservation bins
Calculate site conservation metrics
Calculate BLS of each site for sites to broadly conserved miRNAs
Assign conservation status using BLS thresholds
Calculate P_{CT} from BLS
Score features of mRNAs
Score features of sites
Score features of miRNA families
Calculate context++ score for each site
Weight each context++ score by its affected isoform ratio (AIR)
Summarize target predictions
Calculate cumulative weighted context++ scores (CWCSs)
Calculate aggregate P_{CT}s (for sites to broadly conserved miRNA families) for reference 3′ UTRs
For each miRNA family, tally the number of sites of each type per target
Load database and create download files
Load all data into MySQL database
Create download files
Create web interface
Design scripts to access database and display results by miRNA family or gene/transcript ID for each organism
Provide options to rank targets for each miRNA and miRNAs targeting each mRNA

2.1.4 Results

This model performs significantly better than existing programs and is characterised the best high-throughput method in vivo crosslinking approaches. This is evident due to the fact that no more false-positive predictions are generated than the best experimental datasets and the union of the CLIP supported targets. In addition, it confirms that miRNAs bind to non-canonical sites despite their inefficacy. In case of canonical sites only, it is observed that the set of canonical CLIP-supported targets outperforms the predictions of Targetscan in some cases, while the repression of the predictions of its contex++ model, with the aid of pulldown-seq or IMPACT-seq data, is more effective than using mRNAs identified biochemically without crosslinking. Finally, it achieves high performance in terms of sensitivity and precision.

2.2 PACCMIT

PACCMIT (Prediction of ACcessible and/or Conserved MIcroRNA Targets) [93, 94] is an algorithm that filters potential miRNA binding sites in 3'UTR regions by their conservation, accessibility, or both and then ranks the predictions according to the over-representation of sites complementary to the miRNA seed, using a Markov-based model.

The algorithm filters the seed matches according to one or the two following criteria and allows looking for shorter or longer matches as well as for matches with varying starting position:

- Accessibility (unrestricted location): for all 4-mers in the seed match, the probability P_u that the 4-mer is accessible (is contained within a single-stranded region of RNA) is calculated. Only the "partially accessible" seed matches (contain at least one 4-mer with P_u>= P_{cutoff}) is chosen.
- Accessibility (restricted location): seed matches are considered as "partially accessible" only in cases where the 4-mer opposite to positions 2 to 5 of the miRNA has P_u>= P_{cutoff}.
- Conservation: Only the seed matches that are conserved in the human, chimp, rhesus and mouse 3'UTRs are selected.

However, the results of this study take into account 7-mers complementary matching to the positions 2–8 of miRNA. As a result, the above possible filters take the form:

- Accessibility: 'access' (only accessible 7-mers are counted) or
- Conservation: 'cons' (only conserved 7-mers are counted or
- Accessibility-Conservation: 'cons+access' (only 7-mers that are both conserved and accessible are counted)
- No filter: whole 3'UTR and all seed matches are considered

Then, for each target sequence the number c_{filter} of seed matches that meet the filter requirements is computed. In addition, the probability $P_{SH,}$ that there are at least c_{filter} seed matches in a random background sequence with the same dinucleotide composition as the real sequence, is calculated. The probability P_{SH} measures the over-representation of the seed matches in a given target sequence. As a result, lower P_{SH} values (higher over-representation) demonstrate higher chances of biological functionality.

2.2.1 Training data

In order to evaluate the conservation filter, the proteomics dataset of Baek et al. [58] and Selbach et al [56] are used. The latter dataset covers three highly conserved miRNAs. In particular, an arbitrary classification of the miRNA-gene pairs between functional and non-functional is performed based on the $\log_2$ fold changes ($\log_2$FC) in protein expression. Specifically, miRNA-gene pairs with $\log_2\text{FC}\leq-0.2$ are considered functional targets, while the remaining pairs are considered non-functional targets.

2.2.2 Test data

Positive and negative datasets using the binding sites reported in the PAR-CLIP experiments [79] are formed, in order the effect of different filters on sensitivity and precision of target predictions for highly and weakly conserved miRNAs, to be tested.

2.2.2.1 Positive datasets

Only the 100 most abundant miRNAs from [79] are utilized and are divided into two groups containing 74 highly conserved and 26 weakly conserved miRNAs. In each group, in functional miRNA-gene pairs at least one 7-mer matching miRNA positions 2–8 is found between positions 21 and 30 of the regions mapped to the 3′UTRs. Overall, 3,698 such positive interactions are found.

2.2.2.2 Negative datasets

Initially, all genes, for which there is no evidence of AGO binding in the entire transcript but they contain seed matches in their 3′ UTR, are gathered. Then, for each group of miRNAs, all possible combinations between the miRNAs and the previous genes are produced. As a result, the negative datasets of non-functional pairs are constructed by randomly selecting N pairs from the previously generated combinations, where N equals the number of functional pairs found for the same group of miRNAs. In the case of highly conserved miRNAs, N=3,586, while in the case of weakly conserved miRNAs, N=112. Given to the abundance of negative interactions, only 3,698 randomly chosen negative interactions are used for the analysis for the purpose of balancing positives and negatives sets.

Due to the analysis of the statistical significance of the precision and sensitivity values of the different methods, each dataset of 2*N validated pairs (N functional and N non-functional) is partitioned into three smaller sets. The subsets of highly conserved miRNAs contain 25, 25 and 24 miRNAs. The subsets of weakly conserved miRNAs contain 9, 9 and 8 miRNAs. The statistical significance of the difference between various methods is evaluated with the one-sided t_{test}.

2.2.3 Computation of accessibility

Accessibility is used in two of the aforementioned criteria that filter the seed matches. Indeed, any 7-mer in the 3′UTR sequence (including seed matches) is labeled as accessible if it contains at least one 4-mer unpaired with a probability $P_{free} \geq P_{cutoff}$, where P_{cutoff} has an optimized value of 0.2. Calculation of P_{free} values for all 4-mers in all the human 3′UTR sequences is implemented with the program RNAplfold with parameters W(window)=80 and a maximum pairing distance L=40.

2.2.4 Scoring scheme

The predicted miRNA-3′UTR interactions are ranked according to the single hypothesis P value (P_{SH}), which considers simultaneously single and multiple binding sites as well as facilitates accessibility and/or conservation filters. P_{SH} is defined as an approximate probability that a given oligomer (e.g., a 7-mer), complementary to the miRNA seed, is found by chance at least c times in the corresponding 3′UTR. Lower values of P_{SH} imply that the interaction is more likely to be functional. P_{SH} is computed as:

$$P_{SH} = \sum_{i=c_{filter}}^{t_{filter}} \binom{t_{filter}}{i} P^{i}(1-P)^{t_{filter}-i} \qquad (1)$$

Where t_{filter} is the number of 7-mers in a 3′UTR sequence that meet the filter requirement and c_{filter} is the number of seed matches that meet the filter requirement.

In case both accessibility and conservation filters are used, P_{SH} is computed using Eq. (1) where t_{filter} equal to the total number $t_{cons+access}$ of 7-mers in the 3′UTR that are both conserved and accessible (regardless of their complementarity to the seed) and c_{filter} equals to the number $c_{cons+access}$ of conserved and accessible seed matches.

2.2.5 Results

The conservation of the binding site is counted by the number of species with the same sequence and/or by the phylogenetic distance between the species sharing the same sequence. In site conservation, the following approaches are used:

- “Any-species” (Any-S) approach: the seed match must be present in the aligned sequences of at least S species (including the human), regardless of their distance from the human.
- “Selected-species” (Selected-S) approach: the seed match must be present in the aligned sequences of specific S species. The (S+1)st added species is pre-selected and is more distant from the human than the preceding S species. Only the species, in which the seeds of the eight miRNAs from the proteomics datasets that are conserved, are included.

As a result, for both implementations and for all levels of stringency of the conservation filter, PACCMIT has better performance with the conservation filter than without it and S=12 is chosen to be the cut-off.

In addition, it is found that for highly conserved miRNAs, the conservation filter finds more true targets per miRNA and the precision is higher than in the accessibility filter, as both sensitivity and precision are higher among the top predictions. However, in the case of weakly conserved miRNAs, the conservation filter has worst performance not only compared to the accessibility filter but in many cases also to the algorithm with no filter at all, proposing that the site conservation is not equally effective in rejecting false positive predictions for all miRNAs. Consequently, criteria such as accessibility should be considered. Alltogether, 48.5 targets per highly conserved miRNA and only 4.3 targets per weakly conserved miRNA are obtained. This difference is justifiable due to the fact that highly conserved miRNAs possibly accumulate more targets throughout evolution. Moreover, the combined filter of conservation and accessibility appears to slightly improve the sensitivity, precision, and quality (measured by downregulation of targets) of the top predictions of highly conserved miRNAs compared to the application of each filter separately. Finally, it is concluded that the top ranking predicted miRNA-gene interactions correspond to more downregulated proteins than do the lower ranking predicted miRNA-gene interactions.

2.3 PACCMIT-CDS algorithm

PACCMIT–CDS [93, 95] is an unbiased algorithm, which predicts both conserved and non-conserved miRNAs targets within coding sequences (CDS) by searching for conserved motifs that are complementary to the microRNA seed region. Then, miRNA-gene interactions are ranked according to overrepresentation of conserved seed matches preserving the codon usage and the amino acid sequence. Indeed, the best scoring interactions are mapped to complementary sites that are preserved only for gene regulation. The algorithm succeeds a lower rate of false positives and better ranking of predictions than existing methods in the 3′ UTR.

2.3.1 Training data

Genomic coordinates of Ensembl human genes in hg18 are used to derive the human coding sequences available at the UCSC Table browser subject to the formation of the mRNA sequences. Overall, 21,426 coding sequences are examined. Similarly, 1,919 miRNA sequences used in this study are extracted from the miRBase v18 [96] (http://www.mirbase.org).

2.3.2 Important parameters

The extent of overrepresentation is calculated by defining the probability P_{SH}, which denotes that a specific seed match would be found in a specific sequence at least c times by chance. c declares the number of seed matches in the real sequence.

Lower P_{SH} values and therefore stronger overrepresentation contain a higher likelihood of biological functionality. Specifically, probability P_{SH} is computed as:

$$P_{SH} = \frac{N_c}{N_{total}} = \frac{number\ of\ random\ sequences\ with\ at\ least\ c\ seed\ matches}{total\ number\ of\ random\ sequences}$$

The precision of the algorithm is further increased by implementing the conservation filter of the seed match. In particular, if c_{cons} is the number of conserved seed matches observed in the real sequence, the formula is modified as:

$$P_{SH} = \frac{N_{ccons}}{N_{total}} = \frac{number\ of\ random\ sequences\ with\ at\ least\ ccons\ conserved\ seed\ matches}{total\ number\ of\ random\ sequences}$$

P_{SH} intervals are declared as $10^{-(n+1)} \leq P_{SH} < 10^{-n}$, for n = 0, 1 …, 7. Due to the fact that the best resolution of P_{SH} with 10^8 random sequences is 10^{-8}, the last interval is simply defined as $P_{SH} < 10^{-8}$.

Another crucial parameter in this analysis is the signal-to-noise ratio, which is defined for a given P_{SH} interval: the ratio between the predictions within this P_{SH} interval in two genomes: the real genome (signal) and the random background (noise).

Moreover, due to the fact that conservation of seed matches in the human genome among different species is considered, a 28-species alignment file (MAF file) is utilized, which is available at the UCSC Table browser. In particular, as in the case of PACCMIT, the "Any-species" approach introduced by Marin and Vanicek [94] is employed. Indeed, a seed match is conserved if it is displayed in the aligned sequences of at least S species (including the human), regardless of their phylogenetic distance from the human. The threshold S = 12 is concluded to be the ideal cut-off, as is shows that conservation in 12 species increases the performance of the algorithm in terms of precision.

The online predictions are available at *https://lcpt.epfl.ch/PACCMIT-CDS*, consider 7-mer seed matches and use the PS+CU protocol.

2.3.3 Generation of random background sequences

4 different protocols are utilized in the production of random background sequences:

NR: No restrictions imposed. Each codon is accidentally replaced by any other codon of the genetic code, even if it was present or absent in the original sequence

PS: Preservation of the protein sequence. Each codon is replaced randomly only with synonymous codons

CU: Preservation of codon usage by shuffling all the codons present in the real sequence

PS + CU: Preservation of protein sequence and codon usage. Shuffling only of codons that encode the same amino acids.

The shuffling in CU and PS + CU is implemented by a modified version of the Fisher-Yates shuffling algorithm (Knuth [97]). In an array with N elements, the following steps are carried out:

- Traversal of array elements with indices i from 1 to N−1
- Selection of a random integer j, for each index i, satisfying $i \leq j \leq N$
- Swap of i_{th} and j_{th} elements

The shuffling algorithm achieves linear complexity in N in contrast to the original quadratic implementation of Fisher and Yates. More significantly, all N! different permutations are generated with equal likelihood.

2.3.4 Test data

For the computation of the precision and sensitivity of the algorithm, 4,376 interactions, which are included in the positive and negative data sets and use the binding sites reported in the PAR-CLIP experiments, [79] are considered. In particular, PAR-CLIP experiments provide direct information about physical binding between miRNA and mRNA. In addition, following the procedure from Marin and Vanicek [94], from a set of 100 most abundant miRNAs, 74 conserved miRNAs are selected. As a result, 2,188 highly reliable positive interactions are obtained, defining true targets as genes whose coding region contains at least one seed match, overlapping with an AGO-bound region.

As far as the negative data set is concerned, genes for which no seed match overlaps with any region of the whole transcript (5′ UTR, CDS, or 3′ UTR) are taken into account. Among these unbound genes, only those, which contain at least one 7-mer complementary to positions 2–8 of any of the 74 conserved miRNAs, are retained. As a result, the negative data set of nonfunctional pairs is formed by randomly selecting 2,188 pairs from the list of unbound pairs.

Moreover, proteomics data of [56], which provide the protein fold changes calculated after overexpression of five conserved miRNAs, are also tested.

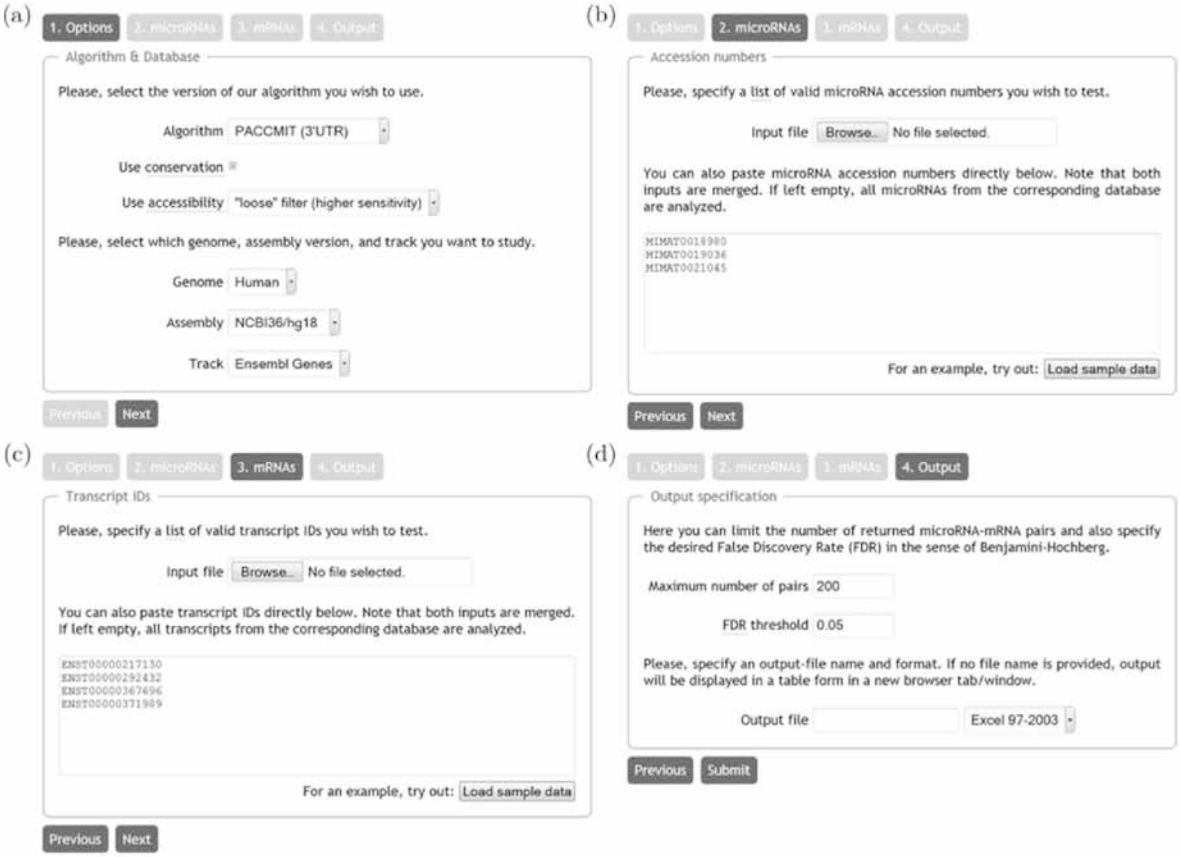
2.3.5 Results

In the study of PACCMIT-CDS, PS + CU protocol is the background model which is finally employed due to the least contaminated noise. In addition, applying conservation of binding sites throughout evolution increases the signal-to-noise ratio as well as the likelihood that the predictions in that P_{SH} interval are highly probable to be functional. As far as the precision curves of the algorithm are concerned, the PS + CU protocol, as a background model, removes more false positives than other randomization schemes and site conservation increases the precision of the algorithm by as much as ~20%, independently of the background used. Another experiment includes the effect of the alteration of the length of the seed match on sensitivity and precision of the algorithm. In examining 6-, 7-, and 8-mers, it is found that 7-mers outperforms the 6-mers and 8-mers in terms of precision in most of the sensitivity range and that 7-mers are more reliable than 8-mers. The same behavior is observed both in PAR-CLIP and proteomics data sets, requesting always site conservation.

2.3.6 Web Server

Both PACCMIT and PACCMIT-CDS algorithms are incorporated in a web server at *http://paccmit.epfl.ch*. Initially, the user must select one of the two existing algorithms, either PACCMIT for transcripts targeted in the 3′ UTRs or PACCMIT-CDS for transcripts targeted in the CDSs. As an option, the user can also increase the precision of the selected algorithm by enforcing conservation (available for both PACCMIT and PACCMIT-CDS) and/or accessibility (available for PACCMIT only) filters. In addition, the user defines the miRNAs and mRNAs to be analyzed. Then, the web server detects candidate miRNA-target pairs, ranks them according to a P-value, assesses the statistical significance according to the false discovery rate and provides the predictions, containing a list of miRNA/mRNA pairs in multiple downloadable formats. The algorithm only predicts if the given mRNA as a whole is possible to be targeted by the given miRNA. An overview of the web server is showed in **Figure 10**.

Figure 10. Review of the main user interface of the PACCMIT/PACCMIT-CDS web server. **(a)** Selection of the algorithm: either PACCMIT or PACCMIT-CDS. Application of filters: accessibility (PACCMIT only) and/or conservation (PACCMIT and PACCMIT-CDS. Selection of the genome, assembly and track of interest. PACCMIT supports the subsequent database combinations **i.** NCBI36/hg18, Ensembl genes, **ii.**GRCh38/hg38, Ensembl genes, **iii.**GRCh38/hg38, RefSeq genes. PACCMIT-CDS supports only the NCBI36/hg18, Ensembl genes option. **(b)** Insertion of miRNAs of interest. miRNAs are pasted as accession numbers directly into the web site or are uploaded as a text file. Duplicates are automatically eliminated. **(c)** Insertion of mRNAs of interest. mRNAs are inserted either as transcript IDs or RefSeq IDs **(d)** Specification of the output parameters. The predicted miRNA/mRNA pairs can be exported to Microsoft Excel (version 97 and newer) and CSV files. In case the format of the output file is not specified, the predictions are displays within the browser as HTML format. In the results generated from the database, apart from the unique miRNAs and mRNAs (if any), items not found in the search ('discarded') and miRNA accession numbers or mRNA transcript IDs present in the database and thus in the search, appear in tabular form. The miRNA target predictions contain the following columns: Transcript ID, Gene ID, Accession, ID, log(p-value), log(p-value adj.), Gene Name, Gene Description, Seed match positions.



Please, select the version of our algorithm you wish to use.

Algorithm PACCMIT (3'UTR)

Use conservation □

Use accessibility "loose" filter (higher sensitivity)

Please, select which genome, assembly version, and track you want to study.

Genome Human

Assembly NCBI36/hg18

Track Ensembl Genes

Please, specify a list of valid microRNA accession numbers you wish to test.

Input file No file selected.

You can also paste microRNA accession numbers directly below. Note that both inputs are merged. If left empty, all microRNAs from the corresponding database are analyzed.

MIMAT0018980
MIMAT0019036
MIMAT0021045

For an example, try out:

Please, specify a list of valid transcript IDs you wish to test.

Input file No file selected.

You can also paste transcript IDs directly below. Note that both inputs are merged. If left empty, all transcripts from the corresponding database are analyzed.

ENST00000217130
ENST00000292432
ENST00000367696
ENST00000371989

For an example, try out:

Here you can limit the number of returned microRNA-mRNA pairs and also specify the desired False Discovery Rate (FDR) in the sense of Benjamini-Hochberg.

Maximum number of pairs 200

FDR threshold 0.05

Please, specify an output-file name and format. If no file name is provided, output will be displayed in a table form in a new browser tab/window.

Output file Excel 97-2003

2.4 MIRZA-G

MIRZA-G (MIRZA-Genome-wide) [98] constitutes a method that can predict canonical, non-canonical miRNA targets and siRNA off-targets with similarly high accuracy, using evolutionary conservation. A fundamental part of the model includes the miRNA–target interaction energy predicted by the MIRZA biophysical model that had been inferred from Argonaute 2 crosslinking and immunoprecipitation (Ago2-CLIP) data [99]. This model's implementation assigns base binding energies on the candidate miRNA-mRNA duplexes. Along with the MIRZA-predicted energy of interaction, the model includes features such as the nucleotide (nt) composition around putative target sites, their structural accessibility and location within 3′ untranslated regions (3′ UTRs), predictive for functional miRNA target interactions. The training and testing of the algorithm was performed using two generalized linear models (GLMs) with the logit function (logistic regression) against miRNA/siRNA transfection microarray and proteomics datasets.

2.4.1 Training Data

For the training and the evaluation of the model, 26 experiments were carried out by seven different groups, in which the changes in gene expression are induced by the transfection of individual miRNAs, are measured. A summary of the experimental data sets is presented in **Table 5**. Data are processed to obtain the log2 fold changes in gene expression levels upon transfection of individual miRNAs. Similarly, the changes in gene expression, induced by 12 different siRNA transfected individually, are measured by Birmingham et al. [83] and processed by van Dongen et al. [100] to infer siRNA off-target presence.

Table 5. Summary of the experimental data sets utilized for training and assessing the model.

Reference	Data source (Gene Expression Omnibus (GEO) accession / URL)	miRNAs in the data set
Dahiya et al. [101]	GSE10150	miR-200c, miR-98
Frankel et al. [102]	GSE31397	miR-101
Gennarino et al. [103]	GSE12100	miR-26b, miR-98
Hudson et al. [104]	GSE34893	miR-106b
Leivonen et al. [105]	GSE14847	miR-206, miR-18a, mir-193b, miR-302c

Linsley et al. [80]	GSE683	miR-103, miR-215, miR-17, miR-192, let-7c, miR-106b, miR-16, miR-20a, miR-15a, miR-141, miR-200a
Selbach et al. [56]	http://psilac.mdc-berlin.de/download/	miR-155, let-7b, miR-30a, miR-1, miR-16
Jackson et al. [86]	http://www.ncbi.nlm.nih.gov/geo/	10 siRNAs (PIK3CB-6338, PIK3CB-6340, MAPK14–193, MAPK14-pos2-mismatch, MAPK14-pos3-mismatch, MAPK14-pos4-mismatch, MAPK14-pos5-mismatch, MAPK14-pos6-mismatch, MAPK14-pos7-mismatch and MAPK14-pos8-mismatch) *Preparation of samples 24 h after transfection*

2.4.2 Trained Features

2.4.2.1 MIRZA target quality score

The computation of the MIRZA target quality score is essential for the prediction of miRNA/siRNA target sites. Windows of fixed length, 50 nts, in 3′ UTRs are utilized due to the dependence of target quality score on the length of the putative target site. Initially, a minimization of target quality score that results from the reanalysis of 2,998 sites from Khorshid et al. [99] is crucial for training the MIRZA model. Then, for each site, the miRNA with the highest target quality score is detected and the calculation of the highest-scoring hybrid structure between this miRNA and the CLIPed site takes place. What is more, each putative site is classified as canonical or non-canonical and those with target site quality scores greater or equal to 50 are retained.

2.4.2.2 Position of the target site in 3′ UTRs

The minimum between the distance from the beginning of the seed-complementary region to the stop codon and to the poly-A tail.

2.4.2.3 Nucleotide content

'Flanks G content': proportion of G nts within 50 nt upstream and 50 nt downstream of the miRNA seed-matching region.

'Flanks U content': U nts, respectively, within 50 nt upstream and 50 nt downstream of the miRNA seed-matching region.

2.4.2.4 Accessibility

Structural accessibility of the target site is declared as the probability that the 21 nucleotide long region (located on the right-hand side by the nucleotide matching the 5′-most nucleotide of the miRNA seed) is in single-stranded conformation, across all possible secondary structures. The calculation of this probability is implemented with CONTRAfold, a method for RNA secondary structure prediction that is based on conditional log-linear models (CLLMs). In addition, the indispensable energy which opens the secondary structure of the target region is computed with the RNAup program from the Vienna package. Accessibilty is also computed in the seed-complementary region and in extended target site.

2.4.2.5 Branch length score

For the computation of the branch length score, the following alignments are necessary:

- Alignment of the 3′ UTR sequences to the human genome (hg19) with GMAP (A Genomic Mapping and Alignment Program for mRNA and EST Sequences)
- Pairwise alignments of the human genome (hg19) to genomes of 41 other species from UCSC (*http://hgdownload.cse.ucsc.edu/downloads.html#human*)
- Assessment of the level of evolutionary conservation of putative target sites through the genomic region of the human 3′ UTRs serving as anchor
- Phylogenetic tree of 46 species (including Homo sapiens) (*http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/*)
- Elimination of species for which pairwise alignments to human are unavailable

For each putative target site in humans the following steps are performed:

- Extraction of the region that corresponds to the putative target site in the human 3′ UTR in all other species.
- Computation of the target quality score of the putative target sites with the human miRNA and consideration of the target site when the target quality score was at least 50.

- Computation of the fraction of the total evolutionary distance in the phylogenetic tree along (branch length score). All manipulations of the phylogenetic tree were performed with DendroPy package.
- Comparison of the estimates of selection pressure obtained beforehand with the posterior probabilities that individual putative target site are under evolutionary selective pressure, calculated with the ElMMo method [106]. The aforementioned method supports seed-matching miRNA-complementary sites only.

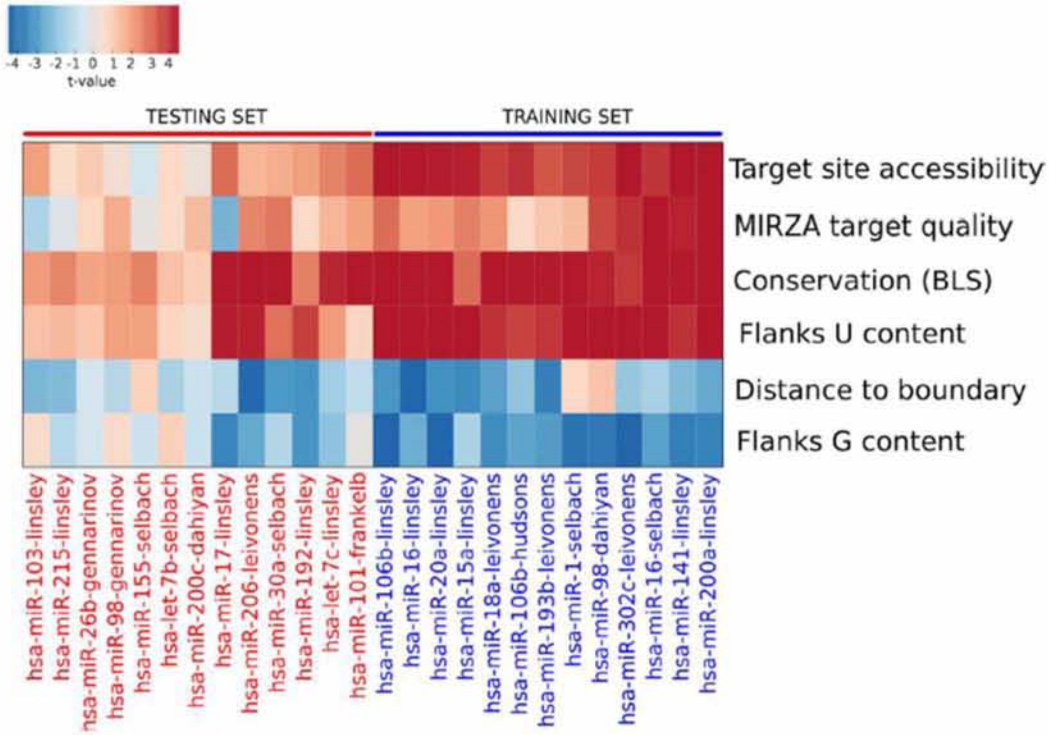
2.4.3 Machine Learning Models - Training of the generalized linear model

Two generalized linear models (GLMs) are trained with the logit function (logistic regression) to classify the training data. The first model includes the branch length score as a feature, while the second model is not. For each experiment, the 100 most downregulated and the 100 least-changing (whose log fold-change is closest to 0) transcripts are extracted with a single putative miRNA binding site in the 3′ UTR. Therefore, these transcripts provided the 100 positive and the 100 negative target sites in the experiment. Furthermore, for each site the trained features described above are computed. The evaluation of the power of the trained features is implemented with the use of two-sample t-tests for the difference of the mean values of a given feature between the positive and negative target sites in each experiment. Finally, the subset of experiments, in which the differences between the positive and the negative subsets of sites are the most significant (most significant t-values), is used to train the model.

2.4.4 Test data

In the test dataset only putative canonical sites of miRNAs are utilized. The other subset of experiments, in which the differences between the positive and the negative subsets of sites are not the most significant, is used for testing the model.

Figure 11. Value of t-statistic in comparing the mean values of features used in the model (rows) among functional and non-functional miRNA seed-complementary sites across 26 experiments (columns). The data from the experiments labeled in blue were used to train the model and those from experiments labeled in red were used in testing the model.



In this study, four different models are implemented, which lead to the prediction of diverse target site types. The features that each model incorporates and the site type that target are summarized in **Table 6**.

Table 6. Four alternative MIRZA-G models

Model name	Features	Target site type
seed-MIRZA-G	MIRZA target quality score, structure accessibility, nucleotide composition of flanks, distance to boundary	canonical
seed-MIRZA-G-C	MIRZA target quality score, structure accessibility, nucleotide composition of flanks, distance to boundary, evolutionary conservation	canonical
MIRZA-G	MIRZA target quality score, structure accessibility, nucleotide composition of flanks, distance to boundary	canonical and non-canonical
MIRZA-G-C	MIRZA target quality	canonical and non-

	score, structure accessibility, nucleotide composition of flanks, distance to boundary, evolutionary conservation	canonical

As far as the impact siRNAs provoke to pathways is concerned, 3′ UTRs are scanned for matches to the seed regions of all siRNAs, therefore obtaining ∼50 million distinct matches. For each of these putative target sites, the trained features described above are calculated. In the end, per-gene scores for all siRNAs are determined.

2.4.5 Results

The MIRZA-G variant that uses evolutionary conservation performs better than currently available methods, in predicting canonical miRNA target sites. What is more, it undergoes the strongest down-regulation in response to miRNA transfection and predicts non-canonical miRNA target sites with similarly high accuracy. In addition, it is found that the most predictive feature that contributes to the mRNA degradation is the branch length score, which can readily be computed for non-canonical sites. Furthermore, MIRZA-G variants predict siRNA off-target sites with accuracy unmatched by currently available programs, whether only canonical or both canonical and non-canonical sites are taken into account. Especially, in case of evolutionary conservation of the siRNA-complementary sites, a stronger downregulation of the predicted mRNA targets is observed.

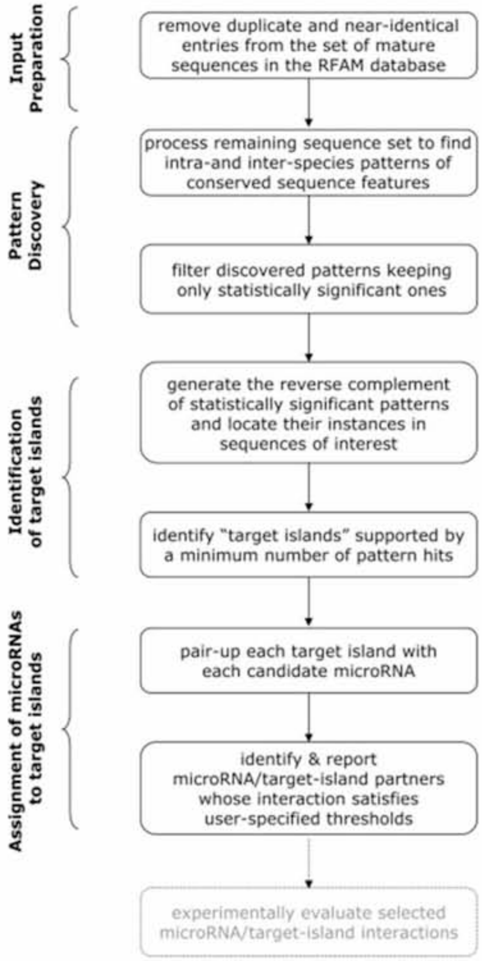
2.5 RNA22

RNA22 [34] consists of a pattern-based approach for the discovery of microRNA binding sites and their corresponding microRNA/mRNA complexes. RNA22 has high sensitivity, does not rely upon cross-species conservation, is resilient to noise and can be applied to the analysis of any genome without requiring genome-specific retraining. Its characteristic of avoiding the use of a cross-species sequence conservation filter, permits the discovery of microRNA binding sites that may not be present in closely related species. Putative microRNA binding sites can be identified when the identity of the targeting microRNA is unknown, thus enabling the identification of binding sites, even if the targeting microRNA is not among those currently known.

In addition, the method’s fundamental idea was extended to a low-error microRNA-precursor-discovery scheme. As a result, the true numbers of microRNA precursors of a genome, microRNA binding sites and affected gene transcripts may be substantially higher than currently hypothesized and in along with 3’-UTRs, numerous binding sites likely exist in 5’-UTRs and CDSs.

The flowchart, shown in **Figure 12**, presents the various steps of the RNA22 method. Firstly, during input preparation, all duplicates and seldom identical entries are extracted from the set of mature sequences. Through the pattern discovery step, salient, conserved sequence features are identified and are portrayed with the aid of patterns. As a whole, these patterns reflect the original body of knowledge but in a diverse and superfluous way. During the identification of Target Sites' step, sequence segments from the primary input are illustrated by multiple patterns, each of which appears in two or more sequences. The term "target island" is applied to any hot spot where the reverse complement of mature microRNA patterns aggregated. Due to the high number of patterns clustering around specific UTR locations and the "guilty-by-association" approach [107], such "hot spots" are matched with putative microRNA binding sites. Finally, during the association of MicroRNAs with Target Islands step, putative microRNA binding sites, in the form of a target islands, are allocated and the establishment of the identity of the microRNA(s) that will bind to it, is taking place. Indeed, each one of the available microRNAs is paired with each generated target island, for all possible relative offsets.

Figure 12. Flowchart shows the Various Steps of the Method.


Input Preparation
remove duplicate and near-identical entries from the set of mature sequences in the RFAM database
Pattern Discovery
process remaining sequence set to find intra-and inter-species patterns of conserved sequence features
filter discovered patterns keeping only statistically significant ones
Identification of target islands
generate the reverse complement of statistically significant patterns and locate their instances in sequences of interest
identify "target islands" supported by a minimum number of pattern hits
Assignment of microRNAs to target islands
pair-up each target island with each candidate microRNA
identify & report microRNA/target-island partners whose interaction satisfies user-specified thresholds
experimentally evaluate selected microRNA/target-island interactions

2.5.1 Training data - Input Preparation

644 mature microRNA sequences, obtained from Release 3.0 (January, 2004) of RFAM [108] were processed. The 2-year-old release of the RFAM database was necessary in order to measure the ability of the method to deduce from a small repository of available knowledge. Before processing, identical and near-duplicate entries from this collection were eliminated using a scheme [109] in reference to BLASTN [110]. Indeed, no two remaining sequences from the ultimate set of 354 ones advocate on more than 90% of their positions.

2.5.1.1 Extension to the Discovery of MicroRNA Precursors

The 719 microRNA precursor sequences, contained in Release 3.0 of RFAM (01/2004), were utilized as training dataset. A no redundant set of 530 sequences arose posterior to the extraction of identical and near-duplicate entries. Moreover, the experiment was repeated by considering the precursor sequences contained in the December 2005 instance of RFAM as training data.

2.5.2 Test data

As far as testing data are concerned, three main validation methodologies were applied. The first one includes the prediction of the microRNA binding sites of previously reported heteroduplexes. The aforementioned was facilitated by the training of RNA22, utilizing a January 2004 instance of the RFAM database. As a result, any binding sites that RNA22 forecasted and appeared in the literature after January 2004, were equivalent to fitting de novo predictions. Indeed, RNA22 effectively distinguish 81%, or 17 out of the 21 full-length binding sites, for instance sites with base-pairing that extends beyond the microRNA's seed or nucleus region.

The second method aims at the identification of the correct microRNA for previously reported heteroduplexes. After the determination of all target islands in a given UTR, it was crucial to shape complexes between the islands and each candidate microRNA as well as report the microRNA which satisfies the user-specified M, G and E parameters. For this purpose, this process was employed to the 17 previously reported, full-length binding sites. Consequently, RNA22 precisely identified the primary reported microRNA as the one binding to the found site.

The third method consists of the experimental support for novel predictions of RNA22. Luciferase-reporter assays were implemented to test binding the sites predicted by RNA22. Each predicted microRNA binding site was introduced as a single copy directly downstream of a Renilla luciferase open reading frame (ORF). Three murine microRNAs (mmu-miR-375, mmu-miR-296 and mmu-miR-134) were part of the analysis. Setting a minimum of M = 14 matching base pairs between the microRNA and a target, at most G = 1 unpaired bases in the seed region and binding energies [E = −22 Kcal/mol for the microRNA/mRNA complex], 2292, 271 and 2318 targets for miR-375, miR-296 and miR-134 respectively, were predicted. Due to the inability to test all the predictions, solely 44 predicted targets for miR-375, 24 for miR-296 and 158 for miR-134 were tested. Thus, Luciferase activity was suppressed by at least 30% for 168 out of the 226 tested predictions. Indeed, for more than half of the tested predictions, suppression ranged between 40% and 80%.

Finally, false positive rate, sensitivity and resilience in the presence of random sequences were also calculated.

2.5.3 Machine Learning Models

In the stage of Pattern Discovery, the Teiresias algorithm [111] was applied to discover variable-length motifs (‘‘patterns’’) in the mature microRNA sequences of the cleaned-up input. The aforementioned motifs incorporate a minimum of L = 4 nucleotides, contain at least 30% of their positions specified (i.e.,W = 12) and appear a minimum of K = 2 times in the processed input.

Conserved sequence segments are represented by regular expressions with varying degrees of descriptive power [112]. In this analysis, expressions with a combination of literals (solid characters from the alphabet of permitted symbols), wildcards (each denoted by ‘‘.’’ and representing any character) and sets of equivalent literals, which can occupy the corresponding position, were employed. For instance, [AT][CG].TTTTT[CG]G..[AT] is one such pattern. In particular, all instances of it contain either an A or T on their first position, a C or G on their second position, any nucleotide on their third position and a T on their fourth position. These patterns are called ‘‘rigid patterns’ due to the fact that the distance between two consecutive occupied positions remains intact across all instances of the pattern. Furthermore, a second-order Markov chain was trained in order to estimate each pattern’s statistical significance. All patterns with estimated log-probability ≥ -38 were rejected. As a result, 233,554 mature microRNA patterns abided after the implementation of this stage.

2.5.4 Results

The following results are concluded from the analysis of RNA22 algorithm.

2.5.4.1 Insulin secretion in Murinae

Validated target #2 of miR-375 is included in the 3'UTR of Kv2, a member of the voltage-dependent K+ channel family. In addition, validated target #14 is contained in the 3′UTR of a GLP-2 receptor. Both of these targets are linked to insulin secretion [113, 114].

2.5.4.2 A single microRNA can have numerous targets

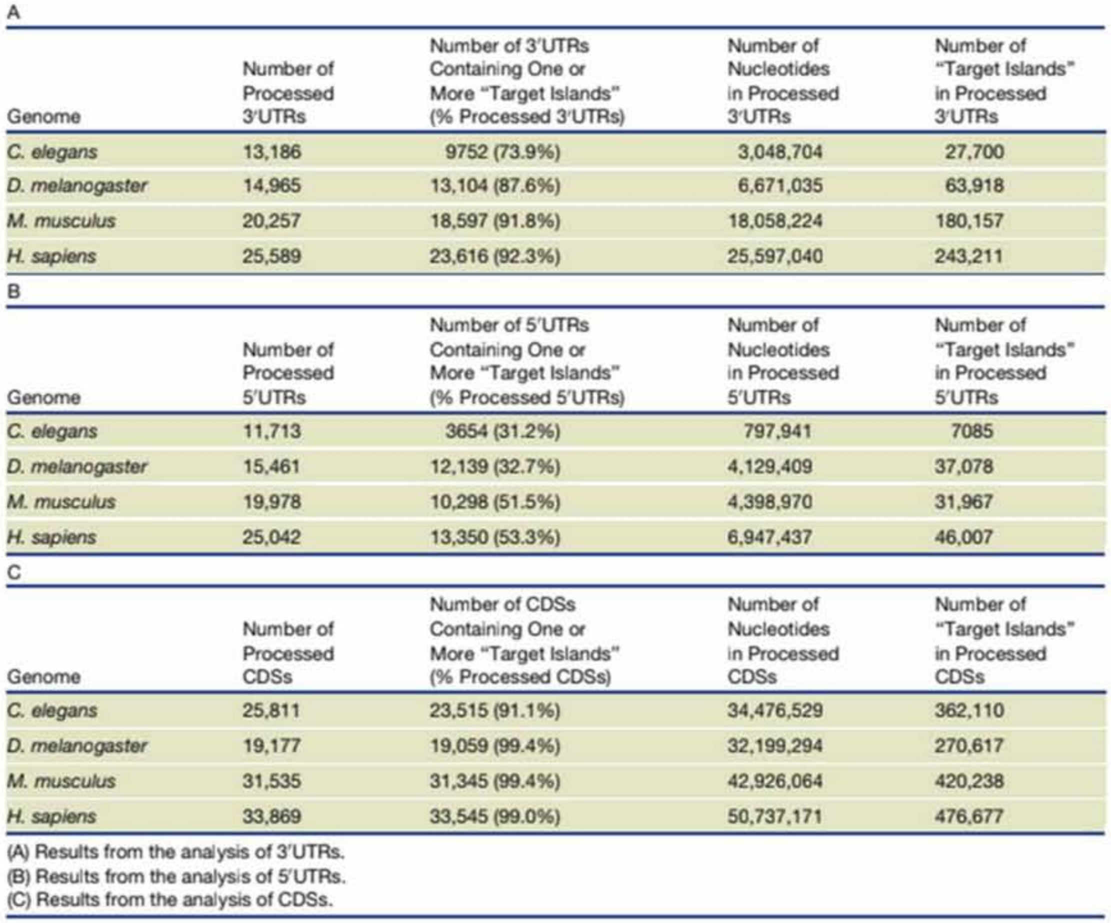
From the studies obtained for miR-134 and miR-375, it was concluded that a huge portion of the predicted miR-134 and miR-375 targets are plausibly true.

2.5.4.3 Revisiting the number and location of microRNA binding sites

Since the presented pattern recognition method is not biased in any way in favor of 3′UTRs, it was also applied to the analysis of the 5′UTRs and CDSs of the four genomes. Between 31% and 53% of the known 5′UTR sequences, they are forecasted to contain one or more targets. Concerning CDSs, almost every amino acid coding sequence was predicted to contain one or more targets. In all regions, namely 5′UTR, CDS and 3′UTR, the number of discovered target islands was roughly 1/100 of the number of examined nucleotides. Thereupon, microRNA regulation may be effected through the 5′UTRs and CDSs of gene transcripts in animals, along with 3′UTRs.

The 3′UTRs of C.elegans, D. melanogaster, M. musculus and H. sapiens were analyzed and the number of microRNA binding sites they contain, was calculated. **Table 7** aggregates the results. Indeed, according to the genome, between 74% and 92% of the known 3′UTRs one or more target islands are contained, each of which corresponds to at least one putative binding site.

Table 7. Summary of RNA22's Predictions for Four Model Genomes [34].



A

Genome	Number of Processed 3′UTRs	Number of 3′UTRs Containing One or More "Target Islands" (% Processed 3′UTRs)	Number of Nucleotides in Processed 3′UTRs	Number of "Target Islands" in Processed 3′UTRs
C. elegans	13,186	9752 (73.9%)	3,048,704	27,700
D. melanogaster	14,965	13,104 (87.6%)	6,671,035	63,918
M. musculus	20,257	18,597 (91.8%)	18,058,224	180,157
H. sapiens	25,589	23,616 (92.3%)	25,597,040	243,211

B

Genome	Number of Processed 5′UTRs	Number of 5′UTRs Containing One or More "Target Islands" (% Processed 5′UTRs)	Number of Nucleotides in Processed 5′UTRs	Number of "Target Islands" in Processed 5′UTRs
C. elegans	11,713	3654 (31.2%)	797,941	7085
D. melanogaster	15,461	12,139 (32.7%)	4,129,409	37,078
M. musculus	19,978	10,298 (51.5%)	4,398,970	31,967
H. sapiens	25,042	13,350 (53.3%)	6,947,437	46,007

C

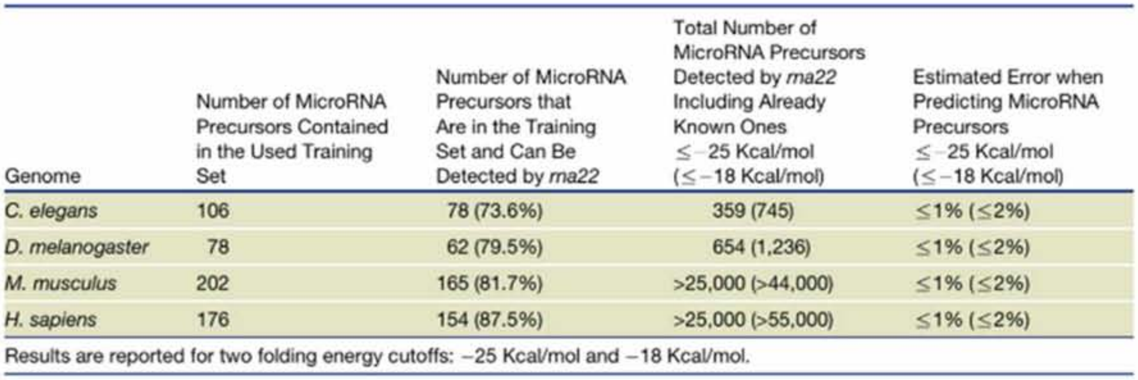
Genome	Number of Processed CDSs	Number of CDSs Containing One or More "Target Islands" (% Processed CDSs)	Number of Nucleotides in Processed CDSs	Number of "Target Islands" in Processed CDSs
C. elegans	25,811	23,515 (91.1%)	34,476,529	362,110
D. melanogaster	19,177	19,059 (99.4%)	32,199,294	270,617
M. musculus	31,535	31,345 (99.4%)	42,926,064	420,238
H. sapiens	33,869	33,545 (99.0%)	50,737,171	476,677

(A) Results from the analysis of 3′UTRs.
(B) Results from the analysis of 5′UTRs.
(C) Results from the analysis of CDSs.

2.5.4.4 Revisiting the number of microRNA precursors

According to the analysis of the four model organisms, the number of endogenously encoded microRNA precursors may in fact be substantially higher than currently hypothesized, in all four studied genomes (**Table 8**). This statement is further buttressed from the method's estimated low false-positive. In addition, for each predicted precursor, RNA22 reports the mature microRNA(s) contained therein.

Table 8. RNA22's Estimates of the Number of MicroRNA Precursors for the Worm, Fruit Fly, Mouse and Human Genomes.



Genome	Number of MicroRNA Precursors Contained in the Used Training Set	Number of MicroRNA Precursors that Are in the Training Set and Can Be Detected by *rna22*	Total Number of MicroRNA Precursors Detected by *rna22* Including Already Known Ones ≤ −25 Kcal/mol (≤ −18 Kcal/mol)	Estimated Error when Predicting MicroRNA Precursors ≤ −25 Kcal/mol (≤ −18 Kcal/mol)
C. elegans	106	78 (73.6%)	359 (745)	≤1% (≤2%)
D. melanogaster	78	62 (79.5%)	654 (1,236)	≤1% (≤2%)
M. musculus	202	165 (81.7%)	>25,000 (>44,000)	≤1% (≤2%)
H. sapiens	176	154 (87.5%)	>25,000 (>55,000)	≤1% (≤2%)

Results are reported for two folding energy cutoffs: −25 Kcal/mol and −18 Kcal/mol.

2.6 TargetRank

TargetRank [115] constitutes a method which enables improved siRNA off-target prediction, permits integrated ranking of conserved and no conserved miRNA targets as well as demonstrates that targeting by endogenous and exogenous miRNAs/siRNAs involves similar or identical determinants. Vertebrate mRNAs are frequently targeted for post-transcriptional repression by microRNAs through mechanisms involving pairing of 3' UTR seed matches to bases at the 5' end of miRNAs. Through analysis of expression array data following miRNA or siRNA overexpression or inhibition, it was found that mRNA fold change increases multiplicatively (i.e., log-additively) with seed match count and that a single 8mer seed match is more susceptible to the mediation of down-regulation comparable to two (2) 7mer seed matches. Thus, several targeting determinants that improve seed match-associated mRNA repression, along with the presence of adenosine opposite miRNA base 1 and of adenosine or uridine opposite miRNA base 9, independent of complementarity to the siRNA/miRNA, were identified. Independently, increased sequence conservation in the ~50 bases 5' and 3' of the seed match and increased AU content 3' of the seed match were each associated with burgeoned mRNA down-regulation.

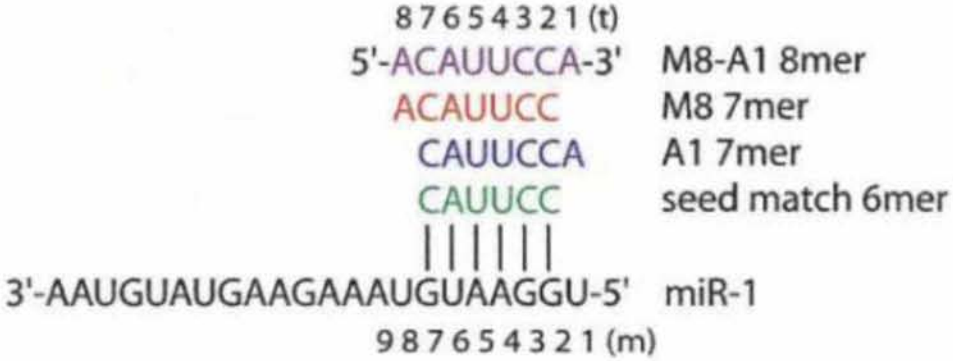
2.6.1 Datasets

2.6.1.1 3'-UTR datasets

Genome coordinates for 3' UTRs were obtained using Refseq annotations and alignments of hg17 with 16 other vertebrate genomes, available from UCSC (*http://hgdownload.cse.ucsc.edu*) for human (hg17, May 2004), mouse (mm5, May 2004) and zebrafish (danRer3, May 2005). Only Refseq transcripts mapping uniquely to the genome were considered. Annotated 3' UTRs shorter than 50 nt were excluded and solely Refseq transcripts mapping uniquely to the genome were taken into consideration.

The 3'-UTR sequences were examined for no overlapping seed matches to relevant miRNAs or siRNAs of the types, as shown in **Figure13**. As far as human and mouse analysis is concerned, multiple alignments were obtained for each 3' UTR by extracting the relevant region from genomic alignments available in multiple alignment format (MAF) from UCSC in order conservation to be assessed (*http://hgdownload.cse.ucsc.edu*, hg17 alignments of 17 vertebrate genomes). Seed matches with perfect conservation in aligned human, mouse, rat and dog (HMRD) UTRs were labeled as conserved.

Figure 13. (A) Seed match types and numbering system, illustrated for miR-1. Positions in the miRNA are numbered 5'-3'. (Seed match 6 mer) WC inverse complement of miRNA bases 2–7; (A1) presence of adenosine opposite miRNA base 1; (M8) WC match to miRNA base 8.



2.6.1.2 miRNA and siRNA transfection datasets

Microarray expression data for miR-1 and miR-124 HeLa transfection experiments [52] were obtained from GEO accession GSE2075. Array platform information for these experiments was obtained from GEO accession GPL1749. Indeed, probes were mapped to the human genome using BLAST and subsequently mapped to Refseq annotated 3' UTRs using Refseq genomic mapping files, available from UCSC (*http://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/*).

Microarray expression data for siRNA HeLa transfection experiments were obtained from *http://www.rii.com/publications/2003/nbt831.html* and from GEO accession GSE5814 [116, 117, 84] and GSE5291 [116, 117, 84]. Only values with Refseq IDs were retained. To remove poorly expressed genes, genes with log2 intensity <4.0 were excluded for both datasets.

2.6.1.3 Zebrafish embryo Dicer knockout datasets

Microarray expression data for zebrafish wild-type and MZdicer mutant embryos [88] were obtained from GEO (accession GSE4201). Probe information for the Affymetrix Gene-Chip Zebrafish Genome Array was also obtained from GEO (accession GPL1319). Probes were mapped to Refseqs using genomic mapping information for zebrafish Affymetrix Exemplar sequences from the UCSC annotation database. Only probes with a present (P) call were considered.

2.6.1.4 bic/mir-155 knockout datasets

Microarray expression data for mouse wild-type and miR-155 deficient Th1 cells [78] were obtained from ArrayExpress (accession E-TABM-232). TargetRank algorithm considered only probes mapping to mouse Refseqs.

2.6.2 Methods

2.6.2.1 Conditional Dicer knockout mice and MEFs

Male mice carrying one copy of the pCAGGCre-ER allele [118] and one Dicer floxed allele [119] were crossed to Dicer floxed/floxed females harboring also a LacZ reporter (R26R) for detection of Cre activity [120]. Timed-pregnant females were sacrificed at embryonic day 16 and embryos were dissected and dissociated to generate mouse embryonic fibroblast (MEF) primary culture [121]. After 72 h of incubation, cells were frozen in aliquots.

2.6.2.2 Cell culture and treatments

MEFs were thawed prior to experiments, grown in DMEM supplemented with 10% FCS, penicillin/streptomycin and glutamine, split once, and induced for loss of functional Dicer by addition of 4-orthohydroxy). Following 4 d (and daily change of media and drug), total RNA and total protein were extracted. Control MEFs derived from wild-type mice were subjected to the same treatments.

2.6.2.3 RNA extraction

Total RNA was extracted using TRIzol reagent (Sigma) and RNA quality was measured using an Agilent Bioanalyzer.

2.6.2.4 MEF miRNA microarray analysis

MicroRNA microarrays were printed using a Cartesian PixSys 5500 Arrayer on epoxy slides (Corning) using Ambion's miRvana amine-modified DNA oligonucleotide probe set and scanned using an Axon Scanner GenPix 4000.

2.6.2.5 Northern analysis

Thirty micrograms of total RNA was separated in 15% TBE-UREA gels (Bio-Rad), transferred to a GeneScreen Plus membrane (Perkin Elmer) using semidry electroblot apparatus (Owl) in 1× TBE (90 mM Tris-base, 2 mM EDTA, 90 mM Boric acid) at 350 mA for 35 min. Prehybridization and hybridization were carried out in PerfectHyb Plus Hybridization Buffer (Sigma) supplemented with Salmon Sperm DNA (20 µg/mL) for 2 and 16 h, respectively, at 42°C, with a radiolabeled probe added to the latter. Washes were done in 2× SSC + 0.2% SDS (twice), then 1× SSC + 0.2% SDS (once) for 5 min at 50°C. Membranes were exposed to a PhosphorImager cassette for 3 d, then scanned (PhosphorImager, Molecular Dynamics, 445 SI) and quantitated (ImageQuant, Molecular Dynamics).

2.6.2.6 MEF mRNA microarrays

Affymetrix GeneChip Mouse Genome 430_2 Array labeling, hybridization, and scanning were performed. The data were deposited in NCBI's Gene Expression Omnibus (GEO, *http://www.ncbi.nlm.nih.gov/geo/*) and were accessible through GEO Series accession number GSE6046. To map probes on the Affymetrix Mouse 430_2 array to Refseq transcripts, custom CDF file MM430_MM_REFSEQ_6 were downloaded from *http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF*, the custom CDF project site. Refseq transcript expression levels were then calculated using GCRMA (GCRMA package, Bioconductor in R environment) using default settings. Genes with normalized log2 intensity below 3 were excluded from the analysis.

2.6.3 TargetRank Scoring System

TargetRank scores the seed matches in a UTR relative to a given siRNA or miRNA and then calculates an overall score for the mRNA as a whole by summing the scores for all seed matches present in the 3' UTR. The score for each seed match is based on

- its seed match type
- the base composition at position t9
- flanking AU content (of the 50 nt immediately 3' of the seed match)
- flanking conservation (of the 50 nt immediately 5' of the seed match)

In cases where multiple input siRNA/miRNA sequences are provided and the seed match type is ambiguous (i.e. the same 3' UTR sequence can be interpreted as a different seed match type depending on which input siRNA/miRNA is considered), then the TargetRank score for each possible seed match type is calculated separately and the final score for the seed match is the average of the possible scores. Only targets with scores above 0.2 are reported. Also, the relative ranking given by TargetRank is probably more useful than the score itself, since the overall magnitude of miRNA- or siRNA-associated repression will vary in different systems.

2.6.4 Software - Statistical analyses

All test statistics were estimated using R (*http://www.r-project.org*). The Wilcoxon rank sum test was preferable compared to the t-test because it does not assume normality of the underlying distributions, and because it is more intuitive and familiar than nonparametric alternatives such as the Kolmogorov–Smirnov (KS) test. Moreover, a P-value cutoff of 0.05 was employed for all analyses.

2.6.5 Results

The analysis of the effects on global mRNA expression in miRNA and siRNA overexpression studies as well as the outcomes on mRNA expression of endogenous mouse miRNAs, analyzed following knockout of the Dicer1 gene, led to novel rules and determinants for targeting in both endogenously expressed miRNAs and exogenous miRNAs and siRNAs. These rules include:

a) An hierarchy of extended seed match types associated with different degrees of target down-regulation
b) Seed match hierarchy supported by siRNA, comparative genomic and luciferase data

c) Evidence for direct recognition of t1 adenosines by the silencing complex
d) Stronger down-regulation of mRNAs with conserved seed matches
e) Inducible inhibition of endogenous miRNA expression in mouse embryonic fibroblasts
f) Targeting rules inferred from derepression of mRNAs following Dicer knockout
g) Fold change increases multiplicatively with seed match count for both endogenous miRNAs and exogenous miRNAs/siRNAs
h) Evidence for A or U at position t9 as a targeting determinant
i) Increased conservation and AU content flanking siRNA seed matches associated with increased mRNA repression

2.7 mirSVR

mirSVR [33] algorithm contributes to the prediction of the likelihood of target mRNA down-regulation from sequence and structure features in microRNA/mRNA predicted target sites. It uses miRanda algorithm version 2.0 with score cutoff (-sc) of 120, gap opening and gap extension (-go -ge) of -9 and -4 respectively in order to obtain candidate predicted target sites followed by a support vector regression to score them. Each target site is represented by a feature vector that incorporates information on microRNA-mRNA interactions such as site accessibility, AU flanking content, position of the target site within the 3′ UTR, UTR length and conservation. Indeed, accessibility scores were computed with the aid of RNAplfold [122] with the following parameters: w = 80, L = 40 and u = 8 on a window of 160 bases around the target site and phastCons scores [123] were calculated for measuring the conservation of nucleotide positions across multiple vertebrates. The program also generates pre-computed results which are available online and has the potential to identify a significant number of experimentally determined non-canonical and non-conserved sites.

After scrutiny of the results delivered by mirSVR, the following parameters play an important role in its function:

1. mirSVR score: This score is an estimate of the miRNA effect on the mRNA expression level and is based on sequence, context, accessibility, conservation and UTR relevant features. The more negative the score, the greater the effect is.
   a. mirSVR score< 0
2. PhastCons score: 0<PhastCons score< 1

2.7.1 Training data

MirSVR is trained on a set of 9 microRNA transfection experiments from Grimson et al [23] (GEO: GSE8501) dataset including expression arrays from HeLa cells transfected by miR-122a, miR-128a, miR-132, miR-133a, miR-142, miR-148b, miR-181a, miR-

7and miR-9. In transfection experiments, only signal intensities above median are considered. During training, two (2) different models were created : “Canonical-only” mode incorporates training genes that contain a single canonical site in the 3’ UTR while “all-sites” mode incorporates training on genes that contain a single canonical or non-canonical site in the 3’ UTR, enabling non-canonical sites with exactly one G:U wobble or mismatch in the 6-mer seed region.

2.7.2 Test data

MirSVR is tested on 17 independent microRNA transfection experiments using mRNA expression profiling from Linsley et al. [80] data set [GEO:GSE6838] which involves expression data from let-7c, miR-103, miR-106b, miR-141, miR-15a, miR-16, miR-17-5p, miR-192,miR-20, miR-200a and miR-215 all measured after 24 hours. In addition, the algorithm is tested on 5 microRNA transfection experiments using protein expression data from Selbach et al. data set [56] (let-7b,miR-155, miR-16, miR-1, and miR-30a). Changes in protein expression are calculated as the median of the log2 expression changes of its measured peptides between transfection and control experiments. Only proteins with unique peptide count ≥ 10 contribute to unique protein identification.

3 microRNA inhibition experiments in conjunction with mRNA expression profiling are also tested. The datasets consist of miR-106b 2′-O-methyl inhibition, A172 glioma cells treated with anti-miR-21 and LNA inhibition of miR-122 and are derived from GEO: GSM155605] [80], GEO:GSM298113] [124] and [125]. Moreover, the performance of MirSVR is tested on microarray data from AGO IP experiments. The generation of a test set of mRNA targets for endogenously expressed microRNAs includes the use of the Landthaler et al. dataset [126] that incorporates miRNAs from seed families (hsa-miR-16, hsa-miR-30e-5p, hsa-miR-19b, hsa-miR-32, hsa-miR-20a and hsa-miR-21) and searches for their target sites in genes that are enriched in the immunoprecipitates (IP) of the four AGO1-4 proteins in HEK293 cells. Microarray data from the IP experiments can be found in [127], their normalization is implemented with the GCRMA R package and log enrichment values is done with the limma package.

Finally, CLIP data provided by the authors of the publication is used. The sequence traces identified by the CLIP method predict non-canonical target sites given that they match 3′ UTRs regions, while the ones that match coding regions or 5′UTRs are eliminated. Non-canonical sites that overlap with the CLIP-bound sequences in 3′ UTR are labeled as true sites and all other non-canonical candidate sites for the same miRNAs are predicted as false predictions. As a result, a data set of 4,692 negative sites and 883 positive sites for 54 microRNAs is generated.

2.7.3 Results

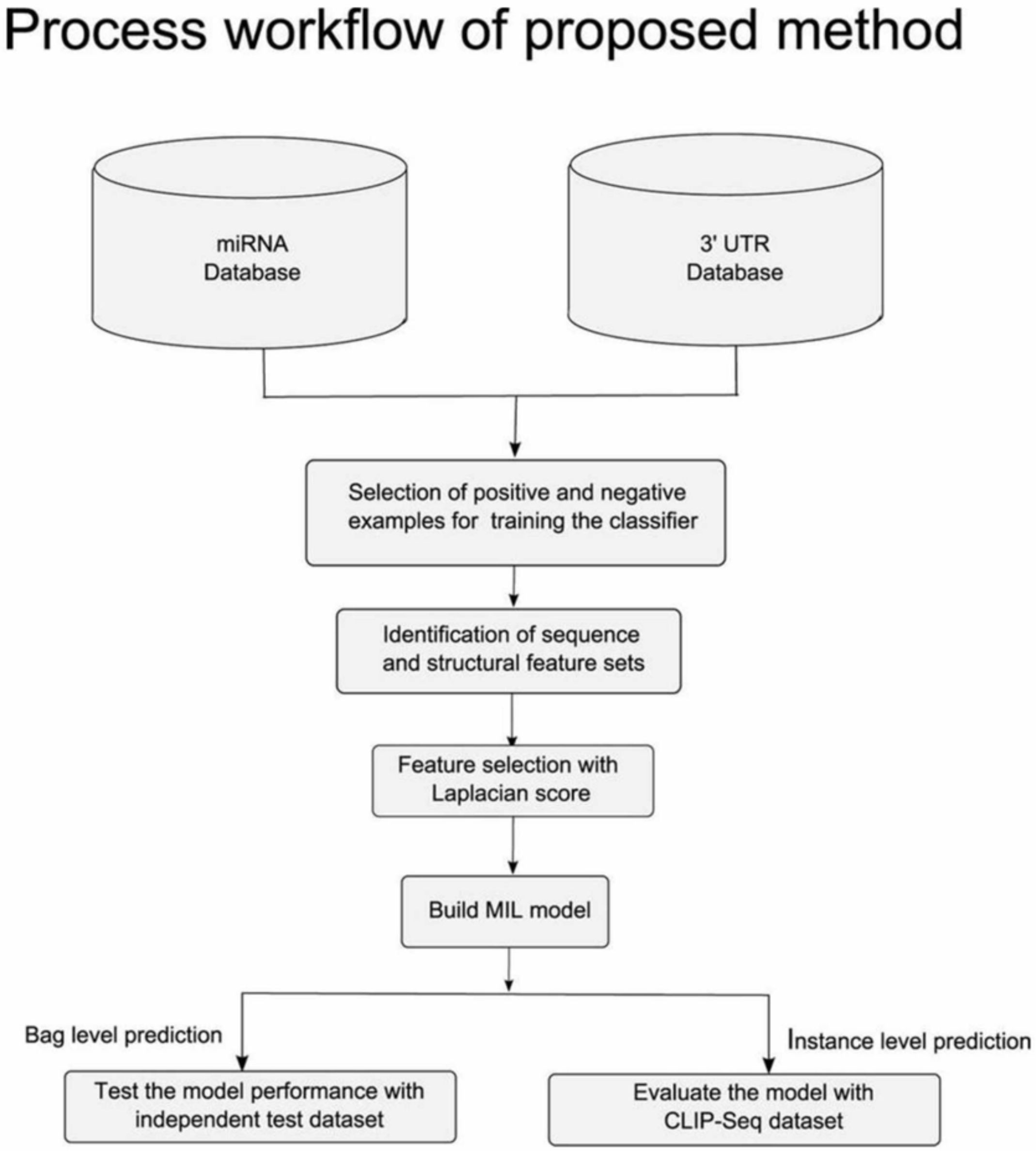
mirSVR is a score-based method, which is independent of seed classification and can be compared with other miRNA target prediction methods when tested on mRNA and protein expression changes as it accurately predicts target site efficiency. Moreover, mirSVR scores ameliorate ranking of canonical sites and identify genes with functional non-canonical sites taking into account data from microRNA transfections and CLIP experiments. The model predicts genes regulated by multiple endogenous microRNAs and considers site conservation as a feature rather than a filter. Finally, the algorithm incorporates a variety of microRNA::site duplex and contextual features and is created so that it avoids overfitting and scores multiple sites.

2.8 MBSTAR

MBSTAR (Multiple instance learning of Binding Sites of miRNA TARgets) [128] algorithm accurately predicts true or functional miRNA binding sites. Due to the fact that the actual binding sites in the target mRNAs are unknown, multiple instance learning is adopted as an approach. In MBSTAR, due to the fact that six different multiple instance learning frameworks are taken into account, a random forest model achieves the highest accuracy within the training set. From a compendium of 31 structural and 340 sequence extracted features, the 40 most relevant features are selected in order the classifier model to be built.

The process flowchart of the complete method of MBSTAR is described in the following **Figure 14**. Firstly, human 3′-UTR genome sequence and 2,042 mature miRNA sequences are extracted. Biologically verified positive examples of miRNA-mRNA pairs are collected as well as non-target examples are taken from previous work [129]. In addition, sequence and structural features are extracted from PBSs of miRNA and transcript pairs. Laplacian unsupervised feature selection is utilized to rank the features on their importance and the top 40 features are selected to train the classifier.

Figure 14. Process flowchart of the proposed MBSTAR.


Process workflow of proposed method
miRNA Database
3' UTR Database
Selection of positive and negative examples for training the classifier
Identification of sequence and structural feature sets
Feature selection with Laplacian score
Build MIL model
Bag level prediction
Instance level prediction
Test the model performance with independent test dataset
Evaluate the model with CLIP-Seq dataset

2.8.1 Training data

In MBSTAR, miRNA-mRNA pairs, corresponding to mature miRNA sequences and 3′UTRs of mRNAs, are considered as training data. The 3'UTR sequence of human assembly hg19 is extracted from UCSC Genome Browser [130], while 2,042 mature human miRNA sequences are extracted from miRBase database [96].

To train the random forest classifier 286 negative (non-target) examples are taken as described in [129]. In particular, 2 data sets containing expressions of both mRNA and miRNA in the same tissue are considered. The pairs that are over-expressed or under-expressed in the same tissue are extracted as potential negative examples, while those that show poor interactions in terms of interaction energy score (> 0 K cal/mol) are also eliminated. Finally, all those pairs that have a high conservation score (≥ 0.5) are removed. Therefore, a set of miRNA-mRNA pairs that are unlikely to be targets, is formed. As far as experimentally validated positive examples are concerned, 286 miRNA and validated transcript pairs from the miRecords database are taken. It is essential to note that the same number of positive and negative examples is considered so that a balanced classifier model is devised.

For each miRNA-mRNA pair, the complementary matching sites in the 3′-UTR of the mRNA, corresponding to the seed site of the miRNA, are first identified. These are called PBSs. The algorithm considers single G-U wobble pair while finding the PBSs. PBSs that are positioned neither too close (≤ 15 nt) to the stop codon nor near the middle of the 39-UTR, are taken as instances in MBSTAR. In addition, feature extraction is carried out from the PBSs and their neighboring regions.

2.8.2 Feature extraction

Regions surrounding a PBS play a significant role in determining binding site accessibility of miRNA. Based on this observation, 630 nucleotides flanking regions around a PBS were taken into consideration to extract 371 features. These features consist of both sequence as well as structural ones. The sequence features comprise single, di-, tri- and quad-nucleotide frequencies from the flanking regions of the PBS. Vienna RNA package version 2.0.723 was used to calculate the duplex structure and estimate features such as (a) internal-loops or interior loops which are found in RNA if non Watson-Crick base pairing between the nucleotides separates the double stranded RNA, (b) bulge loop which is a single stranded region connecting two adjacent base-paired segments in shape of a ‘‘bubble’’ in the middle of a double helix on one side, (c) hairpin loop which is a structure with two ends of a single-stranded region (loop) connecting a base-paired region (stem) and (d) multibranch loop which is a loop that brings three or more base-paired segments in close vicinity forming a multi-furcated structure.

Furthermore, the minimum free energy was another feature, calculated using RNAcofold program for the entire flanking region including the seed matching site.

2.8.3 Feature selection

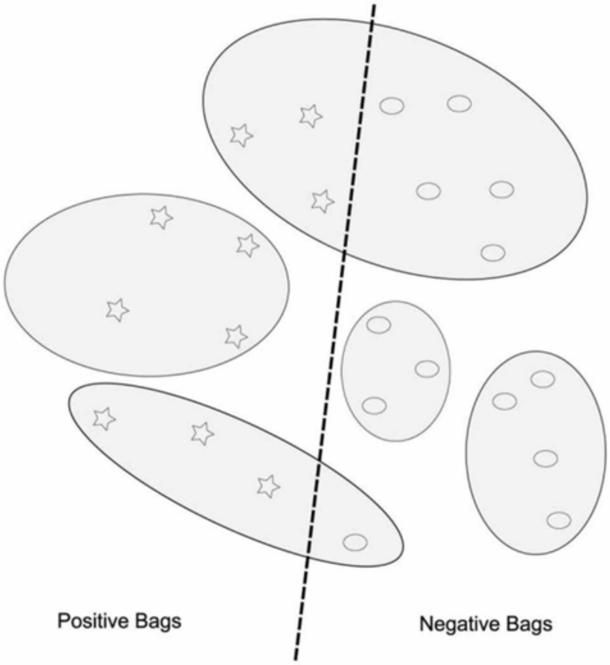
A total of 371 features were extracted for each PBS in both target and non-targets. Feature selection is utilized to extract noisy and redundant features from the extracted feature set for better classification accuracy. Laplacian score based feature selection (LSFS), unsupervised discriminative feature selection (UDFS) and multiclass feature selection (MCFS) techniques were developed to distinguish the best set of features within 5 fold cross validation. All the MIL algorithms were trained on selected features in each fold and the method with the most robust accuracy was selected. It was found that LSFS outperformed the other two methods and with the aid of Laplacian score, the top 40 features from training data, were selected. It is important to mention that Laplacian score depends on the fact that data belonging to the same class are often close to each other.

2.8.4 Machine Learning Models-Training an MIL random forests classifier

With the aid of the aforementioned selected features, an MIL random forests (MIL-RF) classifier with 50 trees was used. The top 40 features, according to the Laplacian scores, out of the 371 initial features were utilized to train the classifier. The cooling parameter of deterministic annealing was set to -0.25. As a result, Hinge loss function and bagging with refine sampling were most likely to make a balanced prediction result with sensitivity 0.755 and specificity 0.685 using 5-fold cross validation.

Multiple instance learning (MIL) constitutes the fourth learning paradigm following supervised, unsupervised and reinforcement learning. A graphical representation of MIL problem is depicted in **Figure 15**. The ellipsoids denote the individual bags, whereas the star and the small ellipsoids represent positive and negative instances respectively. The hyperplane, which divides the instances, is illustrated by a dotted line.

Figure 15. Classification of positive and negative instances by multiple instance learning methodology, when only the bag label is known.



In addition, Diverse Density (DD), Expectation-Maximization DD (EM-DD), Citation kNN and two variations of multiple instance SVM (MI-SVM) classifiers with the same dataset were trained and performed a 5-fold cross validation. Ten different sets seed points for DD and EM-DD algorithm were used and the results were aggregated. As far as Citation kNN is concerned, both Euclidean and cosine distance were measured with varying the values of reference and citers from 1 to 10. It was found that the Euclidean distance with reference 2 and citation 4 presented the best result. For both the SVM variants, linear polynomial and Radial Basis Function (RBF) kernels were used, varying the respective parameters, with RBF kernel with gamma value 0.05 giving the most accurate result. The results are reported in the following **Table 9**. It is observed that the MIL-RF provides the highest accuracy among all MIL classifiers for the given dataset. Citation kNN achieves second highest accuracy and is able to beat SVM based approaches with a high margin.

Table 9. Comparative study of 5-fold cross validation accuracies for different MIL frameworks.

Method	Bag level accuracy
MIL-RF	0.7202
Citation kNN	0.6854
Expectation-Maximization Diverse Density	0.6644
MI-SVM	0.5871
mi-SVM	0.5316
Diverse Density	0.4861

2.8.5 Test data

All biologically verified positive interactions were derived from TarBase 6.0 database [131]. After converting genes to corresponding reference sequence identifiers for NCBI standard, a total of 31,456 unique positive interactions were obtained. Experimental methodologies, such as reporter gene assay, western blotting, northern blotting, microarray analysis, proteomics (such as pSILAC), sequencing (RNA-Seq, HITS-CLIP, PAR-CLIP), qPCR and others (ELISA, RACE, immunohistochemistry, etc.), have been used in order to verify the aforementioned interactions. These unique interactions contained a total of 145 miRNAs and 16,944 mRNAs. This dataset was compared with PAR-CLIP cluster data of the human genome downloaded from starBase [132].

Moreover, this analysis includes Biological complexity (BC) greater than or equal to one. BC of an experiment is a measure of reproducibility between biological experiments. Indeed, in one experiment, at least, the PAR-CLIP cluster is targeted by miRNAs. Due to the fact that for canonical seed matching to occur, at least a 6-mer site, including a possible single G-U wobble pair should be present in the PAR-CLIP cluster, all clusters corresponding to TarBase 6.0 positive dataset were isolated. Furthermore, the clusters which contain at least one 6-mer site, corresponding to high confidence miRNA-mRNA pairs and obtained from TarBase 6.0 database, were chosen. As a result, 16,824 clusters corresponding to 121 miRNAs and 5120 mRNAs were selected. These clusters could be mapped to 9,582 miRNA-mRNA interactions. Any common interactions among the data were removed in the construction of the model and 9,531 interactions and 16,681 clusters were obtained. These 9,531 positive miRNA-mRNA interactions (which had no overlap with the training data) were used as 9,531 bags for independent testing. Also, 973 non-target pairs of miRNA-mRNA from TarBase 6.0 database were extracted.

2.8.6 Results

It is found that MBSTAR achieves the highest number of overlapping binding sites with PAR-CLIP with maximum F-Score of 0.337. Compared to the other methods, MBSTAR also predicts target mRNAs with highest accuracy of 78.24% for the validated positive interactions. Another analysis on biological complexity and number of T → C conversion shows that MBSTAR is able to predict many more relevant binding sites compared to other methods.

Based on its performance, it can be concluded that MBSTAR will play a fundamental role on future laboratory experiments for obtaining functional miRNA binding sites. Apart from the evaluation of binding sites in 3’-UTRs and canonical binding sites, binding sites in the coding regions ought to be investigated. In addition, due to the inaccuracy of PAR-CLIP data and the inability of clusters to identify all of the experimentally verified results, research on the exact set of miRNAs binding to a particular cluster, is planned to occur.

2.9 SVMicrO

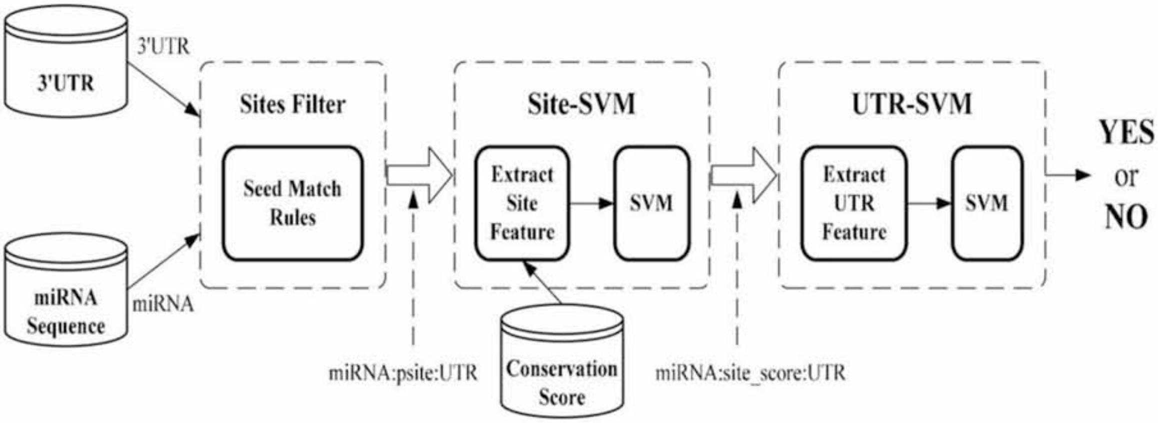
SVMicrO [133] is a machine-learning approach appropriate for mammalian miRNAs that incorporates a 2-stage structure that consists of a site support vector machine (SVM) followed by a UTR-SVM. It considers a set of 113 site and 30 UTR features, selecting as best predictors of miRNA regulation, 21 optimal site features and 18 optimal UTR features during training such as seed match, conservation, free energy, site accessibility and target-site abundance. SVMicrO utilizes these features to predict candidate miRNA:mRNA target pairs. The performance of the algorithm has been implemented on the training data, proteomics data, and immunoprecipitation (IP) pull-down data. Finally, when compared with other available miRNA target prediction programs, SVMicrO appears to enhance specificity, sensitivity and precision in forecasting miRNA targets.

Firstly, the 3'UTR sequence of miRNAs is scanned in order potential binding sites to be identified. This measure improves the efficiency of the algorithm. In particular, it is found that more than 20% true miRNA-site and 20% true miRNA-target pairs do not contain a 6-mer seed match. Therefore, a relatively looser seed match rule should be applied to gain higher sensitivity and the number of false positives to be reduced. After experimentation and investigation, it is concluded that the application of the following 5 seed match rules on regions of the 3'UTR sequence identify the latter as potential sites, achieve near 96% sensitivity on both experiment validated targets and sites as well as succeed less false positive sites:

1. There are more than 4 consecutive W-C matches
2. There are more than 5 consecutive matches (including G:U pair) and more than 2 consecutive W-C matches
3. There are more than 6 matches in total and 3 consecutive W-C matches; no gap allowed
4. 2~4 nucleotides of miRNA are W-C match, there is more than 3 W-C matches and more than 4 matches in total; no gap allowed
5. There are more than 5 matches and 5 W-C matches, and only one gap is allowed on either miRNA sequence or 3'UTR

In addition, the identified sites are entered in the Site-SVM classifier. Consequently, features from each site are extracted and a score is assigned revealing the prediction confidence of the site as a true site. In the end, the scores of the sites along with other UTR features are taken into account by the UTR-SVM classifier in order the final prediction of the UTR, as a target, to be generated.

Figure 16. The block diagram of SVMicrO. SVMicrO includes three steps. First, a site filter is applied to find the potential binding sites of the probing miRNA. Second step, Site-SVM extracts features from each potential site and assigns a score to indicate the prediction confidence of the site as a true site. Final step, the site scores, together with other UTR features are considered by the UTR-SVM to produce the final prediction of the UTR as a target.



2.9.1 Training data

2.9.1.1 Positive data

Positive data are obtained from the most up-to-date miRNA target depository called miRecords. They focus on the records of human (1,020 records), mouse (166 records), and rat (133 records). Thus, 324 miRNA-site pairs are taken from 187 miRNA-target pairs and 709 additional miRNA-target pairs are also extracted but without site information.

2.9.1.2 Negative data

Negative data result from 20 microarray data each produced by over-expressing a different miRNA, taken from NCBI Gene Expression Omnibus [69]. High quality negative data are obtained by considering the most confident up-regulated genes by restricting the differential expression p-value, fold change and consistency of the samples over time whenever available. In the end, 3,542 negative miRNA-target pairs are generated.

2.9.2 Feature Selection

Feature selection selects the most distinctive features for site and UTR. A total of 113 site features and 30 UTR features are extracted.

2.9.2.1 Site Features

7 groups of site features describe the characteristics of target recognition within a site.

2.9.2.2 Perfect seed match features

The existence of a perfect seed match in the site.

2.9.2.3 Pair-wise binding structure features

Overall, 39 pair-wise binding structure features are extracted. These include the binding energy of seed region, the binding energy of 3' region and exact boundaries of each region. Based on the secondary structure, the following types of nucleotide matches as well as the content of each 2-mer are defined:

- W-C match
- G-U match,
- mismatch
- gap

2.9.2.4 Regional binding structure features

The 18 regional features in each binding reagion are considered:

- total number of W-C matches in miRNA
- total number of G-U matches in miRNA
- total number mismatches in miRNA
- total number of gaps in miRNA
- number of bulged structures on mRNA
- number of bulged nucleotides on mRNA

2.9.2.5 Conservation features

The conservation score of each nucleotide in the site is obtained from the phastCons28way in the UCSC [130]. Then, the average conservation scores of seed binding region, 5' context region and 3' context region are formed.

2.9.2.6 Energy features

Due to the fact that a true binding site results from a stable structure, the binding energy features of the seed region, 3' region and total region are assessed by miRNAbind. Moreover, the accessibility is studied as an energy feature.

2.9.2.7 Seed context features

Context region: the contiguous upstream and downstream sequences of the seed region.

To an end of a locally AU-rich context, two 10-nt long sequences on both ends of seed binding regions are isolated as context regions. 20 context features are obtained from the calculation of the single nucleotide and 2-mer compositions for each context regions. Moreover, the nucleotide compositions of all positions in these 2 regions are considered as another 20 context nucleotide type features.

2.9.2.8 Site location features

To examine the fact that binding sites are more frequently observed at the two ends of a 3'UTR but not too close to the stop codon, 3 features are considered:

- the distance from the potential site to stop codon
- the distance from the potential site to the nearest end of 3'UTR
- the ratio of the distance from the potential site to the nearest end over the length of 3'UTR

2.9.2.9 UTR Features

3 groups of UTR features describe the characteristics of target recognition within the 3'UTR.

2.9.2.10 Length of 3'UTR

Since a target 3'UTR includes multiple binding sites, the length of the 3' UTR affects miRNA targeting. On average, the positive targets have longer length than the negative targets.

2.9.2.11 Site density features

Since the efficacy of binding is reduced when the distances among the sites are large, the global site density feature is calculated as the ratio of the number of potential/positive binding sites over the length of 3'UTR. Moreover, a 100-nt window detects a region with the maximum number of potential/positive binding sites and these maximum numbers are considered as two (2) additional features.

2.9.2.12 Binding site score features

The Site-SVM classifier generates a score for each candidate site which is considered the prediction confidence for this site. Consequently, the higher the confidence of site predictions, the higher the possibility that an UTR is a target.

Potential sites: sites identified by the filter.

Positive sites: potential sites predicted positive by the Site-SVM (SVM score >0).

The total score of positive sites, the number of potential sites and the number of positive sites are considered as 25 features.

2.9.3 Machine Learning Models - Training of the model

5-fold cross validation is carried out to train the parameters and select features for both SVMs. In each round of cross validation, a minimal redundancy maximal relevance algorithm is implemented in order the features that best contribute to the forecast of miRNA regulation, to be determined. As a result, 21 optimal site features (**Table 10**) and 18 optimal UTR features (**Table 11**) are extracted from cross-validation. Specifically, two parameters need to be optimized:

- penalty constant C of SVM
- parameter γ of the RBF kernel

Table 10. The group of optimized site features

Features	
6mer seed match	Binding energy of seed region
Conservation score of 3' context region	Seed conservation score
Number of matches in seed region	nucleotide content at the 7th position from the 5' end of miRNA match status
7mer_A1 seed match	Context nt type of the sequence in the end of 5' context region
7mer_m8 seed match	Binding energy of total region
7mer_m1 seed match	conservation score of 5' context region
8mer_A1 seed match	Number of mismatches in seed region
Accessibility energy	nucleotide content at the 5th position from the 5' end of miRNA match status
8mer_m1 seed match	nucleotide content at the 2th position from the 5' end of miRNA match status
6th 2mer status	nucleotide content at the 12th position from the 5' end of miRNA match status
Number of matches in total region	

Table 11. The group of optimized UTR features

Features	
Top site score	Number of positive sites with 8mer_m1
Total positive score	top score with 8mer_m1
Positive site number	top score 7mer_m1
Max No. of positive sites within 100 nts	top score with 7mer_A1
Density of positive sites	top score with 6mer
Number of potential sites with 8mer_A1	top score without perfect seed
Number of positive sites with 8mer_A1	Number of potential sites with 7mer_A1

Top score with 8mer_A1	Number of postive sites with 7mer_A1
Number of potential sites with 8mer_m1	length of utr

2.9.4 Test data

Genome-wide target prediction is tested on proteomics data such as human miR-1, miR-16, miR-30a, miR-124, miR-155 and let-7b. Moreover, the same validation is carried out for miR-16, miR-30a, miR-155 and let-7b. However, due to the presence of noise in protein quantification, the prediction of miR-124 and miR-1 is further evaluated with the aid of IP pull-down data. In particular, in the experiments each miRNA is transfected in 293 cells, immunoprecipitation of the AGO2 protein is done and the gene expression of possible miRNA targets is analyzed by microarrays. Particularly, 388 and 56 highly expressed genes are treated as the true targets of miR-124 and miR-1, respectively.

2.9.5 Results

After a close investigation of the set of the optimal trained features resulted after cross-validation, it is found that 7mer and 8mer seed matches are sufficient for miRNA site recognition along with 6-mer seed match. Moreover, conservation of the 3'context region or 10 nts downstream of the seed region is more important than the seed conservation. Among energy features, accessibility energy and binding energy of seed regions are the most important features. However, accessibility energy feature is considered to be more crucial, pointing that the secondary structure of potential site influences the ability of miRNA binding. As far as UTR features are concerned, the UTR length does not affect the miRNA target prediction as proposed by the histograms of UTR length in this study. Another outlook resulting from the analysis of SVMicro is the fact that the more accurate the sites are predicted, the more accurate the UTR prediction is. Indeed, the most crucial feature in the UTR-SVM classifier appears to be the Site-SVM score, a higher value of which increases the possibility that the 3'UTR is a real target.

The validation on training data demonstrates that SVMicrO retains the largest AUC and achieves the highest True Positive Rate especially for low False Positive Rate. As a result, SVMicrO achieves improved prediction sensitivity, specificity and precision compared to available algorithms.

2.10 MirMark

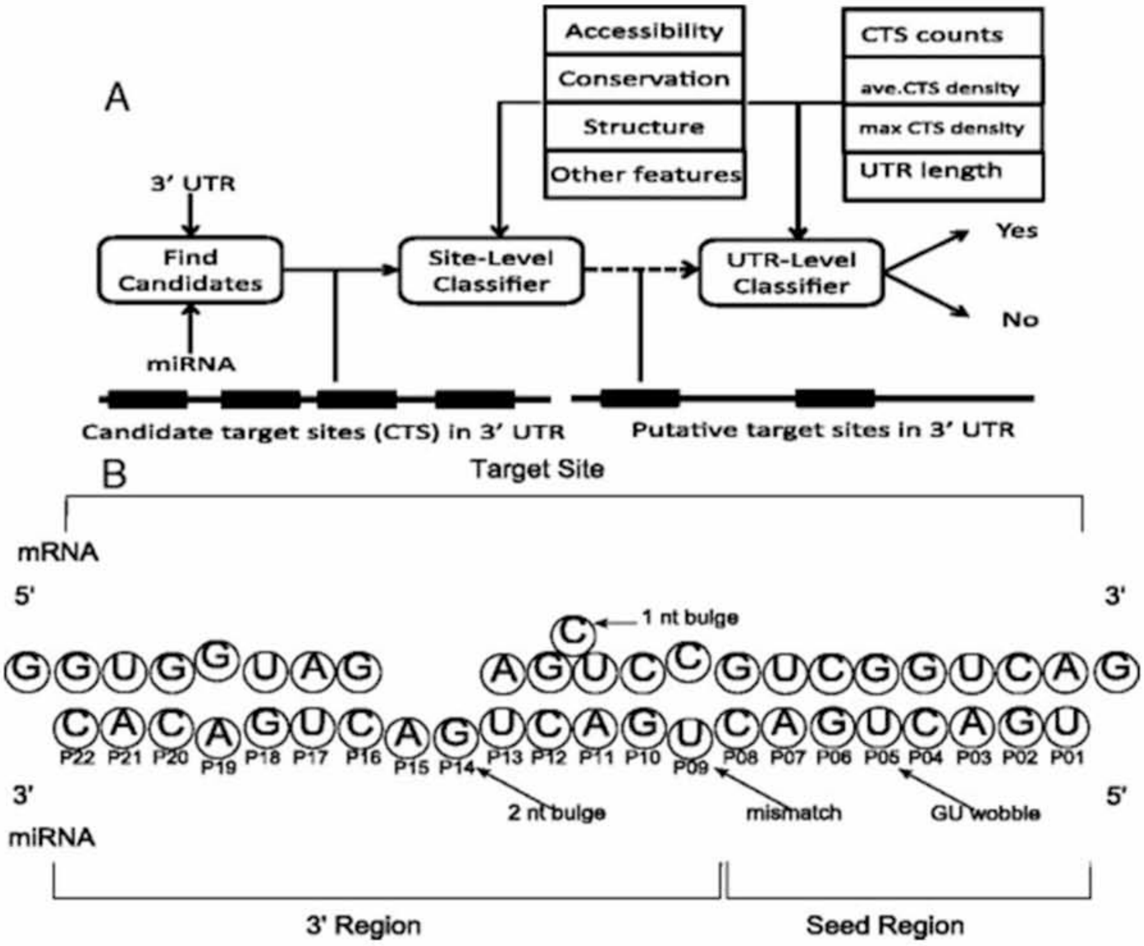
MirMark [134] predicts putative targets at both the site and UTR level by considering predictive features such as complementarity, evolutionary conservation, structural accessibility and composition as well as using the latest experimentally verified miRNA target data. It incorporates MiRanda for the initial identification of candidate miRNA binding sites. The algorithm extensively uses several statistical or machine learning methods with a random forest model to outperform all the other classifiers for target-site and miRNA-UTR interaction evaluation. MirMark's performance is assessed with the use of PAR-CLIP data.

The structure of miRNA target predictors of MirMark is illustrated in **Figure 17A**. To begin with, the identification of candidate target sites (CTSs) of the miRNA on the 3′ UTR of the mRNA takes place. CTSs are discovered using the alignment algorithm implemented in MiRanda [135]. The alignment favors, but does not require, seed matches to allow for weak seed targets such as 3′ compensatory target sites.

Having the list of CTSs of the miRNA, together with their predicted alignments (**Figure 17B**), the site-level classifier assigns a posterior probability that the given CTS is a target site of the miRNA. This prediction is made on the basis of features such as the presence of a seed match, free energy of the duplex and the accessibility of the target site.

Finally, due to the CTSs and their posterior probability of being a true target as computed by the site-level classifier, the UTR-level classifier assigns a posterior probability that the miRNA targets the mRNA overall. This prediction can be made on the basis of features such as the number of CTSs, the number of CTSs of a particular seed type and the length of the 3′ UTR. The above step allows the integration of the information, provided by the set of CTSs, improving the accuracy of the prediction (**Figure 17A**).

Figure 17. Structure of mirMark and miRNA-target region duplex. **(A)** mirMark consists of two levels of classifiers, site-level and UTR-level, depending on the type of prediction desired. First, candidate target sites (CTSs) of the miRNA on the 3′ UTR are found. The alignment of the CTSs and various other features concerning accessibility, conservation and structural information are then used by the site-level classifier to find the strongest CTSs. On the other hand, the UTR-level classifier integrates the CTSs to determine if the gene is a target of the given miRNA. **(B)** An illustration of the site-level binding between miRNA and target regions of the 3′ UTR. Information about the type of bindings that occur in the seed region is particularly predictive.



2.10.1 Training data

2.10.1.1 Positive data

The positive data appropriate for training are obtained from miRecords [136] and miRTarBase [137]. Indeed, at the site-level, only human miRNA-mRNA pairs with validated target site information are taken from miRecords. At the UTR level, experimentally validated human miRNA-gene pairs are obtained from (1) all human gene and miRNA pairs from miRecords and (2) the subset of miRNA-gene pairs that are not labelled as weakly supported from mirTarBase. Consequently, a list of 507 miRNA-target site pairs is utilized as the site-level positive set as well as a list of 2,937 miRNA-gene pairs are utilized as the UTR-level positive set.

2.10.1.2 Negative data

The negative data are produced using mock miRNAs. A mock miRNA is a random mutation of a real mature miRNA sequence that does not overlap with the seed sequences from known miRNAs. At the site level, for each real miRNA-gene pair in the positive dataset, a corresponding mock miRNA-gene pair is produced and replaces the positive miRNA in the miRNA-gene pair. At the UTR level, mock miRNA-gene pairs are generated for each real miRNA-gene pair in the UTR-level positive dataset. Overall, 520 mock miRNA-site pairs corresponding to the real miRNA-site pairs are generated using MiRanda's predicted alignments. This dataset is split and 80% is used for training and 10-fold cross-validation and the rest 20% reserved as a test set for independent evaluation.

2.10.2 Test data

From the previously discussed dataset, the rest 20% is used for independent evaluation. In addition, PAR-CLIP data are utilized for evaluation of the algorithm.

2.10.3 Feature selection

The most relevant, yet least redundant, set of features for site- and UTR-level prediction are taken into consideration. 151 site-level features are considered, which belong to the following categories:

Total minimum free energy

- Seed match type:
  - Seed_match_8mer: p1-p8 Watson-Crick (WC) match
  - Seed_match_8merA1: p1 match/mismatch to A, p2-p8 WC match
  - Seed_mach7mer1: p1-p7 WC match
  - Seed_match7mer2: p2-p8 WC match
  - Seed_match7merA1: p1 match/mismatch to A, p2-p7 WC match
  - Seed_match6mer1: p1-p6 WC match
  - Seed_match6mer2: p2-p7 WC match
  - Seed_match6mer1GU: p1-p6 WC match allowing only one GU wobble
  - Seed_match6mer2GU: p2-p7 WC match allowing only one GU wobble
- miRNA pairing: Information of the type of target duplex pairing for the first 20 nt of the miRNA in seed region, 3′ region and total miRNA region
  - number of G-C matches
  - number of A-U matches
  - number of GU wobbles
  - number of mismatches
  - number of bulges
  - number of nucleotides in bulges in the seed region of the miRNA
- Target site accessibility. The following characteristics are considered:
  - accessibility of entire seed region
  - accessibility of the 5′ half of the seed region

  - accessibility of the 3′ half of the seed region
  - position-wise accessibility of each seed position of the CTS
  - accessibility of the regions 10 nt upstream of seed region
  - accessibility of the regions 10 nt downstream of seed region
  - position-wise accessibility of the regions 10 nt upstream of seed region
  - position-wise accessibility of 10 nt downstream of seed region
- Target site composition. The following characteristics are considered:
  - nucleotide and dimer composition of CTS
  - flanking 70 nt regions upstream and downstream of CTS
  - flanking AU score, which is a weighted count of AU composition flanking the seed region

Target site conservation (Per base conservation scores of human 3′ UTRs are obtained from PhastCons46way [51]). The following characteristics are considered:

- average per base conservation score of the seed region of CTS
- average per base conservation score of the entire CTS
- average per base conservation score of 70 nt upstream and downstream flanks of the CTS

Location of target site

- Computation of distance of the CTS to the closest 3′ UTR end point/ length of the 3′ UTR for scale

From the above features, 12 site-level features are selected by feature selection, performed on the training set and are summarized in **Table 12**.

Table 12. Selected site-level features by correlation-based feature selection.

Feature	Description
miR_match_P01	Match status of miRNA position 1
miR_match_P03	Match status of miRNA position 3
miR_match_P04	Match status of miRNA position 4
miR_match_P08	Match status of miRNA position 8
miR_match_P15	Match status of miRNA position 15
Seed_bulge	Number of bulges in seed region
Total_AU	Number of AU matches in target site
Total_mismatch	Number of mismatches in target site
Total_bulge	Number of bulges in target site
Total_bulge_nt	Number of nucleotides within bulges in target site
Seed_P01_acc	Accessibility score position 1 of seed region
Seed_cons_score	Conservation score of seed region

624 UTR-level features are considered which belong to the following categories:

- Summary of site-level features
  - Calculation of total, minimum, maximum and mean values of the 151 site-level features of the CTSs of a miRNA-gene pair
  - Total, minimum, maximum and mean values of the posterior probability from the random forest-based site-level classifier
  - MiRanda alignment score
  - CTS start and end positions
- Length of the 3′ UTR
- Number of CTSs for a miRNA-gene pair
- CTS density is calculated as a) number_sites/UTR_length or b) counting the maximum number of CTSs that lie within 100 nt of each other

From the above features, 15 UTR-level features are selected by feature selection performed on the training set and are summarized in **Table 13**.

Table 13. Selected UTR-level features by correlation-based feature selection.

Feature	Description
Miranda_score.max	Maximum alignment score between miRNA and target sites
Seed_match_6mer2.mean	Proportion of target sites with P02-P07 WC match
miR_match_P01.min	Match status of miRNA position 1
Seed_match_7mer2.max	Proportion of target sites with P02-P08 WC match
Seed_match_7mer1.mean	Proportion of target sites with P01-P07 WC match
Seed_MFE.min	Minimum MFE of seed region of miRNA:site duplexes
X3p_MFE.mean Target_UC_comp.mean	Mean MFE of 3′ region of miRNA:site duplexes UC dimer composition of the CTS
miR_match_P09.mean	Match status of miRNA position 9
miR_match_P02.min	Match status of miRNA position 2
Seed_GU.mean	Mean number of GU matches in target site seed regions
miR_match_P07.mean	Match status of miRNA position 7
Start_position.min	Minimum distance of target sites to the 5′ end of the 3′ UTR

miR_match_P19.min	Match status of miRNA position 19
miR_match_P15.min	Match status of miRNA position 15

2.10.4 Software

In this study, RNAduplex, RNAfold and RNAplfold [138] in Vienna RNA package are used for energy and accessibility calculations. Nucleotide composition is computed using BioPerl [139]. Weka 3 data mining software [140] and entropy package [141] in R are used for features selection, classifier training and evaluation.

2.10.5 Results

MirMark selects the most relevant and minimally redundant features from a set of over 700 features and exhibits a significantly improved predictive performance at both the site and UTR levels. In comparison with existing available tools for human miRNA target prediction, MirMark is advanced in terms of evaluation, using the random forest classification method.

2.11 ChimiRic (miRNA target site detection: CLIP-seq analysis)

ChimiRic [142] is a model for miRNA target prediction which uses discriminative learning on transcriptome-wide AGO CLIP [79, 61] and CLASH [63] profiles. The algorithm predicts biochemical miRNA-target site interactions, instead of the extent of regulation, in order to increase the sensitivity of miRNA target prediction. In fact, chimiRic provides more accurate predictions than state-of-the-art methods based on indirect measurements. In addition, novel features of miRNA-mRNA interactions such as potential collaboration with specific RNA-binding proteins are incorporated in the model.

2.11.1 Training data

The training set of miRNA-mRNA interactions includes many non-canonical pairings captured by chimeric reads in the CLASH data, which is combined with canonical AGO binding sites identified by CLIP. In particular, the training of the duplex model uses CLASH data in HEK293 cells [63], Argonaute PAR-CLIP data in HEK293 cells [61], focusing on the top 59 expressed miRNAs in 21 miRNA seed families, HITS-CLIP in mouse CD4+ T cells [77] and HITS-CLIP in HeLa cells [55, 143]. In all the above datasets, interactions are retained according to the following criteria:

- the binding sites are located in the 3'UTR
- the binding sites complementary match to the 6-mer seed of the corresponding miRNA with edit distance of 0 or 1
- interactions are advocated by non-chimeric reads

In the training of the AGO binding model, the corresponding examples are 3'UTR sites that match to the 6-mer seed of one of the highly expressed miRNAs and are not bound. For instance, in human HEK293 PAR-CLIP cells, 59 miRNAs from 21 miRNA families are considered and in mouse CD4+ T cells, 58 miRNAs from 24 miRNA families are considered.

2.11.1.1 Positive data

The duplex model contains 1,727 (miRNA, site) pairs supported by chimeric reads CLASH data, of which 1,228 are non-canonical sites and 11,211 are AGO CLIP binding sites, containing a 6-mer seed match (or longer seed) for each highly expressed miRNA and interacting with the corresponding miRNAs. As far as the AGO binding model is concerned, the overlapping of a seed match with an Argonaute binding site in the CLIP data is categorized as a positive example for the analogous miRNA. Otherwise, if a seed overlapped with no Argonaute CLIP reads, it is consireded to be a negative example.

2.11.1.2 Negative data

The duplex model randomly selects a subset of negative examples from a total of 25,411 (miRNA, site) pairs for the training procedure. Negative examples constitute sites that are paired with another miRNA from the same seed family based on CLASH chimeric read data or canonical miRNA seed matches with no AGO CLIP confirmation and are unlike to interact with the miRNAs. As far as the AGO binding model is concerned, the separation of a seed match with Argonaute CLIP data, infers that the example is negative.

2.11.2 Test data

iPAR-CLIP data in C. elegans [144] and CLEAR-CLIP data in mouse brain [62] are used for testing the performance of the algorithm. What is more, as test data are also used array data, which involve the change in the expression of genes in eight individual miRNA transfection experiments in HCT116 cells (miR-15a, miR-16, miR-215, miR-17, miR-20a, let-7c, miR-106b and miR-103a with their corresponding GEO data sets GSM156545, GSM156546, GSM156548, GSM156553, GSM156554, GSM156557, GSM156576 and GSM156580).

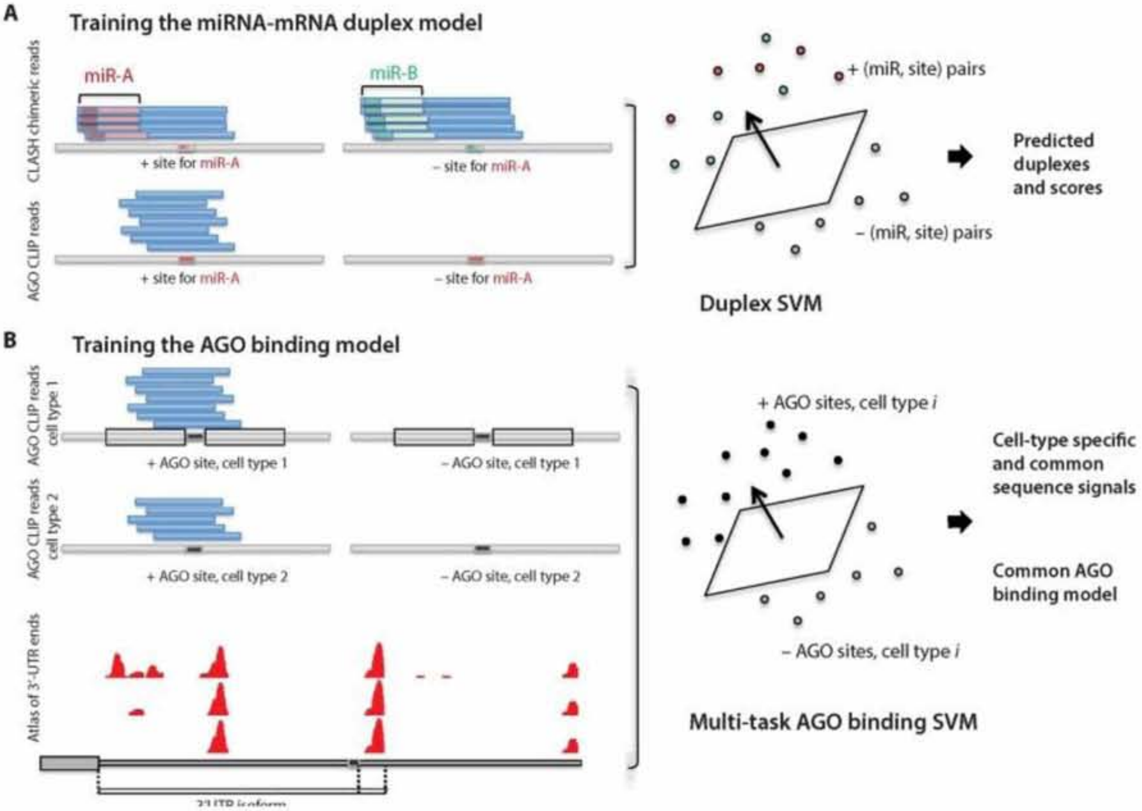
2.11.2.1 Positive test data

From the training data, all HEK293 CLASH interactions for a single miRNA seed family are extracted.

2.11.2.2 Negative test data

Targets sites interact with other miRNAs based on chimeric reads.

Figure 18. Outline of the chimiRic prediction model. **(A)** The first segment of the chimiRic model is the duplex SVM, which is trained to predict and score miRNA-mRNA duplex alignments from CLASH and CLIP-seq data. The optimal alignments of the training examples (miRNA, site pairs) are generated through an iterative training process (right). **(B)** The second segment of chimiRic is the AGO binding SVM, which utilizes the positional bias of the AGO binding sites and the local positional k-mer sequence features. Mouse and human ApA atlases based on 3' end sequencing data (bottom) give the coordinates of 3' ends, essential in this study.



2.11.3 Feature selection

The representation of features in this study is similar with that of MIRZA [99]. Specifically, the following features are contained in the description of structures between miRNA-mRNA interactions:

- the type of base pair (GU, UG, AU, UA, GC, CG) at each position in the alignment
- the bases where a loop is opened, symmetrically extended or asymmetrically extended in the duplex structure
- binary variables for each position in the miRNA sequence representing if it is paired to an mRNA base or not.
- the first base in the miRNA is paired with an Adenine in the mRNA sequence

mRNA sites contain two types of UTR features:

- local sequence context
- global positional context

The sequence context is depicted by positional k-mer features (k = 1, …, 6) from 30 nt sequences upstream and downstream of the miRNA seed match. For each site, the following positional context features are calculated:

- the distance to the nearest stop codon
- the distance to the next end of a 3'UTR isoform
- the distance to the previous end of a 3'UTR isoform. The re-normalization is achieved by a radial basis kernel.

2.11.4 Training and testing duplex and context models

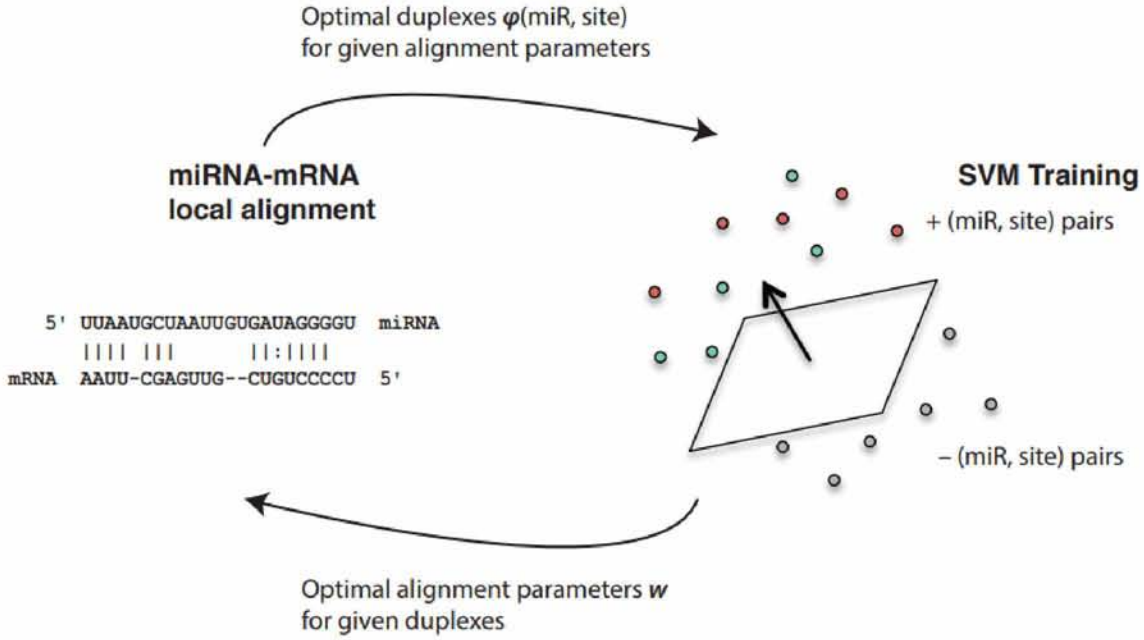
An innovative miRNA target prediction model is trained on CLASH and on AGO CLIP interactions based on 6-mer seed complementarity. The model combines two SVM classifiers, the first forecasts and ranks miRNA-mRNA duplex alignments and the second learns AGO's local UTR sequence preferences and positional bias in 3'UTR isoforms. As far as the duplex SVM model is concerned, it facilitates the prediction of non-canonical target sites and is advantageous due to the utilization of weights as parameters for local pairwise alignment. Given the description of features φ(miRNA, site) for a duplex alignment, the alignment score can be depicted as: w* φ (miRNA, site). The conversion of this score to an SVM discriminant function, requests the definition of a match/mismatch score that relies on the position in the miRNA sequence along with the aligned nucleotides, the penalties for loop opening as well as symmetric and asymmetric loop extensions.

To assess the performance of the duplex model, for each different miRNA family, chimiRic produces and scores the duplexes between miRNAs in the seed family and mRNA site sequences in the test set. Another experiment that shows the superiority of the duplex model is the prediction of duplexes for non-canonical miRNA target sites, which have already been evaluated in the past.

The evaluation of the combined chimiRic model is implemented by extracting for each miRNA seed family, all HEK293 positive and negative site sequences. In the positive site sequences, canonical and non-canonical sites, advocated by chimeric reads from CLASH data as well as canonical sites with AGO CLIP data that are undeniably assigned to the specific seed family, are included.

Moreover, in an attempt to predict the extent of mRNA downregulation of miRNA targets, the performance of chimiRic is tested on eight miRNA transfection experiments in HCT116 cells.

Figure 19. Depiction of the iterative procedure of the duplex model. The following procedure is repeated twelve times until the convergence of the model. **a)** Iterative optimization of w given the currents alignments. **b)** Computation of optimal alignments given current w (simultaneous optimization of duplexes and scoring model) **c)** Initial duplex structure for each pair is predicted by duplexfold (in ViennaRNA package) **d)** Corresponding duplex feature vectors are used to train a linear SVM classifier e) w were used as local alignment parameters to update the duplex structure between the miRNA and mRNA site sequences.


Optimal duplexes φ(miR, site)
for given alignment parameters
miRNA-mRNA
local alignment
SVM Training
+ (miR, site) pairs
5' UUAAUGCUAAUUGUGAUAGGGGU miRNA
|||| ||| ||:||||
mRNA AAUU-CGAGUUG--CUGUCCCCU 5'
– (miR, site) pairs
Optimal alignment parameters w
for given duplexes

Concerning the AGO binding model, when its training embodies CLIP data from a single cell type, a regular SVM classifier is applied to the UTR kernel matrix. On the other hand, the mix of data sets from various and different cell types requires the multi-task learning approach. After an alteration of the kernel matrix, the multi-task SVM is created as follows:

$$Kst(x, z) = (\mu + \delta st)\, K(x, z)$$

The free parameter μ manages how close are the task-specific models to the average model, while its optimal value is established through five-fold cross-validation.

In the study of sequence features in the AGO binding model, positional oligomer importance matrix (POIM) [145] approach is utilized in order to determine the more significant positional k-mers in Argonaute binding sequences.

Consequently, the enrichment of these k-mers in RNAcompete data across all RNA binding protein motifs takes place and the significance relative to an empirical null model based on training SVMs on random permutations of the class labels, is computed.

2.11.5 Results

It is found that the duplex model more accurately distinguishes true from false interactions in comparison to MIRZA [99] algorithm, which is trained in the same HEK293 PAR-CLIP data set as chimiRic. Similarly, the duplex model outperforms again MIRZA in the prioritization of observed interactions for each miRNA seed family against interactions with targets sites of other miRNAs in C. elegans and mouse brain. The dominance of the duplex model is again demonstrated as not only the interacting miRNAs are correctly detected above the other highly expressed miRNAs, in spite of the absence of exact 6-mer seed matches, but also the predicted non-canonical modules include GU wobbles, mismatches and bulges in the seed region as well as complementary base pairings in the 3’ region.

The results derived from the validation of the combined chimiRic models manifest that chimiRic’s top-ranked predictions provide at least the same level of accuracy as other available methods trained on AGO CLIP data sets.

According to the extent of mRNA downregulation of miRNA targets, it is corroborated that chimiRic presents alike amount of regulation compared to TargetScan, whereas it achieves better performance than mirSVR.

Taking into account the results from previous studies, it is concluded that there is a plethora of AGO-bound sites near the stop codons and near the end of the 3'UTR, compared to miRNA seeds with no AGO binding in CD4+ T cells across mouse transcripts. Furthermore, for multi-UTR transcripts, abundant AGO-bound sites in the region upstream of internal 3' cleavage sites (as mapped by PolyA-seq) are observed. It is also reported an increased number of ~200nt positive site examples downstream of internal cleavage sites. Additionally, HEK293 AGO binding sites are discovered to be in abundance upstream of internal 3' cleavage sites based on the human 3' end atlas (mapped by 3'-seq) and also downstream.

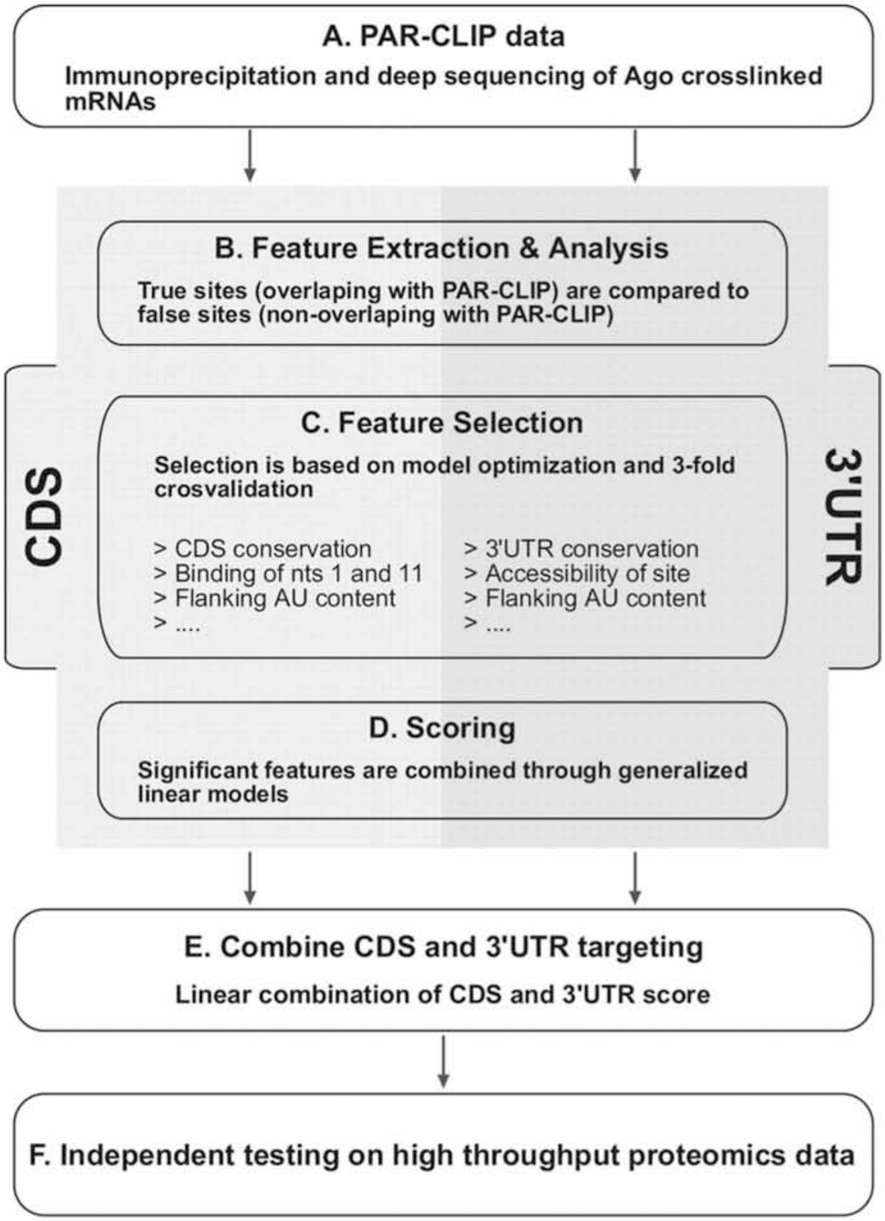
When examining the effect of POIMs method, it is found that the AU content, flanking the miRNA seed matches, presents high value and more complicated sequence features are ready to be explored.

In the search of potential RNA binding protein motifs near miRNA target sites, it is verified that the position-specific k-mers in upstream and downstream sequences are compatible with known RBP motifs. In the common AGO-binding model, an AC-rich motif upstream of the seed match that corresponds to an AGO RNAcompete experiment, is detected and in the downstream component of the common model, Pumilio is identified. Finally, it is verified that Pumilio plays a pivotal transcriptome-wide role in the AGO binding. Indeed, the comparison of HEK293 AGO CLIP to PUM2 PAR-CLIP in the same cell type [79] results in a 16.4% overlap between the AGO sites in HEK293 and PUM2 binding sites.

2.12 DIANA microT-CDS

DIANA microT-CDS [146] is the 5th version of the DIANA microT algorithm that incorporates General Linear Models and Logistic regression to identify miRNA targets extracted from photoactivatable-ribonucleoside-enhanced cross-linking immunoprecipitation (PAR-CLIP) data by [79]. Indeed, the target prediction program is trained on a positive and a negative set of miRNA Recognition Elements (MREs) and learns the features associated with miRNAs, whose binding site is located both in coding sequences (CDs) and 3′UTRs. Features that are included in the miRNA target prediction are the identification of miRNAs and their predicted location of MREs both in the coding sequences (CDS) and in the 3′-UTR, the binding category weight, conservation of MREs targeted CDS or 3′-UTR in 16 and 39 species respectively, the distance to the nearest end of the region (CDS or 3′UTR) or to an adjacent binding site, the predicted free energy of the duplex and AU content. It is crucial to mention that this algorithm selects for analysis the longest annotated transcript, which is the one with the longest 3′- UTR sequence for each gene. In the following flowchart (**Figure 20**), the implementation of DIANA microT-CDS is described in detail.

Figure 20. Flowchart of the analysis on the PAR-CLIP data. MREs specified by the PAR-CLIP data, are divided in two categories according to the genomic region in which they lie **(A).** For these two datasets, 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 cross-validation **(C).** For each identified miRNA MRE, the selected features (depending on the gene region it lies in) are combined into an MRE score through generalized linear models **(D).** For each gene, the CDS score and the 3′-UTR score is defined by summing the MRE scores that lie in CDSs and 3′-UTRs, respectively. These two scores are linearly combined into a final score **(E).** To test for the overall performance of this scoring approach, an independent test on the high-throughput proteomics data of Selbach et al. [56] is performed **(F)** [146].


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

After scrutiny of the results delivered by DIANA microT-CDS, the following parameters play an essential role in its function (**Table 14**):

Table 14. Important parameters of DIANA microT-CDS.

Feature	Range	Description
miTG score	0<miTG score<1	Combined score that measures the potency of each miRNA-gene interaction. The greater the score and close to 1, the greater the confidence
Binding type	6mer, 7mer, 8mer, 9mer and miRNA bugle	The matching sites between the miRNA and the mRNA
Score	0<Score<1	The site contribution score in the miTG score
Conservation	≥0	Number of species where the predicted interaction is conserved
Signal-to-noise ratio	>0	This score calculates the capacity of identification of the miTG score of each interaction without background noise
Precision	from 0 to 1	The score indicates the false-positive rate in a miTG interaction

Apart from PAR-CLIP data, microarray data, proteomics data and HITS-CLIP data are also used in the analysis of this study. In particular, 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 datasets that are used come from Gennarino et al. (2009) [103]: E-GEOD-12091 (mir-26b), E-GEOD-12092 (mir-98); from Wang and Wang (2006) [147]: E-GEOD-6207 (miR-124), E-GEOD-9586 (miR-335); from Linsley et al. (2007) [80]: GSM155604 (miR-106b); from Grimson et al. (2007) [23]: GSM210897 (miR-7), GSM210898 (miR-9), GSM210901 (miR-122a), GSM210903 (miR-128a), GSM210904 (miR-132), GSM210909 (miR-142), GSM210911 (miR-148b), GSM210913 (miR-181a). Proteomics data, as calculated in

Selbach et al. (2008) [56], are downloaded from http://psilac.mdc-berlin.de and HITS-CLIP data are downloaded from Chi et al. (2009) [55].

As far as feature extraction is concerned, a dynamic programming algorithm that identifies the optimal alignment between the miRNA-extended seed sequence (1–9 nt from the 5′ end of miRNA) and every 9 nt window on the 3′ UTR or CDS, is being implemented. A separate prediction model is built for the 3′ UTR and CDS regions and then these are combined to calculate the final miRNA:mRNA interaction score.

2.12.1 Training data

The algorithm is trained on a positive and a negative set of MREs defined by PAR-CLIP data of Hafner et al. (2010) [79]. The true set of MREs consists of MREs which contain the highest number of Watson Crick binding nucleotides, in case there are multiple possible miRNA bindings in the same region. On the other hand, the false set contains all aligned locations that do not overlap with the PAR-CLIP data. In particular, it is found that out of the 17,310 PAR-CLIP peaks throughout the genome, 5,075 overlap with MREs in 3′-UTRs and 6,057 overlap with MREs in CDSs.

2.12.2 Test data

The algorithm is tested on a group of experimentally supported targets for five miRNAs, identified through a high-throughput proteomics method and HITS-CLIP data as defined in Selbach et al. (2008) [56].

2.12.3 Results

The novel computational model provides a significant increase in sensitivity compared to the 3′-UTR-only region model (65% vs 52%), keeping the specificity at the same level of 32%. Indeed, 293 additional correctly predicted targets were identified. When compared with other miRNA target prediction programs, DIANA microT-CDS demonstrates the highest sensitivity at any level of specificity. Interestingly, a high increase in sensitivity is observed at lower specificity values. In addition, it was found that genes with shorter 3′-UTR (3′-UTRs <500 nt) have significantly more targets in coding regions and subsequently higher CDS target score.

Chapter III

Output Description of Algorithms

3.1 TargetScan

For the analysis, the files **Conserved site context++ scores**, and **Non conserved site context++ score**s in all predictions for representative transcripts section have been obtained. These files consist of the precomputed predictions, derived from TargetScan. The columns of the respective files are described below.

Columns of Conserved site context++ scores file

Gene ID, Gene Symbol, Transcript ID, Species ID, miRNA, Site type, UTR start, UTR end, 3' pairing contribution, local AU contribution, position contribution, context++ score, context++ score percentile

Columns of Nonconserved site context++ scores file

Gene ID, Gene Symbol, Transcript ID, Species ID, miRNA, Site type, UTR start, UTR end, 3' pairing contribution, local AU contribution, position contribution, context++ score, context++ score percentile

3.2 PACCMIT

PACCMIT algorithm contains precomputed predictions. In particular, its predictions are based on accessibility filter and on accessibility and conservation filter and are described below (**Table 15**).

Table 15. Description of PACCMIT prediction files

Dataset	Organism	Method	Description	File Name
1	Human	PACCMIT Accessibility (restricted location)	These predictions have been obtained by restricting the location of the nucleation region, i.e., only seed matches with accessible 4-mers pairing miRNA	**paccmit_access_2_5.txt**
2	Human	PACCMIT Access + Cons (restricted location)		**paccmit_access_cons_2_5.txt**

			positions 2-5 are considered. This was shown to increase the precision of the algorithm with respect to the case, in which any accessible 4-mer within the seed match was enough to label the site as accessible. An optimized P_{cutoff} of 0.2 was used.	

3.3 PACCMIT-CDS

PACCMIT-CDS contains precomputed predictions. The file **predictions_human.txt.zip** includes the final predictions obtained using the PACCMIT-CDS program without conservation and the file **predictions_human_cons.txt.zip** embraces the final predictions obtained using the PACCMIT-CDS program with conservation. Furthermore, **paccmit-data.xz** holds the full genome and miRNA data files, which can be decompressed on Linux with *tar -xJvf paccmit-data.xz*. In this analysis, only the **predictions_human.txt.zip** and **predictions_human_cons.txt.zip** predictions have been used.

Below, the basic usage of PACCMIT-CDS is portrayed. A Linux-like operating system running on x86-64 architecture is assumed.

1) The PACCMIT-CDS program package **paccmit-cds.tgz** should be downloaded.
2) It can be decompressed via the command *tar -xzvf paccmit-cds.tgz*
3) The program can be compiled with the command *make*
4) The program should run in two basic ways:
   - ➢ if the conservation of target sites is not required:

     ```
     ./paccmit-cds -g ./data/genes_noncons_example.fa -m ./data/miRNAs_example.fa\
     -i 8
     ```

   - ➢ if the conservation of target sites is required (e.g., in at least 12 species):

```
./paccmit-cds -g ./data/genes_cons_example.fa -m ./data/miRNAs_example.fa -i
8\ -x 27 –M”>11”
```

The Concise User's manual (**paccmit-cds_concise_manual.pdf**) contains all the details concerning the installation and the run of PACCMIT-CDS algorithm.

3.4 MIRZA-G

The files MIRZA-G per-gene scores for miRNA target sites found with MIRZA itself (**mirza-g_all_mirnas_per_gene_scores.tab**) and MIRZA-G per-gene scores for miRNA target sites found with seed scan (**seed-mirza-g_all_mirnas_per_gene_scores.tab**) contain the precomputed predictions for the MIRZA-G algorithm.

The aforementioned files contain 4 columns:

1. Gene (correspond to GeneID in the NCBI database)
2. miRNA (miRNA name in MirBase 20)
3. Total score without conservation
4. Total score with conservation

As far as the installation of the pipeline of MIRZA-G is concerned, the instructions can be found in this online manual.

Dependences useful for the installation of the pipeline include the MIRZA package, the CONTRA fold package, Jobber python library for workflow management as well as other python packages such as drmaa (useful for submission to the cluster), statsmodels, pandas, BioPython, dendropy, numpy and scipy.

3.5 RNA22

This application uses RNA22 v2 to find putative microRNA binding sites in each sequence and then identifies the targeting microRNA. For the analysis, the full set of predictions (***Homo Sapiens, mRNA, ENSEMBL 78, mirBase 21 and RNA22v2***) has been obtained. The columns of the respective files are described below. Each line of the miR file is a predicted target site for that miR.

Column 1 (and filename): name of miR

Column 2: Ensembl Gene ID, Ensembl Transcript ID, chromosome, and strand (-1 for reverse, 1 for sense)

Column 3 & 4: Both of these columns are used to calculate the start/end location of the predicted target site that the miR targets – see section below.

Column 5: always test.seq (this can be ignored)

Column 6: binding energy in -Kcal/mol of the predicted heteroduplex between microRNA and the targeted messenger RNA

Column 7: target site sequence

Column 8: miR sequence

Column 9-10: shows you where the base pairings are for the formed heteroduplex

Column 11 & 12: always the same, the number of paired nucleotides in the heteroduplex

Column 13: the span/length of the predicted target (will be used to calculate the start/end location of the predicted target site that the miR targets)

Column 14-15: please ignore and do not use

Column 16: for RNA22v1.0, this column (quality_estimate) should be ignored. For RNA22v2.0 this column represents the p-value. The p-value represents the likelihood that the target site loci is random. In particular, a lower p-value represents a greater chance that the loci contains a valid MRE

Column 17: region information (5′UTR, CDS, 3′UTR) for the predicted target site. It is crucial to note that a predicted target could overlap in both parts in which both parts will be specified

In order to calculate the local (1-based index) coordinates (these coordinates are relative to the start of the cDNA of the transcript) of the predicted target site, the following procedure is executed:

1. Use columns 3, 4, and 13
2. The 4th column is an offset. So for the example above:
    a) the start location is 3557 (from column 3) + 321 (from column 4) = 3878
    b) the end location is 3557 (from column 3) + 321 (from column 4) + 18 (from column 13) – 1 = 3895

3.6 TargetRank

The output file of TargetRank lists the ranked targets for each miRNA available in miRBase version 10.0. The file for the human species has only been used. Each human miRBase miRNA gene name has been extracted from **hsa_miRBase_miR_ranked_targets.txt** file and entered into the web application due to the fact that the aforementioned file did not contained the score value. The output of the web server is formatted as described below.

Rank

Each 3' UTR is assigned a rank based on its TargetRank score. The 3' UTR with the highest score (predictive of greatest down-regulation in the presence of the siRNA/miRNA) is given a rank of 1. 3' UTRs with identical scores are given the same rank.

Gene Name

The gene name for the corresponding Refseq (field links to the Entrez Gene database entry).

Gene/Isoform Description

Descriptive name for the corresponding Refseq.

Refseq ID

ID for the mRNA isoform scored by TargetRank.

TargetRank Score

Score assigned by TargetRank to a Refseq's 3' UTR.

Seed Match Total

Combined count of conserved and non-conserved 8mer, 7mer and 6mer seed matches.

Conserved Seed Matches

Seed Matches perfectly conserved across aligned human, mouse, rat and dog genomes. If multiple input sequences/miRNA names are provided and a seed match type is ambiguous (i.e. the same 3' UTR sequence can be interpreted as a different seed match type, depending on which input siRNA/miRNA is considered), then all seed match types are assigned a value of 'NA'.

Non-conserved Seed Matches

Seed Matches with at least one nucleotide difference in one of the other three aligned genome regions (e.g. for a human seed match, at least one nucleotide difference in the aligned mouse, rat, or dog sequence). If multiple input sequences/miRNA names are provided and a seed match type is ambiguous (i.e. the same 3' UTR sequence can be interpreted as a different seed match type, depending on which input siRNA/miRNA is considered), then all seed match types are assigned a value of 'NA'.

3.7 MirSVR

The microRNA target predictions and expression data for precomputed results are available as tab delimited files. In the analysis, the files Good mirSVR score, Non-conserved miRNA (**hg19_predictions_S_0_aug2010.txt**) and Good mirSVR score, Conserved miRNA (**hg19_predictions_S_C_aug2010.txt**) are used. “Good” mirSVR score refers to miRNA targets with <−0.1 score. In particular, there is a large overlap between score ranges for 8-mer sites and the 7 (m8) sites and only a subtle difference between the 7 (A1) and 6-mer distributions. In addition, MirSVR contains executable code, which is described below.

Basic Installation instructions

The simplest way to compile this package is:

1) cd to the directory containing the package's source code (**miRanda-3.3a** directory)
2) Give execute permission to your script:
    i. *chmod +x configure*
3) type *./configure* to configure the package for your system. Running configure takes a while. While running, it prints some messages telling which features is checking for.
4) Type *make* to compile the package.

5) Optionally, type *make check* to run any self-tests that come with the package.
6) Type *make install* to install the programs and any data files and documentation.
7) You can remove the program binaries and object files from the source code directory by typing *make clean*. To also remove the files that configure created (so you can compile the package for a different kind of computer), type *make distclean*.

Documentation is contained in the man subdirectory. Some Example files are contained in the examples subdirectory.

The algorithm is running as:

Generally

miranda file1 file2 [-sc score] [-en energy] [-scale scale] [-strict] [-go X] [-ge Y] [-out fileout] [-quiet] [-trim T]

[-noenergy] [-restrict file] miRanda reads RNA sequences (such as microRNAs) from file1 and genomic DNA/RNA sequences from file2. Both of these files should be in FASTA format.

OPTIONS

--help -h

Displays help, usage information and command-line options.

--version -v --license

Display version and license information.

-sc score

Set the alignment score threshold to score. Only alignments with scores >= score will be used for further analysis.

-en energy

Set the energy threshold to energy. Only alignments with energies <= energy will be used for further analysis. A negative value is required for filtering to occur.

-scale scale

Set the scaling parameter to scale. This scaling is applied to match / mismatch scores in the critical 7bp region near the 5' end of the microRNA. Many known examples of miRNA:Target duplexes are highly complementary in this region. This parameter can be thought of as a contrast function to more effectively detect alignments of this type.

-strict

Require strict alignment in the seed region (offset positions 2-8). This option prevents the detection of target sites which contain gaps or non-cannonical base pairing in this region.

-go X

Set the gap-opening penalty to X for alignments. This value must be negative.

-ge Y

Set the gap-extend penalty to Y for alignments. This value must be negative.

-out fileout

Print results to an output file called fileout.

-quiet

Quiet mode, omit notices of when scans are starting and when sequences have been loaded from input files.

-trim T

Trim reference sequences to T nucleotides. Useful when using noisy predicted 3'UTRs as reference sequences.

-noenergy

Turn off thermodynamic calculations from RNAlib. If this is used, only the alignment score threshold will be used. The -en setting will be ignored.

-restrict file

Restrict scans to those between specific miRNAs and UTRs. *file* should contain lines of tab separated pairs of sequence identifiers: miRNA_id <tab> target_id.

If we consider the example files as input, miRanda is runned as:

miranda ./examples/bantam_stRNA.fasta

./examples/ hid_UTR.fasta – sc 120 -en 1 -go -9 -ge -4 -out miranda_out.txt

miRanda algorithm produces **miranda_out.txt** as output.

3.8 MBSTAR

Download Instructions:

- Download the MBStar package
- Unzip and extract all its files.
- Go to MBStar directory. Type *'chmod +x MBStar'* and *'chmod +x MIL-Forest'*.
- Type*: ./MBStar -test < filename > -3utr < filename > -mir < filename >*to execute MBStar.

Usage:

Type: *./MBStar --help to see its usage.*

Parameters:

test file: A Tab delimited, two columns input file. First column contains the miRNA name (ex: >hsa-miR-155-5p) and the second column contains the Refseq id (ex: NM_182715) of mRNA. The last line of file should end with a '>'.

- 3utr file: A fasta format file, containing the 3'UTR sequences.
- mir file: A fasta format file, containing the miRNA sequences.

Download zip file contains:

- hg19_3utr.txt: Human 3'UTR database file.
- mirbase_hsa_latest.txt: MicroRNA sequence file.
- positive_train_instance_index: An index file to train the classifier.
- negative_train_instance_index: An index file to train the classifier.

- positive_train40: Data file for training the classifier.
- negative_train40: Data file for training the classifier.
- config.conf: Configuration file for classifier.
- MBStar_example.txt: Some example interactions for MBStar.

The result is obtained by the following command:

```
./MBStar -test MBStar_example.txt -3utr hg19_3utr.txt -mir mirbase_hsa_latest.txt
```

The result is saved in the file named: **MBStar_predicted_binding_sites.txt**

The result columns are as follows:

Column 1: miRNA Name

Column 2: mRNA Name

Column 3: Binding sites with flanking regions

Column 4: miRNA Sequence

Column 5: Site type

Column 6: Binding position

MBSTAR also contains precomputed predictions, which can be downloaded from the option***: Genome wide target prediction for MBStar***.

The result columns are as follows:

Column 1: miRNA Name

Column 2: mRNA Name

Column 3: Binding sites with flanking regions

Column 4: miRNA Sequence

Column 5: Seed type

Column 6: Binding position

Column 7: Score

3.9 MirMark

Installation

The source code of MirMark is runned on Linux. Firstly, if any dependencies are missing, they should be installed as follows:

```
sudo add-apt-repository ppa:j-4/vienna-rna
sudo apt-get update
sudo apt-get -y install python-software-properties software-properties-common
sudo apt-get -y install gbrowse
sudo apt-get -y install default-jdk
sudo apt-get -y install vienna-rna wget
cd /tmp && wget http://mn.eng.hawaii.edu/~garmire/weka.jar && sudo cp weka.jar /usr/lib/jvm/default-java/jre/lib/ext/.
sudo apt-get -y install build-essential
cd /tmp && wget http://cbio.mskcc.org/microrna_data/miRanda-aug2010.tar.gz && tar zxvf miRanda-aug2010.tar.gz && cd miRanda-3.3a && ./configure && make && sudo make install
cd /tmp && wget http://mn.eng.hawaii.edu/~garmire/MirMarkRNApl.tgz && tar zxvf MirMarkRNApl.tgz && sudo cp MirMarkRNApl/*.pl /usr/share/perl5
cd /tmp && wget http://mn.eng.hawaii.edu/~garmire/MirMark.tgz && tar zxvf MirMark.tgz && sudo cp MirMark/*.pl /usr/local/bin
```

In addition, the following repository should be downloaded to the local machine:

```
git clone --depth 1 https://github.com/lanagarmire/MirMark.git
```

After downloading the repository, in the `Core/` folder all the scripts are shown:

```
cd MirMark/Core
ls
```

In order for them to run properly, they should be copied into the local `bin` directory:

```
cp * /usr/local/bin/
```

Now, the scripts on the test files under **`MirMark/Test`** folder can be tested.

There exist two separate scripts for predicting targets in site and UTR level respectively.

1) The script for predicting targets in site level contains the following arguments: *siteFeaturesARFF.pl <mir-fasta-file><utr-fasta-file><phastcon-utr-scores><pair-file><output-prefix> MirMark/Test/rf.site.model*
   For `<**mir-fasta-file**>`, it can be retrieved from mirbase.org.
   For `<**utr-fasta-file**>`, queries from the UCSC table browser are necessary.
   The format for `<phastcon-utr-scores>` is:
   ... *see MirMark/Test/fast.txt for example*
   The `<**Pair fi**le>` is a TSV of miR and UTR ids, corresponding to fasta file IDs.
   The `<**output-prefix**>` is the output files without the extensions.

2) The script for predicting targets in UTR level contains the following arguments:
   utrFeaturesARFF.pl <mir-fasta-file><utr-fasta-file><phastcon-utr-scores><pair-file><output-prefix> MirMark/Test/rf.site.model

Execution

In MirMark/Core the following commands are run:

siteFeaturesARFF.pl ../Test/mirs.fa ../Test/utrs.fa ../Test/fast.txt ../Test/pairs.txt ../Test/site_features ../Test/rf.site.model

utrFeaturesARFF.pl ../Test/mirs.fa ../Test/utrs.fa ../Test/fast.txt ../Test/pairs.txt ../Test/utr_features ../Test/rf.utr.model

The output files are embraced In **Results_siteFetauresARFF** and **Results_utrFeaturesARFF** folders.

Input_files involves the files that were used for the run of the algorithm.

The input files involve the following files:

fast.txt (file with conservation): this file contains records such as:

NM_006009_utr3_0_0_chr12_49578578_r 0.987 0.916 0.957 0.995 1.000 1.000 1.000 1.000 0.998 0.998 1.000 1.000 1.000 1.000 1.000 0.993 0.990 0.996 0.993 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 0.995 0.999 1.000 1.000 1.000 0.997 0.998 1.000 1.000 0.988 0.937 0.932 0.938 0.960 0.902 0.561 0.432 0.043 0.038 0.066 0.963 0.988 0.990 0.992 0.982 0.542 0.537 0.482 0.361 0.393 0.390

```
0.200 0.043 0.019 0.000 0.000 0.000 0.017 0.039 0.361 0.754 0.885 0.893 1.000 1.000
1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000
1.000 1.000 1.000 1.000 0.999 0.997 0.999 0.998 1.000 1.000 0.993 0.992 0.996 0.986
0.970 0.973 0.977 0.998 0.999 0.988 0.986 0.987 0.977 0.998 1.000 1.000 1.000 1.000
1.000 1.000 1.000 1.000 0.997 0.988 0.980 0.898 0.950 0.951 0.988 0.996 0.995 0.996
0.979 0.977 0.924 0.116 0.003 0.003 0.003 0.002 0.001 0.000 0.000 0.000 0.011 0.037
0.054 0.060 0.046 0.008 0.006 0.009 0.008 0.454 0.961 0.961 0.896 0.882 0.874 0.679
0.010 0.003 0.006 0.000 0.000 0.004 0.018 0.999 1.000 1.000 1.000 1.000 1.000 0.998
0.993 0.997 0.996 0.998 0.999 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000
1.000 1.000 1.000 1.000 1.000 0.998 0.999 1.000 1.000 1.000 1.000 1.000 1.000 1.000
1.000 1.000 1.000 1.000 1.000 0.996 0.103 0.095 0.095 0.028 0.015 0.000
```

mirs.fa: this file contains orthologous miRNA sequences

utrs.fa: this file contains Refseq mRNA IDs (NM)

pairs.txt: this file contains pairs miRNA-NM

The output files consist of the columns below:

site_features.txt, site_features.csv:

Miranda_score, miR_ID, mRNA_ID, Start_position, End_position, miR_match_P01,miR_match_P03,miR_match_P04, miR_match_P08, miR_match_P15, Seed_bulge, Total_AU, Total_mismatch, Total_bulge, Total_bulge_nt, Seed_P01_acc, Seed_cons_score

site_features.weka:

inst# , actual predicted error prediction

utr_features.txt, utr_features.csv, utr_features.weka:

Miranda_score, miR_ID, mRNA_ID, Start_position, End_position, Seed_match_6mer2, miR_match_P01, Seed_match_7mer2, Seed_match_7mer1, Seed_MFE, X3p_MFE, Target_UC_comp, miR_match_P09, miR_match_P02, Seed_GU, miR_match_P07, Start_position, miR_match_P19, miR_match_P15

3.10 SVMicrO

All the related materials, including source code and genome-wide prediction of human targets of SVMicrO are available here. However, due to forbidden access at their server, neither source code, nor precomputed results were accessible. As a result, SVMicrO was not incorporated into the target prediction analysis.

Chapter IV

4. Overview of miRNA Target Prediction Programs

All the aforementioned computational algorithms for miRNA target predictions are surveyed. Indeed, in **Tables 16 - 19**, the information of each algorithm, including their supported organism, websites, approaches, features, binding sites, assembly type and miRBase version, is summarized.

Table 16. Algorithms for computational target prediction

Programs	Website	Type	Organisms	Reference
TargetScan	*http://www.targetscan.org* (version 7.1)	Web-basedªĦ	h, m, r, d, cn, c, rh, cw, o, fr, z, f, w	[75]
PACCMIT	*http://paccmit.epfl.ch/* or*https://lcpt.epfl.ch/MicroRNA_target_predictions*	Web-based□	h	[93, 94]
PACCMIT-CDS	*http://paccmit.epfl.ch/* or*https://lcpt.epfl.ch/PACCMIT-CDS*	Web-basedªn	h	[93, 95]
MIRZA-G (Mirza Analysis)	*http://www.clipz.unibas.ch/mirzag/*	Web-basedªʕx	h	[98]
MIRZA-G (Seed Analysis)	*http://www.clipz.unibas.ch/mirzag/*	Web-basedªʕx	h	[98]
RNA22	*https://cm.jefferson.edu/tools/*	Web-basedªɣ	h, m, f, w,n	[34]
TargetRank	*http://hollywood.mit.edu/targetrank/*	Web-based□	h, m	[115]
MirSVR	*http://www.microRNA.org*	Web-basedªV	h, m, r, f, n, w	[33]
MBSTAR	*https://www.isical.ac.in/~bioinfo_miu/MBStar30.htm*	Web-basedªV	h	[128]
SVMicrO	*http://compgenomics.utsa.edu/svmicro.html*	ª	h	[133]
MirMark	*http://www2.hawaii.ed*	Web-basedªĦ	h	[134]

	u/~lgarmire/software.html *https://github.com/lanagarmire/MirMark*			
ChimiRic	*https://bitbucket.org/leslielab/chimiric*	$\mathbf{a}_{x}\varsigma$	h,m	[142]
DIANA microT-CDS	*http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=microT_CDS/index*	Web-based^{a}	h,m, f, w, n	[146]

Organisms: h, human; m, mouse; r, rat; d, dog; cn, chicken; c, chimpanzee; rh, rhesus; cw, cow; o, opossum; fr, frog; z, zebrafish; f, fly; w, worm; p, pufferfish; mq, mosquito; n: caenorhabditis elegans (nematode)

Type: ^{a} source code available; □ source code not available; $^{\varsigma}$Github; ^{V}: C; ^{Ħ}:PerlScript; ^{n}: C++; ^{Y}:Java; ^{x}:Python

Table 17. Target Prediction Duplex and Local Context Features

	Duplex Features					**Local Context Features**	
Programs	Seed Match	3'Contribution	SPS*	Heteroduplex Free Energy	P-Value	SA **	Flanking AU
TargetScan	✓	c,s	✓	✓		✓	✓
PACCMIT	✓			✓		✓	
PACCMIT-CDS	✓			✓		✓	
MIRZA-G (Mirza Analysis)	✓	c		✓		✓	

MIRZA-G (Seed Analysis)	✓	c		✓		✓	
RNA22	✓	c		✓	✓	✓	
TargetRank	✓						✓
MirSVR	✓	c		✓		✓	✓
MBSTAR				✓			
SVMicrO	~	c		✓		✓	✓
MirMark	✓	c		✓	✓	✓	✓
ChimiRic	✓						✓
DIANA microT-CDS	✓	c		✓		✓	✓

* Seed Pairing Stability, ** Site Accessibility

3’Contribution : c,3' Compensatory Pairing; s,3’ Supplementary Pairing

Seed Match: ✓, perfect seed match; ~, partial seed match

Table 18. Target Prediction Global Context Features, Binding Region

	Global Context Features								
Programs	TA*	Pairing Position	3’UTR Lenght	Sequence Lenght	Conservation	Global site density	Binding Type	Region	Machine Learning Model
TargetScan	✓	✓	✓		✓		c,n	3u	SR,MR
PACCMIT					✓		c	3u	Markov Model
PACCMIT-CDS					✓		c	3u, cd	Markov Model
MIRZA-G (Mirza Analysis)					✓		c,n	3u, cd	GLM
MIRZA-G (Seed Analysis)					✓		c,n	3u, cd	GLM

RNA22				✓			n	3u, 5u, cd	Markov Chain (based on pattern recognition)
TargetRank	✓	✓			✓		c	3u	NA
MirSVR		✓	✓		✓		c,n	3u, 5u, cd	SVR
MBSTAR							c,n	3u	RF
SVMicrO	✓		✓		✓	✓	c,n	3u	SVM, UTR-SVM
MirMark	✓		✓		✓	✓	c,n	3u	RF, Gaussian SVM
ChimiRic							c,n	3u	SVM
DIANA-microT-CDS	✓				✓		c,n	3u, cd	GLM

*Target site Abundance

Binding Type: c, canonical sites; n, non-canonical sites: mismatch in seed region; cleavage sites; centered sites; bulges

Region: 3u,3'UTR; 5u,5'UTR; cd,CDS(coding region);

Machine Learning Model: GLM, General Linear Model(Logistic Regression); SVR, Support Vector Regression; RF, Random Forest Classifier; SR,Stepwise Regression; MR, Multiple Regression; NA: No Machine Learning

Prior to the comparison of Target prediction programs, they are adjusted so that they use miRBase version 18. In Table 4, conversion files for each program are summarized.

Table 19. Characteristics of target prediction algorithms.

Programs	**miRBase Release**	**miRBase Source**	**Ensembl Release***	**Coordinates of Binding Region**	**Gene/ Refseq/ Transcript /NCBI****
TargetScan	release 21	mirna_mature.txt.gz (*ftp://mirba*	release 75(GRCH37- Hg19)	Relative	✓/-/-/-

		se.org/pub/mirbase/21/database_files/)			
PACCMIT	release 18	mirna_mature.txt.gz (*ftp://mirbase.org/pub/mirbase/18/database_files/*)	release 54 (NCBI36-Hg18) release 91 (GRCH38-Hg38)	-	✓/-/-/-
PACCMIT-CDS	release 18	mirna_mature.txt.gz (*ftp://mirbase.org/pub/mirbase/18/database_files/*)	release 54 (NCBI36-Hg18)	-	-/-/✓/-
MIRZA-G (Mirza Analysis) siRNA sequences	release 20 supplementary material of Schultz et al. [148a]	mirna_mature.txt.gz (*ftp://mirbase.org/pub/mirbase/20/database_files/*)	release 75(GRCH37- Hg19)	-	-/-/-/✓
MIRZA-G (Seed Analysis) siRNA sequences	release 20 supplementary material of Schultz et al. [148a]	mirna_mature.txt.gz (*ftp://mirbase.org/pub/mirbase/20/database_files/*)	release 75(GRCH37- Hg19)	-	-/-/-/✓
RNA22	release 21	mirna_mature.txt.gz (*ftp://mirbase.org/pub/mirbase/21/database_files/*)	release 78	Relative	✓/-/-/-
TargetRank	release 10	mirna_mature.txt.gz (*ftp://mirbase.org/pub/mirbase/10*	> release 54(NCBI 35-Hg17)	-	-/✓/-/-

		/database_files/)			
MirSVR	release 15	mirna_mature.txt.gz (*ftp://mirbase.org/pub/mirbase/15/database_files/*)	release 75(GRCH37- Hg19)	Absolute	-/✓/-/-
MBSTAR	release 20	mirna_mature.txt.gz (*ftp://mirbase.org/pub/mirbase/20/database_files/*)	release 75(GRCH37- Hg19)	Relative	-/✓/-/-
SVMicrO	release 12	-	-	-	-/-/-/-
MirMark	release 19	-	-	-	-/-/-/-
ChimiRic	-	-	release 75(GRCH37- Hg19)	-	-/-/-/-
DIANA-microT-CDS	release 13	mirna_mature.txt.gz (*ftp://mirbase.org/pub/mirbase/18/database_files/*)	release 69	Absolute and Relative	✓/-/-/-

*Various versions of Ensembl are found *here*.

** Gene: Gene Stable ID, Refseq: Refseq mRNA ID, Transcript: Transcript Stable ID, NCBI: NCBI Gene ID

Table 20. Characteristics of Datasets

Test Datasets	miRBase Release	Ensembl Release
Test Dataset 1	release 18	release 83
Test Dataset 2	release 18	release 83

Table 21. miRNA:mRNA human interactions across all programs

Programs	Predictions		
	Conserved	Non Conserved	All
TargetScan	1,450,123	38,310,305	39,760,429

PACCMIT Access	-	-	1,698,192
PACCMIT Access + Cons	-	-	204,699
PACCMIT-CDS Conservation	-	-	1,916,313
PACCMIT-CDS Without Conservation	-	-	3,672,406
MIRZA-G (Mirza Analysis) Conservation	-	-	4,071,768
MIRZA-G (Mirza Analysis) Without Conservation	-	-	4,757,171
MIRZA-G (Seed Analysis) Conservation	-	-	3,787,775
MIRZA-G (Seed Analysis) Without Conservation	-	-	4,391,410
RNA22	-	-	251,888,722
TargetRank	-	-	376,856
MirSVR	1,097,064	3,320,820	4,417,884
MBSTAR	-	-	47,466,965
DIANA-microT-CDS	-	-	14,611,757

All programs were compared with Test Dataset 1 which contains the positive set of experimentally predicted targets in human (**Human_experimentally_validated_interactions.txt file**) and Test Dataset 2 which contains the exact binding positions of miRNA interactions in experimental methodologies. Additional information is given on the coordinates of the binding positions in the reference genome (**Human_experimentally_validated_interactions_Reporter_Chimeric.txt**).

The columns of Test Dataset 1 are described as follows:

1. **miRNA:** The name of the microRNA. The version, from which the name of the microRNA is taken, is miRBase 18 (*http://www.mirbase.org/*).
2. **Ensembl_Gene_id:** The Gene Stable ID derived from Ensembl genome Browser. The version is Ensembl 83 (*http://dec2015.archive.ensembl.org/index.html*).
3. **Method:** The experimental methodology with which the target miRNA has been confirmed.
4. **MIMAT:** The name of the microRNA. The MIMAT does not change between the miRBase versions. Thus, the results of each program are calculated based on MIMAT.
5. **Gene_name:** The name of the gene. It is used only in cases where Ensembl Gene id is not known as the name of a gene may not be unique.

The columns of Test Dataset 2 are described as follows:

1. **chr:** The chromosome.
2. **start:** The genomic site that initiates the miRNA binding site.
3. **end:** The genomic position that terminates the miRNA binding site.
4. **strand:** The type of chain that the miRNA binds ("+", "-").

Due to the fact that Test Dataset 1 contains Gene Stable ID as a column, all programs should be converted in order to involve Gene Stable ID in their datasets. Consequently, the following conversions have been made:

A. Refseq mRNA ID → Gene Stable ID

   When RefSeq mRNA ID is known and has to be converted to Gene Stable ID, the following steps should be pursued at the BioMart website for Ensembl 91 archive:

   1. Click MARTVIEW (top menu)
   2. Choose ENSEMBL 91 for database and Homo sapiens genes GRCh38.p10 for dataset
   3. Click "Filters" (left menu) and expand REGION
      3.1. Select Chromosome/scaffold: 1-22, MT, X, Y
      3.2. In filters expand GENE, choose in Input external references ID list : "RefSeq mRNA ID " and paste your ID(s) or upload a file of IDs. In particular, a file is uploaded.
   4. Click "Attributes" (left menu) and expand GENE
      4.1. Check Ensembl Gene ID, Gene Name
      4.2. In EXTERNAL : select RefSeq mRNA ID
   5. Click "Results" (top left menu)

B. Transcript Stable ID → Gene Stable ID

   When Transcript Stable ID is known and has to be converted to Gene Stable ID, the following steps should be pursued at the BioMart website for Ensembl 91 archive:

   1. Click MARTVIEW (top menu)
   2. Choose ENSEMBL 91 for database and Homo sapiens genes GRCh38.p10 for dataset
   3. Click "Filters" (left menu) and expand GENE

    3.1. Select Chromosome/scaffold: 1-22, MT, X, Y
    3.2. Choose "Ensembl Transcript ID(s)" and paste your ID(s) or upload a file of IDs. In particular, a file is uploaded
4. Click "Attributes" (left menu) and expand GENE
    4.1. Check Ensembl Gene ID, Transcript ID
5. Click "Results" (top left menu)

C. NCBI Gene ID →Gene Stable ID

When NCBI Gene ID is known and has to be converted to Gene Stable I, the following steps should be pursued at the BioMart website for Ensembl 91 archive:

1. Click MARTVIEW (top menu)
2. Choose ENSEMBL 91 for database and Homo sapiens genes GRCh38.p10 for dataset
3. Click "Filters" (left menu) and expand REGION
    3.1. Select Chromosome/scaffold: 1-22, MT, X, Y
    3.2. In filters expand GENE, choose in Input external references ID list: "NCBI gene ID " and paste your ID(s) or upload a file of IDs. In particular, a file is uploaded.
4. Click "Attributes" (left menu) and expand GENE
    4.1. Check Ensembl Gene ID, Gene Name
    4.2. In EXTERNAL : select NCBI gene ID
5. Click "Results" (top left menu)

Additional information concerning the number of miRNA:mRNA interactions, miRNAs, Genes and the source of Test Dataset 1 is summarized in **Table 22**.

Table 22. Basic information of Test Datasets

	Test Dataset 1	**Test Dataset 2**
Total experimentally verified interactions	25,901	2,380
Unique miRNAs	887	257
Unique Genes	10,638	1600
Source	DIANA-TarBase	DIANA-TarBase

From the analysis of all programs, the following datasets were extracted as shown in **Table 23**:

- **miRNAs Subset**: The common set of miRNAs among all programs that is subset of Test Dataset 1
- **Genes Subset 1**: The common set of Genes among all programs
- **Genes Subset 2:** The common set of Genes among all programs that is subset of Test Dataset 1

Table 23. Common set of miRNAs and Genes across all programs

	Total
miRNAs Subset	360
Genes Subset 1	7,620
Genes Subset 2	7,570

Due to the large number of predictions that some algorithms achieve, the threshold values in the scores of each program, as indicated in **Table 24**, were set. The definition of threshold values for each program aims at increasing the precision and reducing the sensitivity of each program. In **Table 24**, the first column indicates the name of the programs while the second and third ones the cutoff values applied for each program.

Table 24. Thresholds for algorithms

	Threshold in Precomputed Results	**Threshold**
TargetScan	8mer: Score* ≥0.8 7mer-m8: Score* ≥1.3 7mer-A1: Score* ≥1.6[œ]	-
PACCMIT	$P_{cutoff} \geq 0.2$	-
PACCMIT-CDS	P-value ≥ 0.05 (Default)	-
MIRZA-G (Mirza Analysis)	Mirza Score ≥0.12	-
MIRZA (Seed Analysis)	Mirza Score ≥0.12	-
RNA22	≥ 60 pattern instance and Gibbs energy ≤-25 Kcal/mol, ≤-18 Kcal/mol.	P-value < 0.05
TargetRank	P-value ≥ 0.05	-
MirSVR	mirSVR score ≤ -0.1[s]	-
MBSTAR	P-value ≥ 0.5	P-value ≥ 0.5
DIANA-microT-CDS	P < 0.05	P-value ≥ 0.5

œ: 6-mer and offset 6-mer sites are always classified as non-conserved
s: Best predictions of mirSVR present good have PhastCOns score >0.57
*: conservation cutoffs for each site type at different branch-length scores. These cutoffs correspond only to conserved sites.

Due to the extremely long processing of the vast number of prevalent protein coding genes, used by target prediction tools in order to produce the expected results, precomputed results of each program are employed. These precomputed results are downloaded and their datasets are adjusted to the proper format in order the finding of miRNA-gene interactions to be facilitated. For the sake of an example, due to the fact that the source of the 3' UTR sequences of each algorithm is derived from Ensembl [148] database, using BioMart data mining tool *(http://dec2017.archive.ensembl.org/index.html)*, the University of California Santa Cruz (UCSC) [130] database or different assemblies, major differences between the prediction outcomes of the target prediction tools are manifested, rendering their comparison difficult and inaccurate. In particular, the aforementioned problem is due to diverse positive and negative training sets, test sets, feature selection and machine learning models among all tools. As a result, three (3) test cases are examined in order the precise evaluation of the miRNA prediction models to be achieved.

Firstly, according to Test Case I the predictions of each program are filtered at the common set of miRNAs among all programs that is subset of the positive set (Test Dataset 1). In this case, an initial abstract approximation of the performance of the algorithms can be gained because they are by far different from each other and their common set of miRNAs is inadequate to provide a holistic common base. Moreover, Test Case II examines the predictions that are retained after the application of the common set of miRNAs among all programs that is subset of the positive set and the common set of Genes among all programs. In comparison to the previous case, this case presents to be more judicious with the concept of comparing tools under the same base. Test Case III, which filters at the common set of miRNAs among all programs that is subset of the positive set as well as the common set of Genes among all programs that is subset of the positive set, appears to be the most appropriate case because it provides similar results with those that derive after running of all miRNA tools under the same conditions.

Table 25 presents the abbreviations, which are utilized in the following figures.

Table 25. Abreviations concerning certain target prediction features.

Abreviation	Meaning
C	Conservation
WC	Without Conservation
CS	Site Conservation
NCS	No Site Conservation
A	Accessibility

AC	Accessibility & Conservation
CM	miRNA Conservation
NCM	No miRNA Conservation

4.1 Test Case I

The predictions of each program were filtered at the common set of miRNAs among all programs that is subset of the positive set. In addition, a threshold of P-value < 0.05 for RNA22, P-value ≥ 0.5 for MBSTAR and P-value ≥ 0.5 for DIANA microT-CDS is considered.

In **Figure 21**, Total predictions and True positive experimentally validated predictions (shared miRNA-Genes interactions with Test Dataset 1) of all programs are shown in blue and yellow respectively. The results are calculated without taking into account the score of each miRNA-Gene interaction. **Figure 22** illustrates the same results as **Figure 21** with the exception that thresholds for RNA22 and MBSTAR have been applied. The threshold for DIANA microT-CDS has been applied to all test cases. From both **Figures 21, 22** it is evident that TargetScan outperforms the other target prediction programs, while RNA22 has a vast number of predictions and therefore is highly sensitive.

Figure 21. Total and True positive set for Target Prediction Algorithms without considering the score according to Test Dataset 1 (Test Case I).

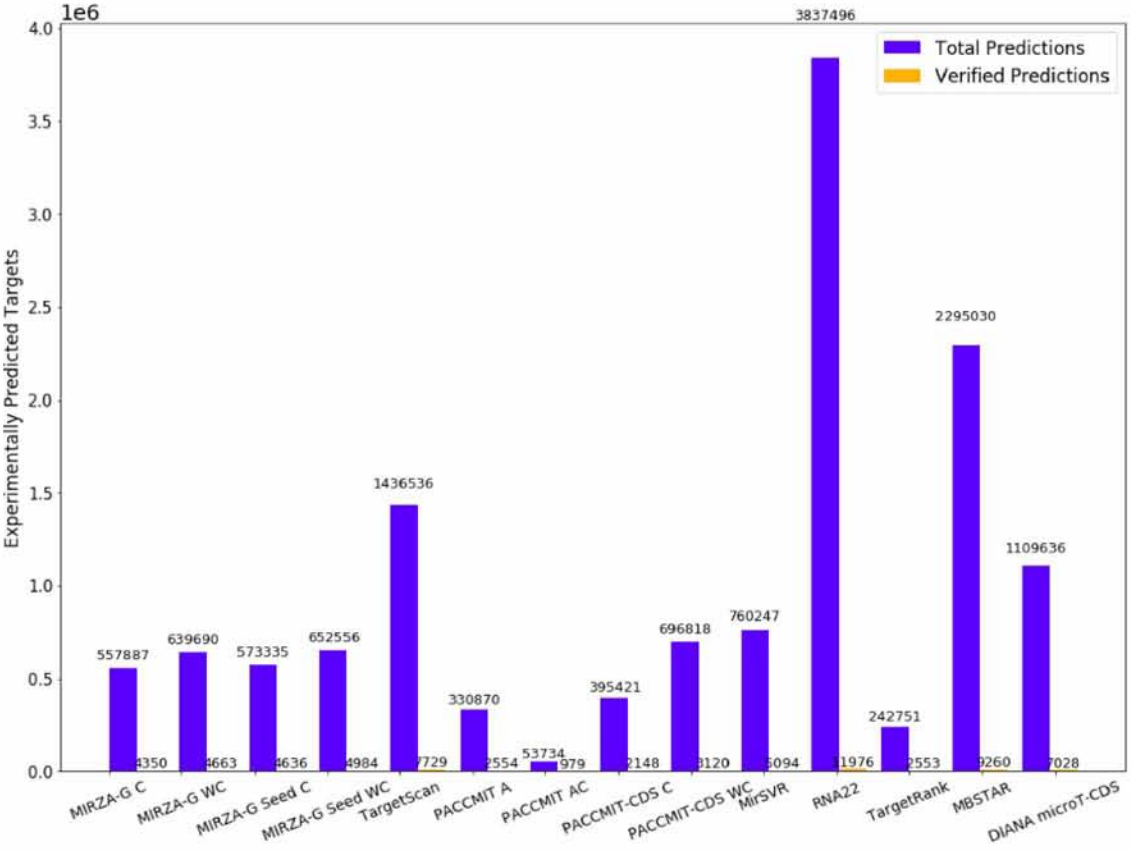


Figure 22. Total and True positive set for Target Prediction Algorithms without considering the score and setting threshold on RNA22 and MBSTAR according to Test Dataset 1 (Test Case I).

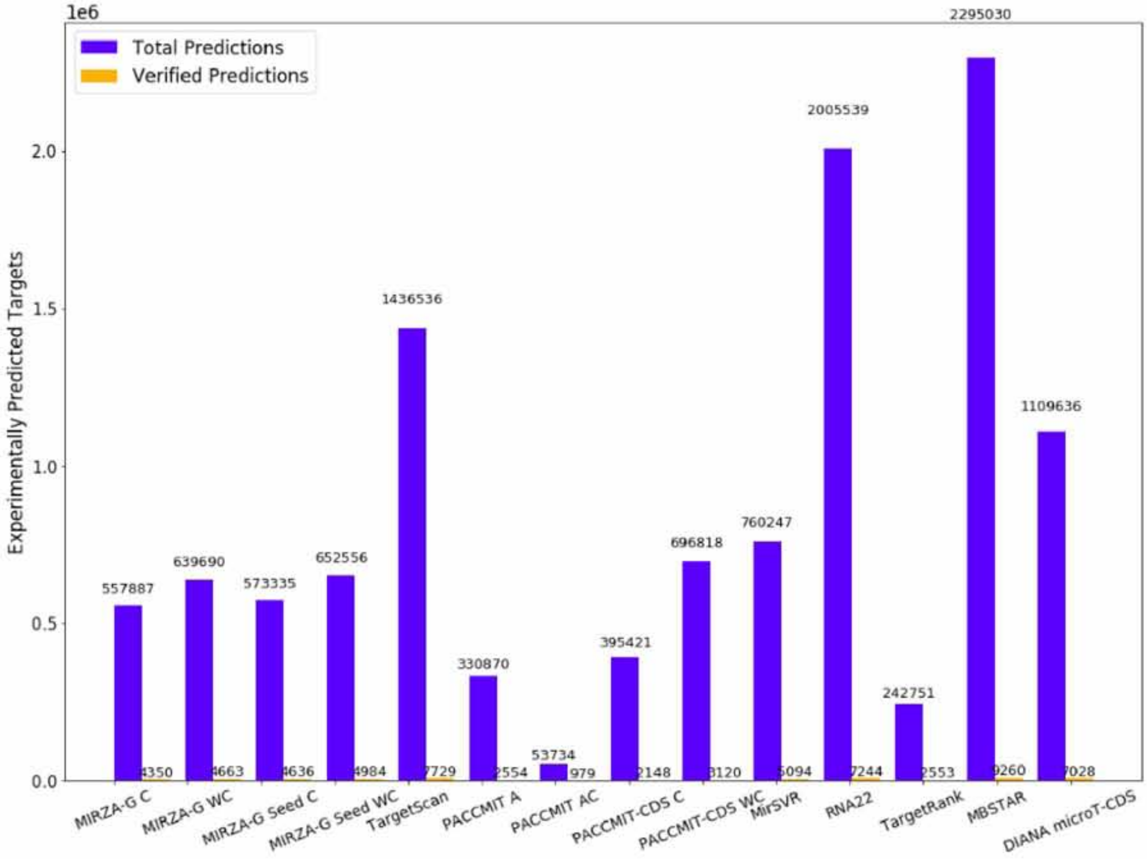
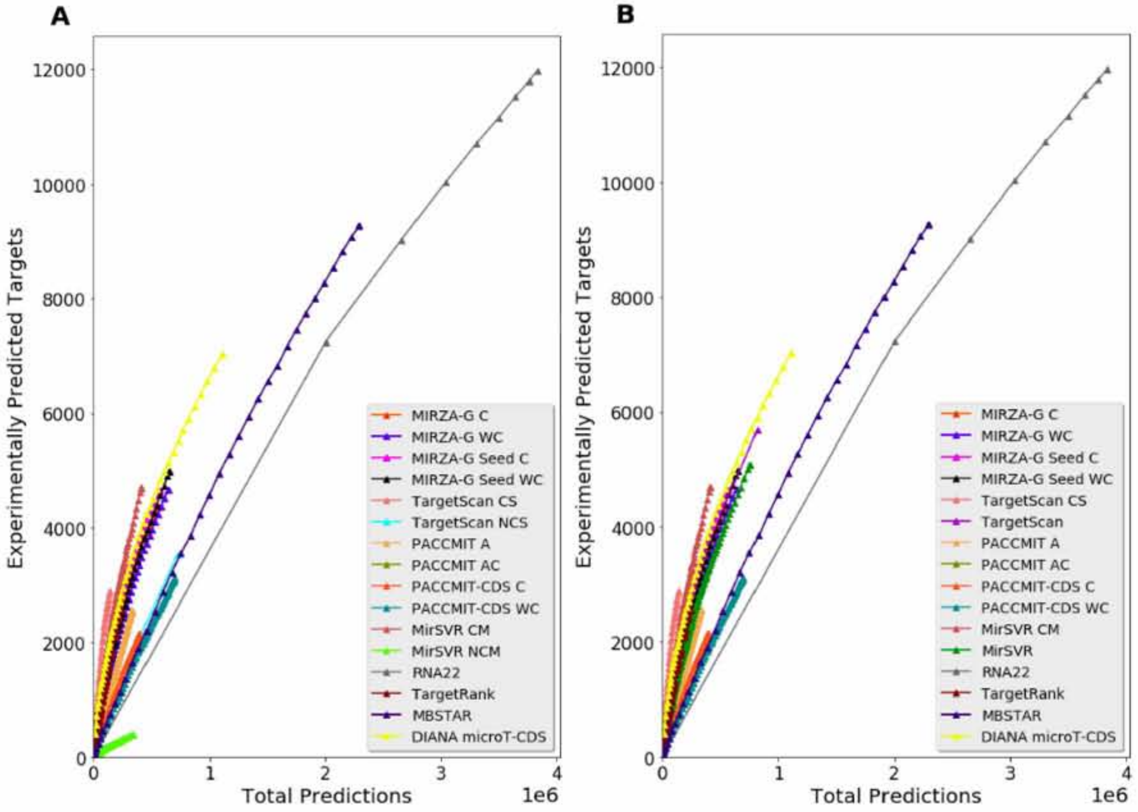


Figure 23 illustrates Total Predictions and Experimentally Predicted Targets of all programs, for corresponding score values. The number of correctly predicted targets is shown by different scores for increasing numbers of total predictions. All predictions, whether conserved, non-conserved or total are filtered on the true positive set of miRNAs.

In **Figure 23A**, it is evident that conserved predictions present an optimized performance compared to non-conserved predictions. In addition, it is observed that TargetScan CS has the optimal performance when compared to other target prediction algorithms due to the fact that it achieves high accuracy and its sensitivity is almost nonexistent. Indeed, the curve that corresponds to Targetscan CS initially starts with a relatively small number of Total predictions and maps them to a large number of experimentally verified predictions. Hierarchically, following TargetScan CS, MirSVR CM, DIANA microT-CDS, MIRZA-G Seed C and MIRZA-G C present the greatest performance among target prediction tools. MBSTAR and RNA22 appear to be very sensitive due to the fact that, although they find a larger number of experimentally supported targets, compared to other programs, at the beginning their Total predictions are proportionally very high. PACCMIT AC, PACCMIT-CDS C, PACCMIT A and PACCMIT-CDS WC seem to hold a small number of Total predictions leading to a tiny number of verified targets.

From **Figure 23B**, conserved miRNA:gene human interactions outperform the Total interactions due to the fact that Total predictions constitute the sum of conserved and non-conserved interactions. For instance, TargetScan CS presents better performance compared to TargetScan. Moreover, conserved miRNA predictions of MirSVR (MirSVR CM) are less sensitive than the one of MirSVR.

Figure 23. Total Predictions vs Experimentally Predicted Targets of all programs when considering step= 0.01 for score values filtered on the true positive set of miRNAs. Between miRNA-gene interactions with the same score, those with the maximum scoring scheme are selected (Test Case I). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.



In **Figure 23**, between miRNA-gene interactions with the same score, the maximum of their scoring scheme is selected. On the other hand, in **Figure 24**, the aggregated score of the corresponding interactions is considered. After careful examination of the plots in the latter case, it is observed that the performance of the algorithms remains intact despite the application of the aggregation filter, with the exception of TargetScan NCS. This is due to the fact that TargetScan assigns a score in each site and not in the entire miRNA-gene interaction as other target prediction programs (e.g DIANA microT-CDS).

In **Figure 25** applies the same graphical analysis as in **Figure 23**, with the sole difference that the predictions of RNA22 with P-value ≥ 0.05 have been cut off. As a result, for a relatively large number of Total predictions, RNA22 achieves a larger number of experimentally verified predictions compared to TargetScan CS. As far as MBSTAR is concerned, the application of the threshold of 0.5 do not alter its predictions, as all miRNA-gene interactions have scores greater or equal to 0.5 prior reaching the stage of implementing the aforementioned threshold.

In **Figure 25**, between miRNA-gene interactions with the same score, the maximum of their scoring scheme is selected. On the other hand, in **Figure 26**, the aggregated score of the corresponding interactions is considered. After careful examination of the plots in the latter case, it is observed that the performance of the algorithms remains intact despite the application of the aggregation filter, with the exception of TargetScan NCS. This is due to the fact that TargetScan assigns a score in each site and not in the entire miRNA-gene interaction as other target prediction programs (e.g DIANA microT-CDS).

Figure 24. Total Predictions vs Experimentally Predicted Targets of all programs when considering step = 0.01 for score values filtered on the true positive set of miRNAs. Between miRNA-gene interactions with the same score, the aggregation of their scoring scheme is selected (Test Case I). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions (Test Case I).

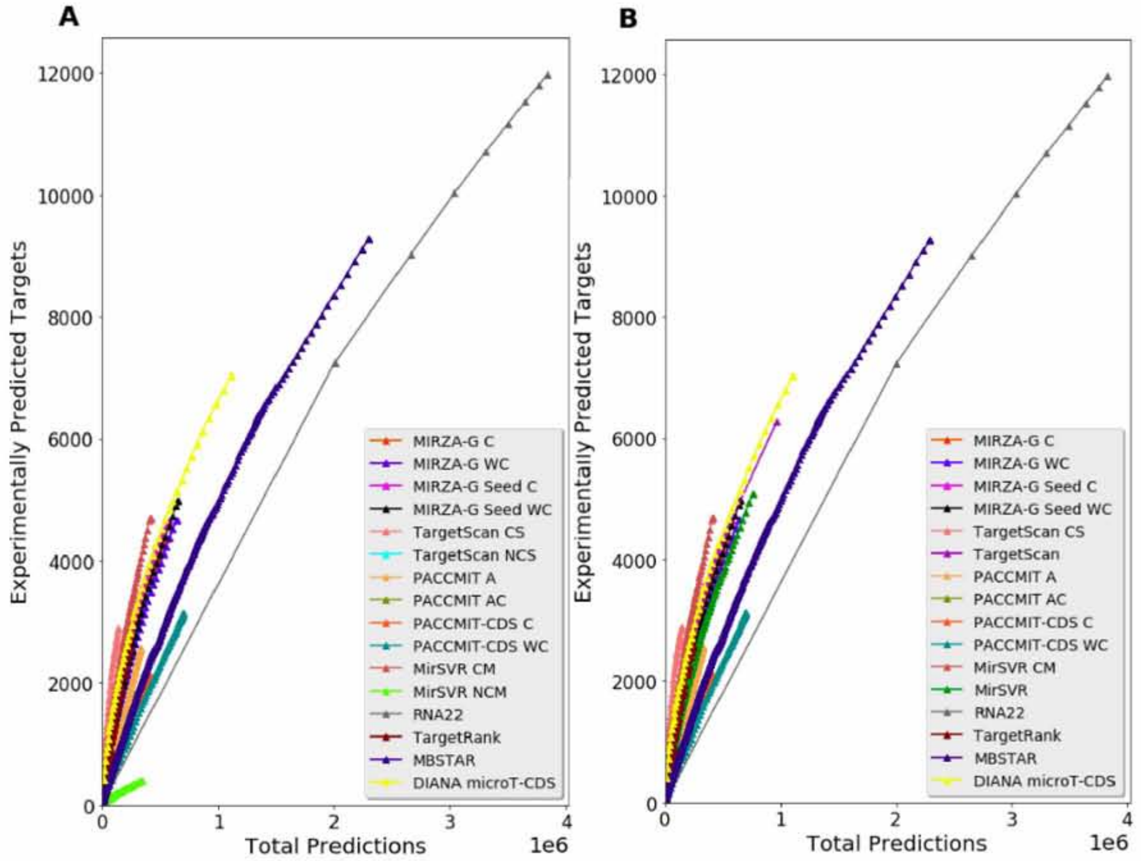


Figure 27 shows the Predicted Targets per miRNA and the corresponding number of correctly predicted targets for different scores. Predicted Targets per miRNA constitute the average value of Total predictions of all miRNAs in each program, having been grouped by score and they indicate the sensitivity of the examined models. Correctly predicted targets are calculated as the average value of experimentally verified targets of all miRNAs in each program, having been grouped by score. **Figures 23 and 27** share the same conclusions as far as the performance of target prediction algorithms is concerned.

In **Figure 28**, despite the selection of the aggregated score of the corresponding miRNA-gene interactions, all algorithms present the same performance as prior to the application of this filter, with the exception of TargetScan NCS.

Figure 25. Total Predictions vs Experimentally Predicted Targets of all programs when considering step = 0.01 for score values filtered on the true positive set of miRNAs and setting threshold on RNA22 and MBSTAR. Between miRNA-gene interactions with the same score, those with the maximum scoring scheme are selected (Test Case I). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.

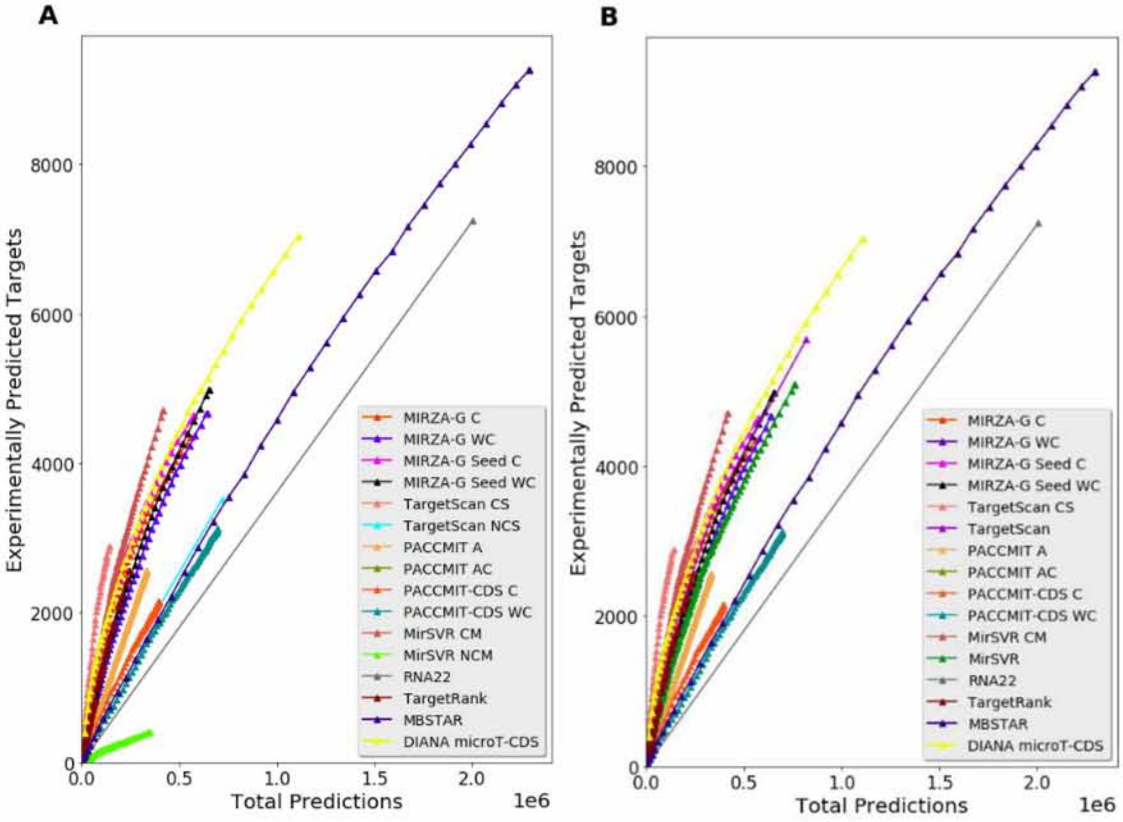


Figure 29 demonstrates the Predicted Targets per miRNA and the corresponding number of correctly predicted targets for different scores after thresholds for RNA22 and MBSTAR have been employed. In **Figure 29** applies the same graphical analysis as in **Figure 27**, with the sole difference that the predictions of RNA22 with P-value ≥ 0.05

have been cut off. Conclusions are compatible with those in **Figure 25**.

In **Figure 30**, the selection of the aggregated score of the corresponding miRNA-gene interactions does not alter the performance of the algorithms, except for TargetScan NCS.

Figure 26. Total Predictions vs Experimentally Predicted Targets of all programs when considering step = 0.01 for score values filtered on the true positive set of miRNAs and setting threshold on RNA22 and MBSTAR. Between miRNA-gene interactions with the same score, the aggregation of their scoring scheme is selected (Test Case I). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.

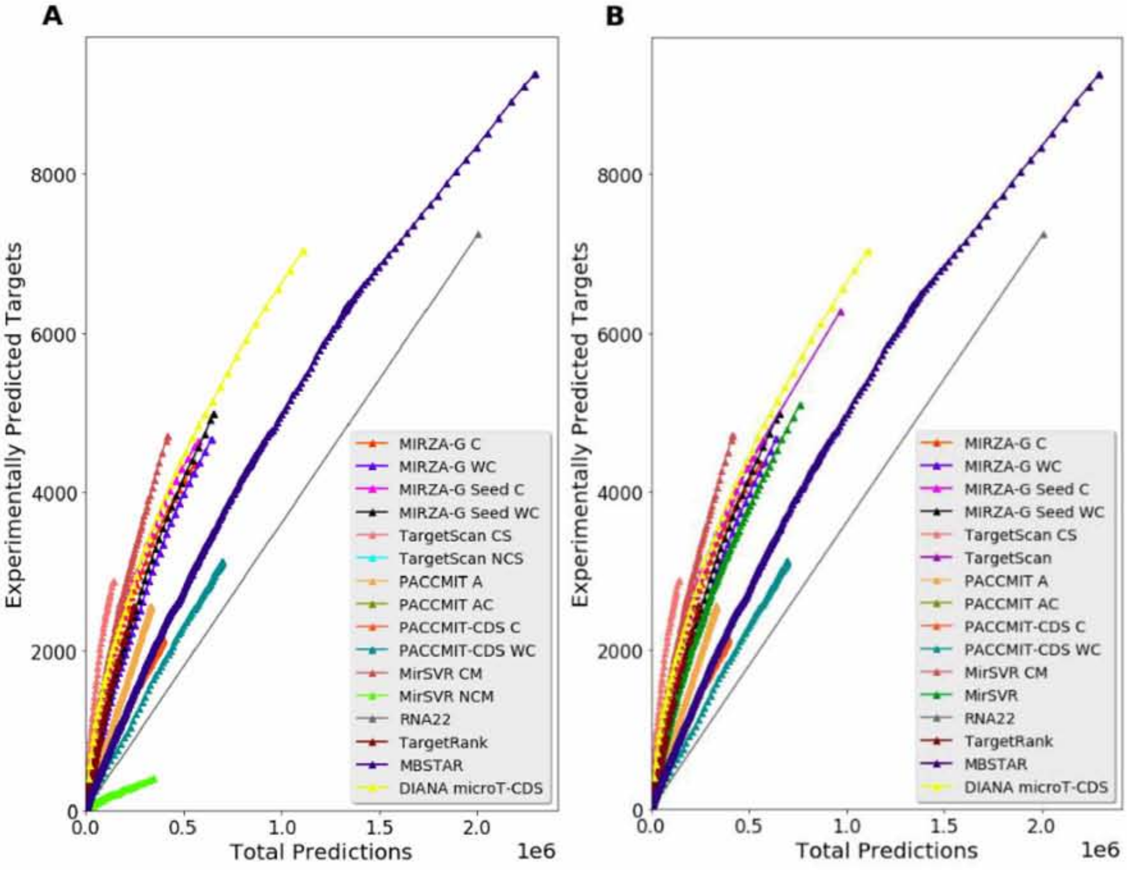


Figure 27. Predicted Targets/miRNA vs Experimentally Predicted Targets. The number of correctly predicted targets is shown by different scores for increasing numbers of predicted targets per miRNA, when considering a step of 0.01. All predictions were filtered on the true positive set of miRNAs. Between miRNA-gene interactions with the same score, those with the maximum scoring scheme are selected (Test Case I). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.

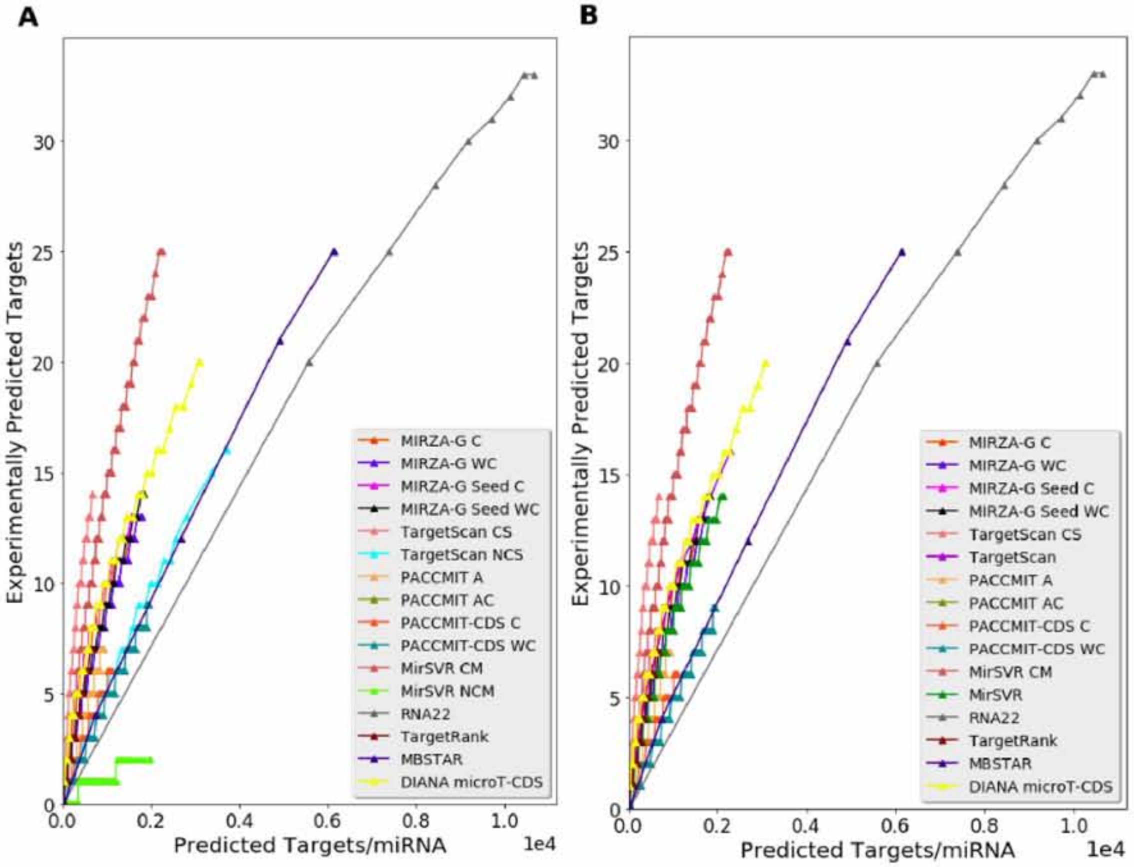


Figure 28. Predicted Targets/miRNA vs Experimentally Predicted Targets. The number of correctly predicted targets is shown by different scores for increasing numbers of predicted targets per miRNA, when considering a step of 0.01. All predictions were filtered on the true positive set of miRNAs. Between miRNA-gene interactions with the same score, the aggregation of their scoring scheme is selected (Test Case I). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.

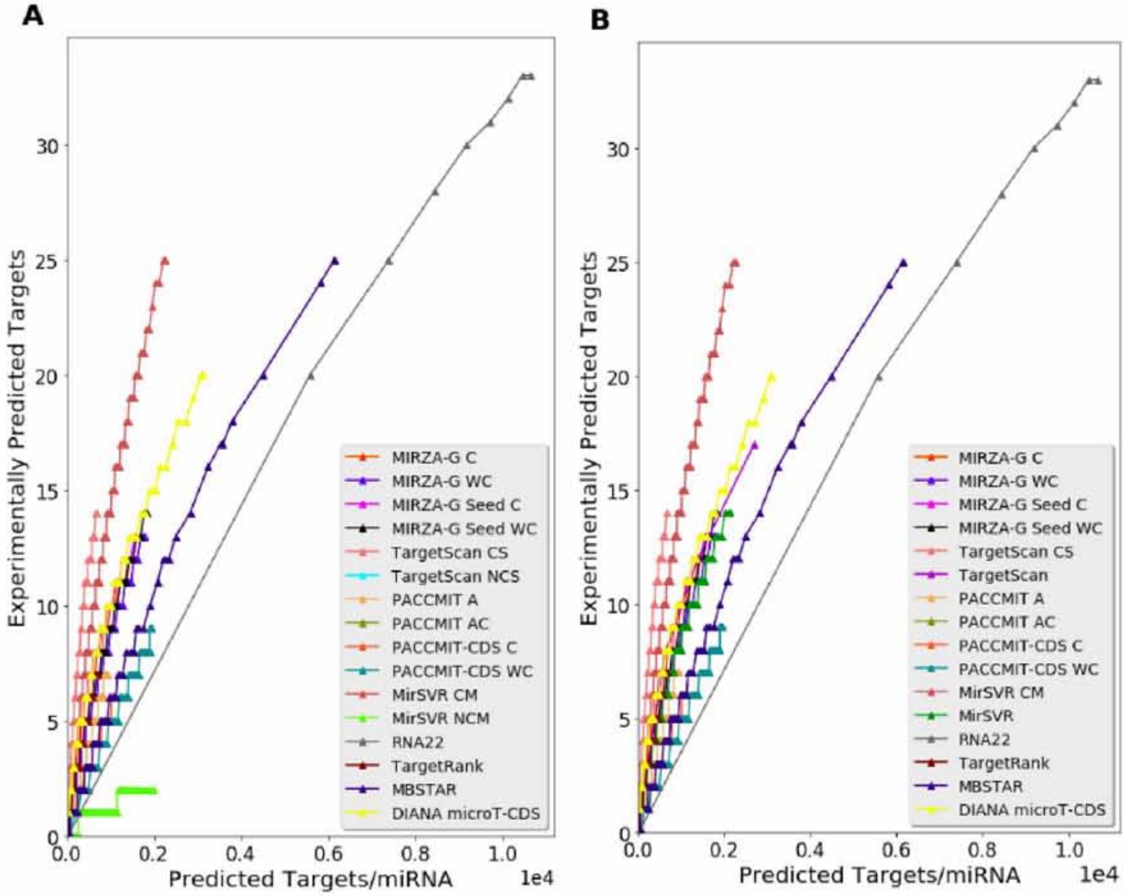


Figure 29. Predicted Targets/miRNA vs Experimentally Predicted Targets by setting threshold on RNA22 and MBSTAR. The number of correctly predicted targets is shown by different scores for increasing numbers of predicted targets per miRNA, when considering a step of 0.01. All predictions were filtered on the true positive set of miRNAs. Between miRNA-gene interactions with the same score, those with the maximum scoring scheme are selected (Test Case I). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.

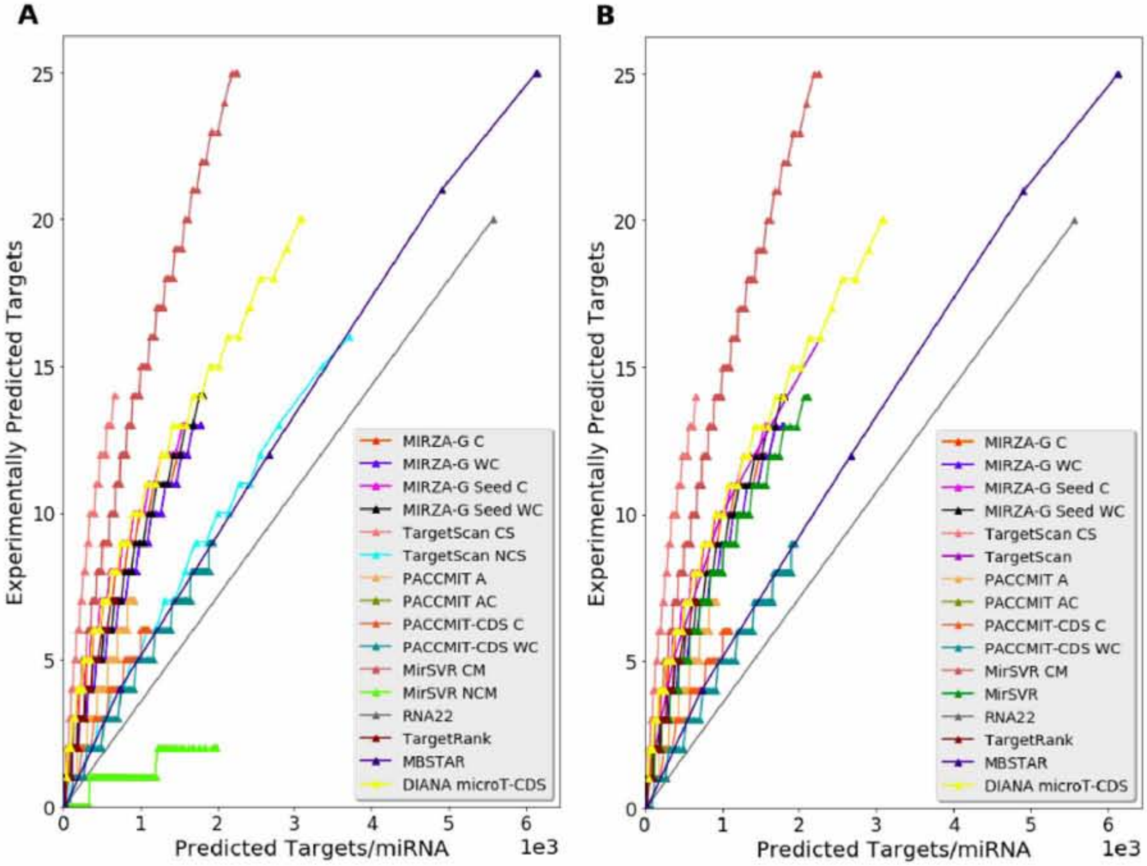
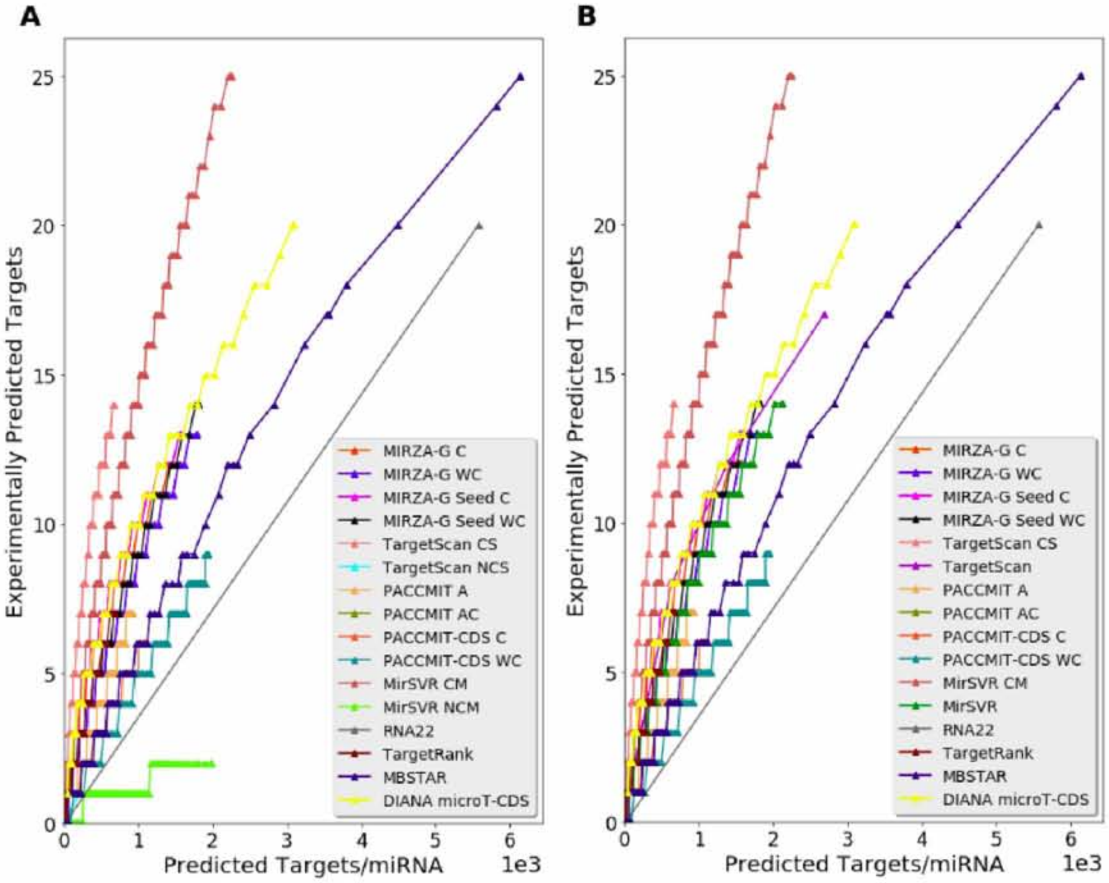


Figure 30. Predicted Targets/miRNA vs Experimentally Predicted Targets by setting threshold on RNA22 and MBSTAR. The number of correctly predicted targets is shown by different scores for increasing numbers of predicted targets per miRNA, when considering a step of 0.01. All predictions were filtered on the true positive set of miRNAs. Between miRNA-gene interactions with the same score, the aggregation of their scoring scheme is selected (Test Case I). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.



In **Figure 31**, it is observed that all algorithms share 142,821 common Total predictions. DIANA microT-CDS contains individually 536,301 total predictions while all algorithms except TargetScan CS share 271,861 predictions. In addition, MIRZA-G C, MIRZA-G Seed C and DIANA microT-CDS share 143,205 total predictions.

In **Figure 32**, it is shown that DIANA microT-CDS, TargetScan and mirSVR share 760,247 common Total predictions. In particular, DIANA microT-CDS and TargetScan contain 349,389 shared initial predictions, while TargetScan has individually 326,900 Total predictions.

In **Figure 33**, all algorithms share 2,893 correctly verified miRNA-gene interactions, which is a subset of Test Dataset 1. DIANA-microT-CDS predicts individually 2,328 miRNA-gene interactions while all algorithms except TargetScan CS predict 1,457 common interactions.

Figure 31. Venn Diagram of the total predictions between MirSVR Conserved miRNAs, TargetScan Conserved Sites, MIRZA-G Mirza with conservation, DIANA-microT-CDS and MIRZA-G Seed with conservation. These algorithms present the best performance according to previous comparisons (Test Case I).

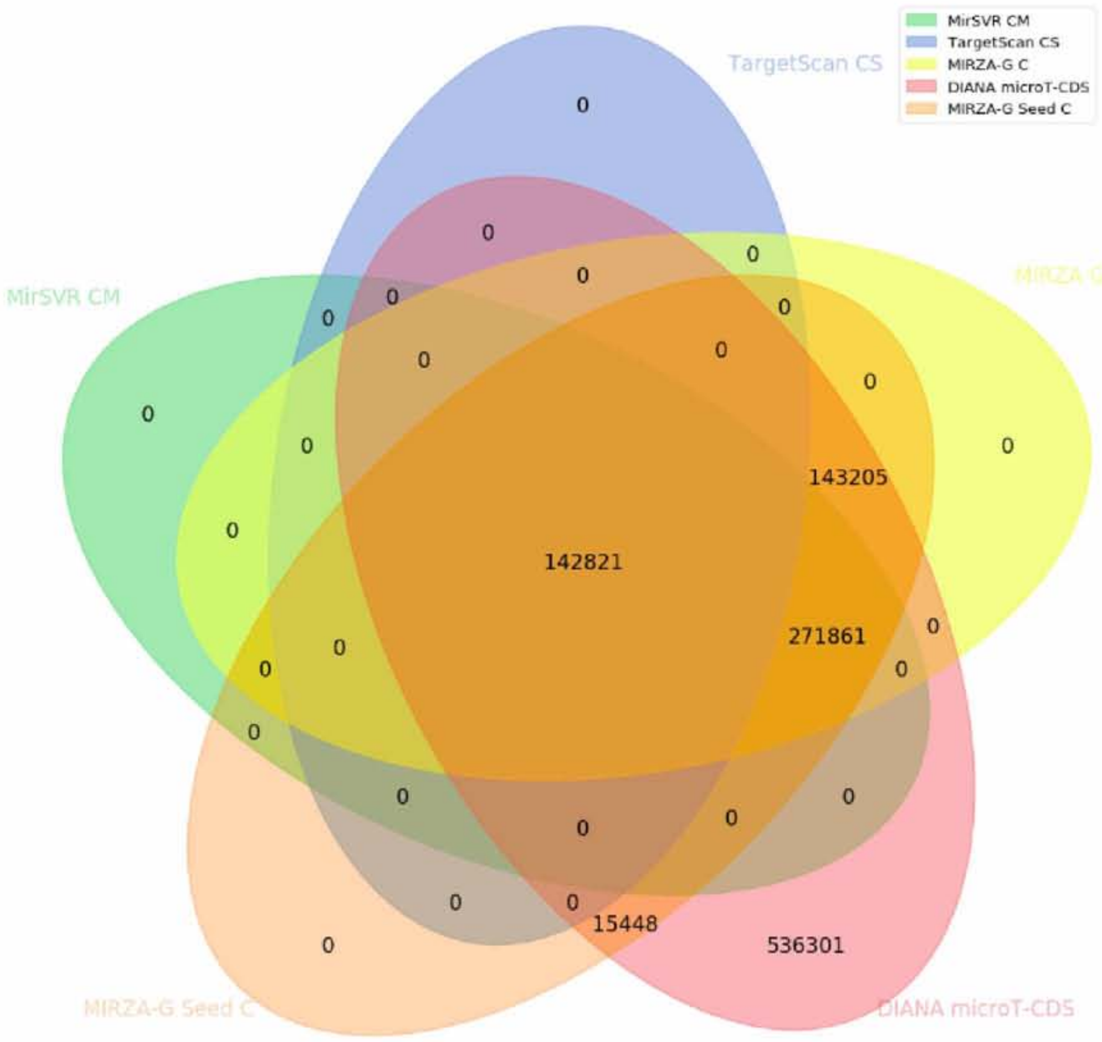


Figure 32. Venn Diagram of the total predictions between MirSVR, TargetScan and DIANA microT-CDS. These algorithms do not differentiate conserved and non conserved predictions (Test Case I).

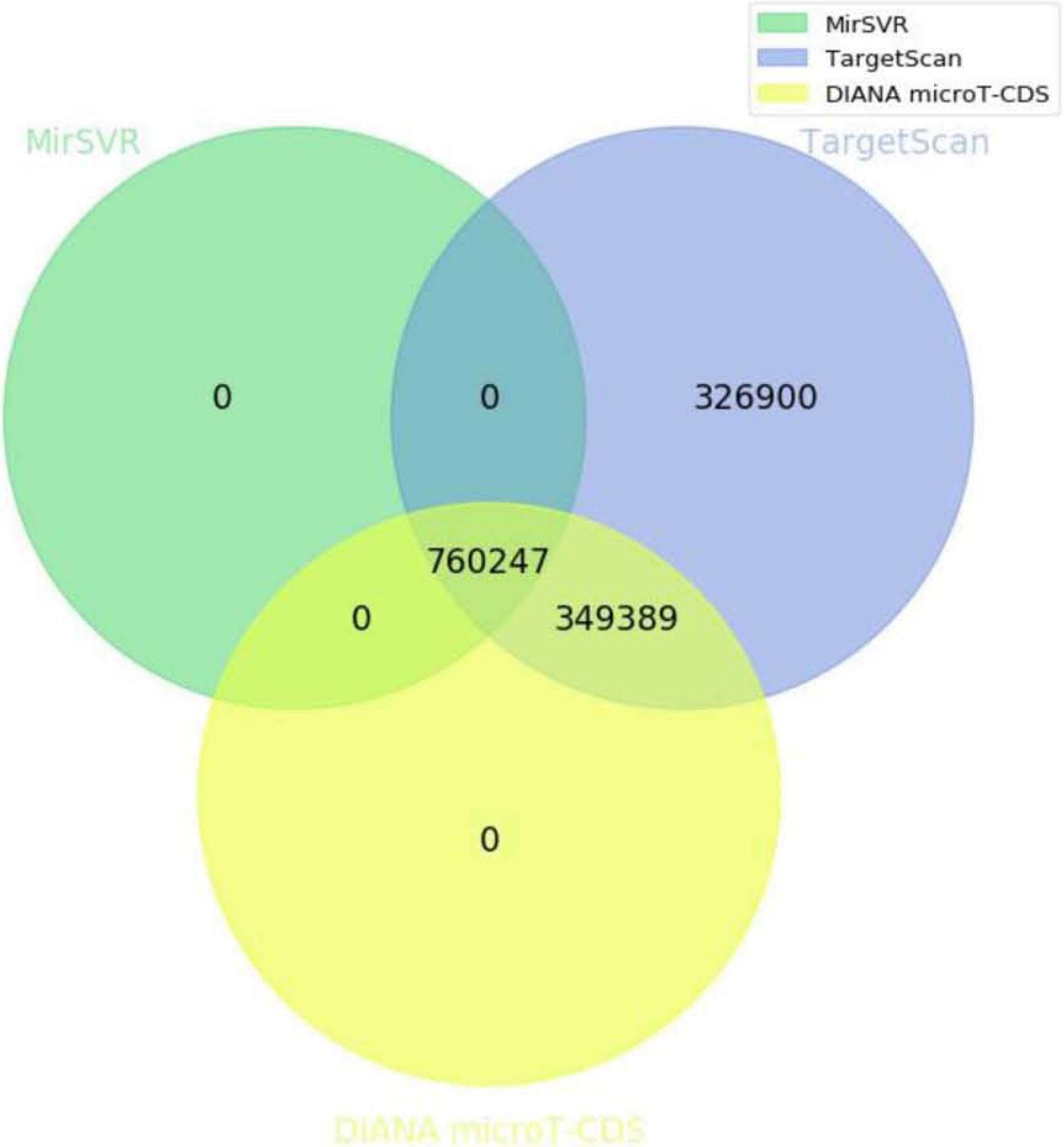
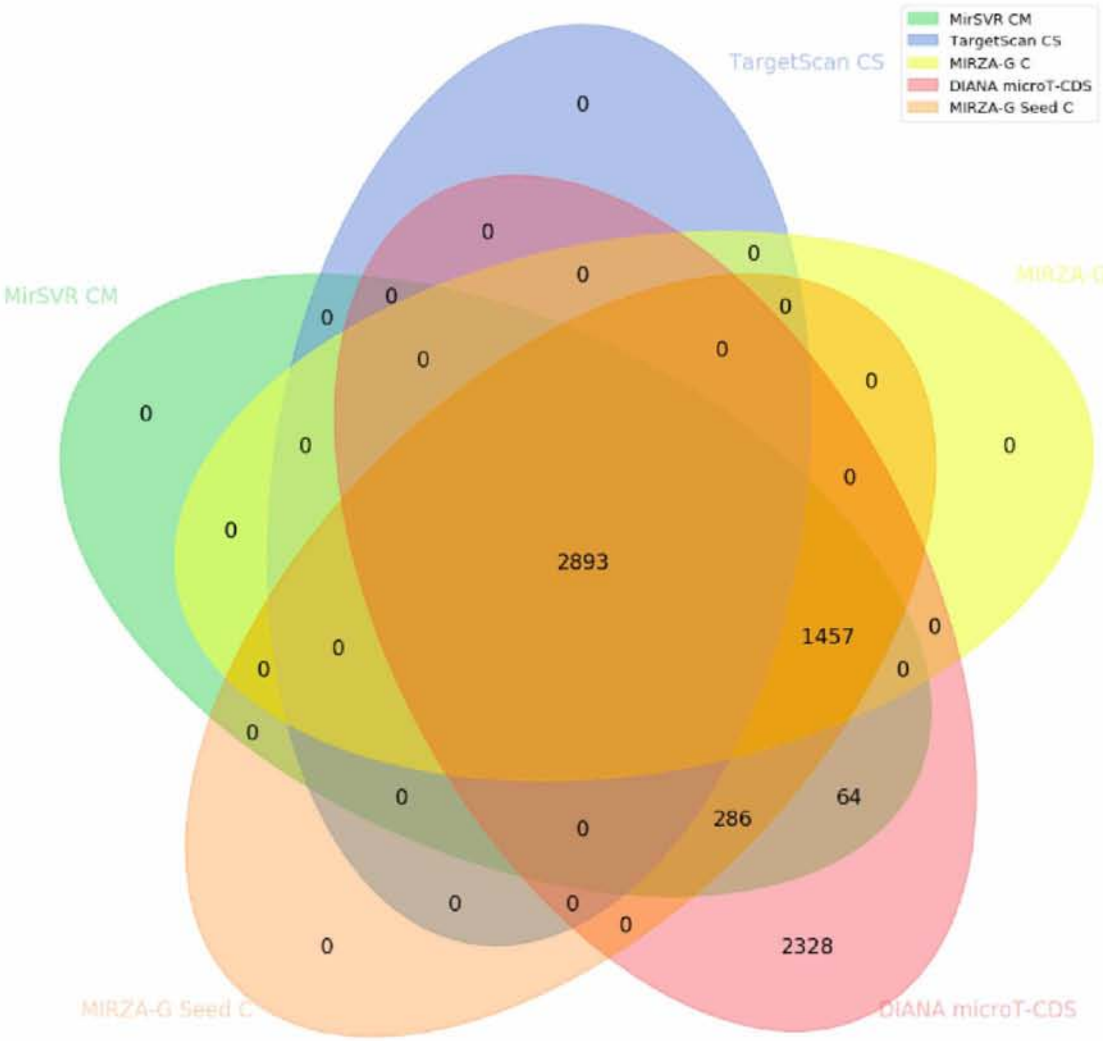
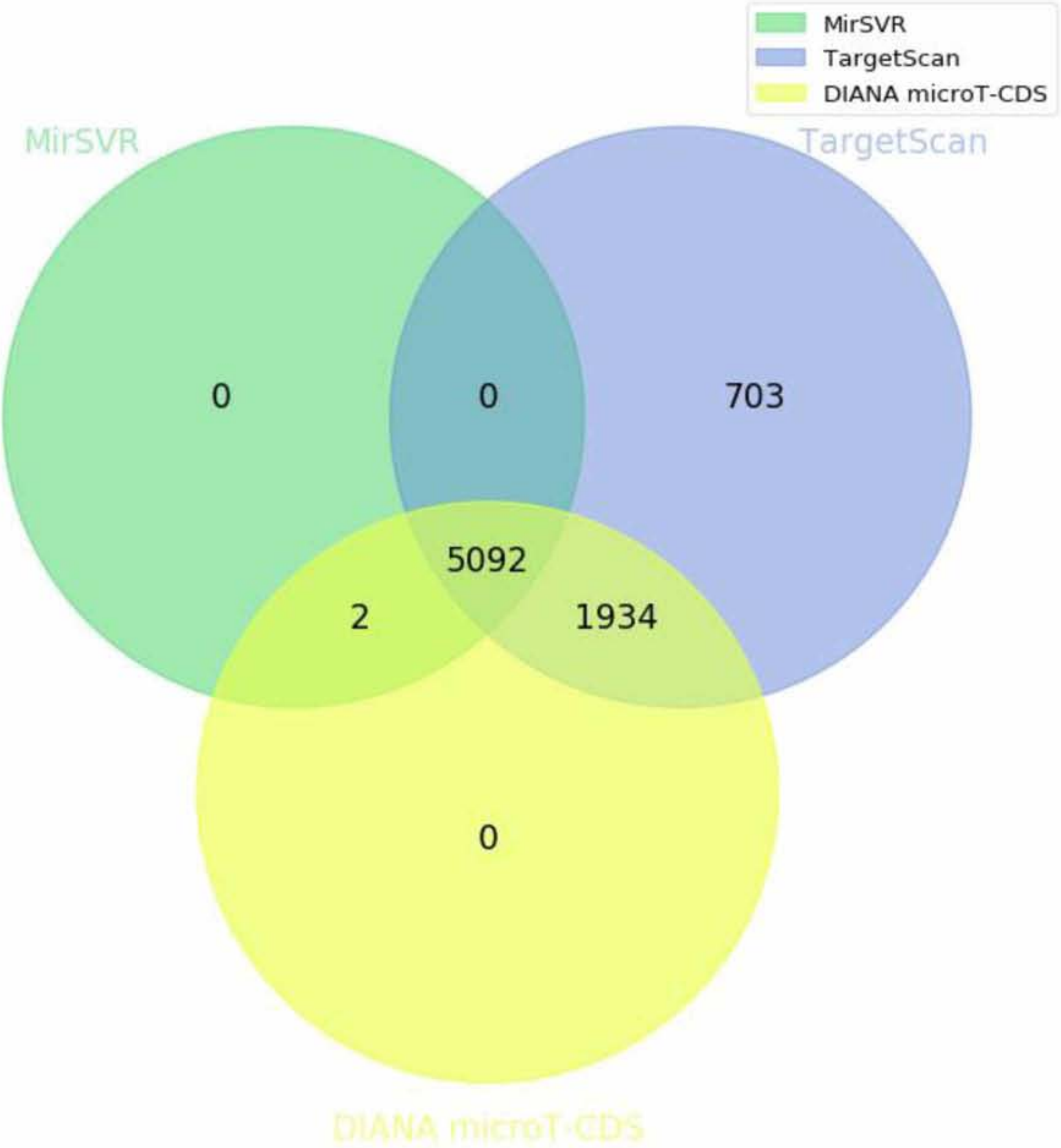


Figure 33. Venn Diagram of the experimentally verified predictions between MirSVR Conserved miRNAs, TargetScan Conserved Sites, MIRZA-G Mirza with conservation, DIANA-microT-CDS and MIRZA-G Seed with conservation. These algorithms present the best performance according to previous comparisons (Test Case I).



In **Figure 34**, it is shown that DIANA microT-CDS, TargetScan and mirSVR share 5,092 common experimentally validated miRNA-gene interactions. Indeed, DIANA microT-CDS and TargetScan intersect in 1,934 interactions, while TargetScan predicts individually 703 miRNA-gene targets. Consequently, it is obvious that TargetScan can forecast correctly both its and other algorithms's interactions, concluding that the combination of itself with other miRNA target prediction programs, would not enhance its performance.

Figure 34. Venn Diagram of the experimentally verified predictions between MirSVR, TargetScan and DIANA microT-CDS. These algorithms do not differentiate conserved and non conserved predictions (Test Case I).



Figures 32, **34** are implemented with the purpose of examining the performance of the most optimal target prediction tools.

4.2 Test Case II

The predictions of each program were filtered by the common set of miRNAs among all programs that is subset of the positive set and the common set of Genes among all programs. A threshold of P-value < 0.05 for RNA22, P-value ≥ 0.5 for MBSTAR and P-value ≥ 0.5 for DIANA microT-CDS, is considered.

In **Figure 35**, Total predictions and True positive experimentally validated predictions (shared miRNA-Genes interactions with Test Dataset 1) of all programs are shown in blue and yellow respectively, after having applied threshold for RNA22 and MBSTAR algorithms. The results are calculated without taking into account the score of each miRNA-gene interaction. The threshold for DIANA microT-CDS has been applied to all test cases. It is evident that TargetScan outperforms the other target prediction programs, while RNA22 has a vast number of predictions and therefore is highly sensitive.

Figure 35. Total and True positive set for Target Prediction Algorithms without considering the score and setting threshold on RNA22 and MBSTAR according to Test Dataset 1 (Test Case II).

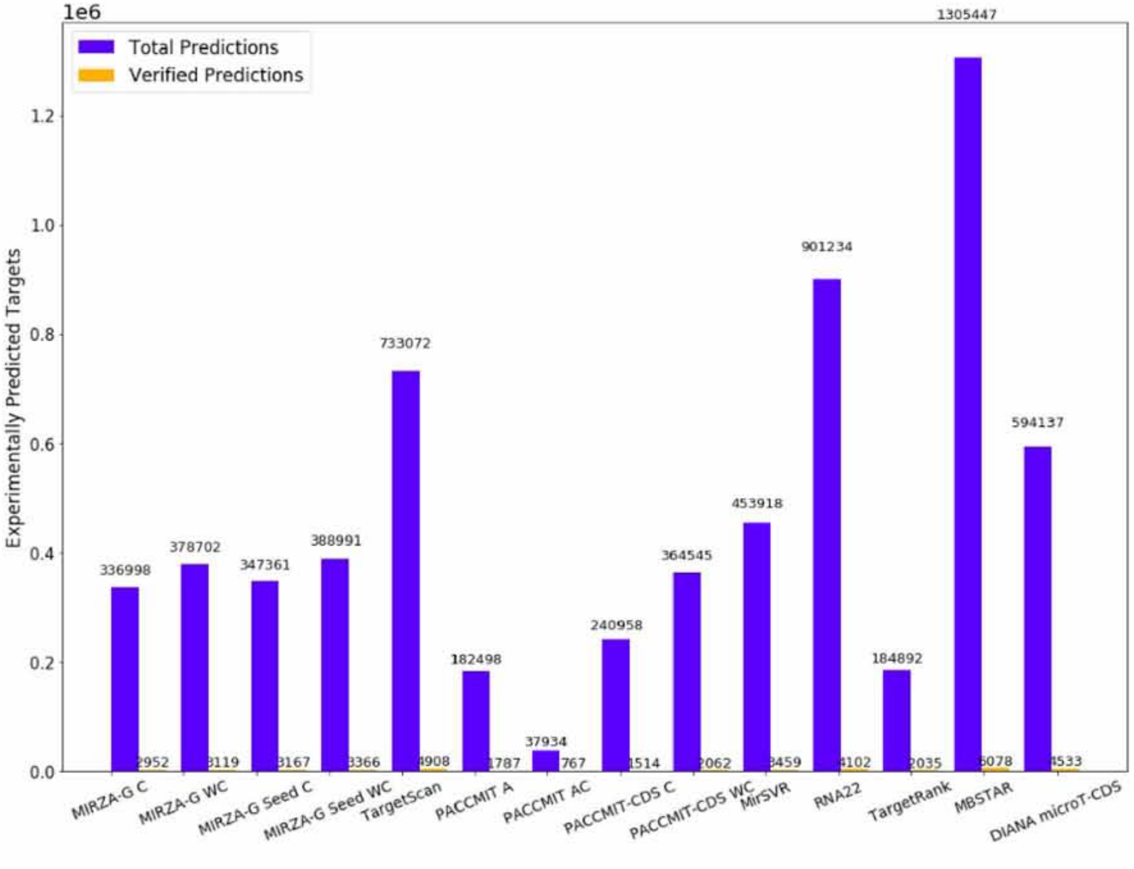


Figure 36 illustrates Total Predictions and Experimentally Predicted Targets of all programs for corresponding score values. The number of correctly predicted targets is shown by different scores for increasing numbers of total predictions. All predictions, whether conserved, non-conserved or total are filtered on the true positive set of miRNAs and the common set of genes among all programs. Also, the predictions of RNA22 with P-value ≥ 0.05 have been cut off. As a result, for a relatively large number of Total predictions, RNA22 achieves a larger number of experimentally verified predictions compared to TargetScan CS. As far as MBSTAR is concerned, the application of the threshold of 0.5 do not alter its predictions, as all miRNA-gene interactions have scores greater or equal to 0.5 prior reaching the stage of implementing the aforementioned threshold.

In **Figure 36A**, it is evident that conserved predictions present an optimized performance compared to non-conserved predictions. In addition, it is observed that TargetScan CS has the optimal performance when compared to the other target prediction algorithms due to the fact that it achieves high accuracy and its sensitivity is almost nonexistent. Indeed, the curve that corresponds to Targetscan CS initially starts with a relatively small number of Total predictions and maps them to a large number of experimentally verified predictions. Hierarchically, following TargetScan CS, MirSVR CM, DIANA microT-CDS, MIRZA-G Seed C and MIRZA-G C present the greatest performance among target prediction tools. MBSTAR appears to be very sensitive due to the fact that, although they find a larger number of experimentally supported targets, compared to other programs, at the beginning their Total predictions are proportionally very high. PACCMIT AC, PACCMIT-CDS C, PACCMIT A and PACCMIT-CDS WC seem to hold a small number of Total predictions leading to a tiny number of verified targets.

From **Figure 36B**, conserved miRNA:gene human interactions outperform the Total interactions due to the fact that Total predictions constitute the aggregation of conserved and non-conserved interactions. For instance, TargetScan CS presents better performance compared to TargetScan. Moreover, conserved miRNA predictions of MirSVR (MirSVR CM) are less sensitive than the one of MirSVR.

In **Figure 36**, between miRNA-gene interactions with the same score, the maximum of their scoring scheme is selected. On the other hand, in **Figure 37**, the aggregated score of the corresponding interactions is considered. After careful examination of the plots in the latter case, it is observed that the performance of the algorithms remains intact despite the application of the aggregation filter, with the exception of TargetScan NCS. This is due to the fact that TargetScan assigns a score in each site and not in the entire miRNA-gene interaction as other target prediction programs (e.g DIANA microT-CDS).

Figure 36. Total Predictions vs Experimentally Predicted Targets of all programs when considering step = 0.01 for score values filtered on the true positive set of miRNAs and the common set of genes among algorithms. A threshold on RNA22 and MBSTAR is set. Between miRNA-gene interactions with the same score, those with the maximum scoring scheme are selected (Test Case II). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.

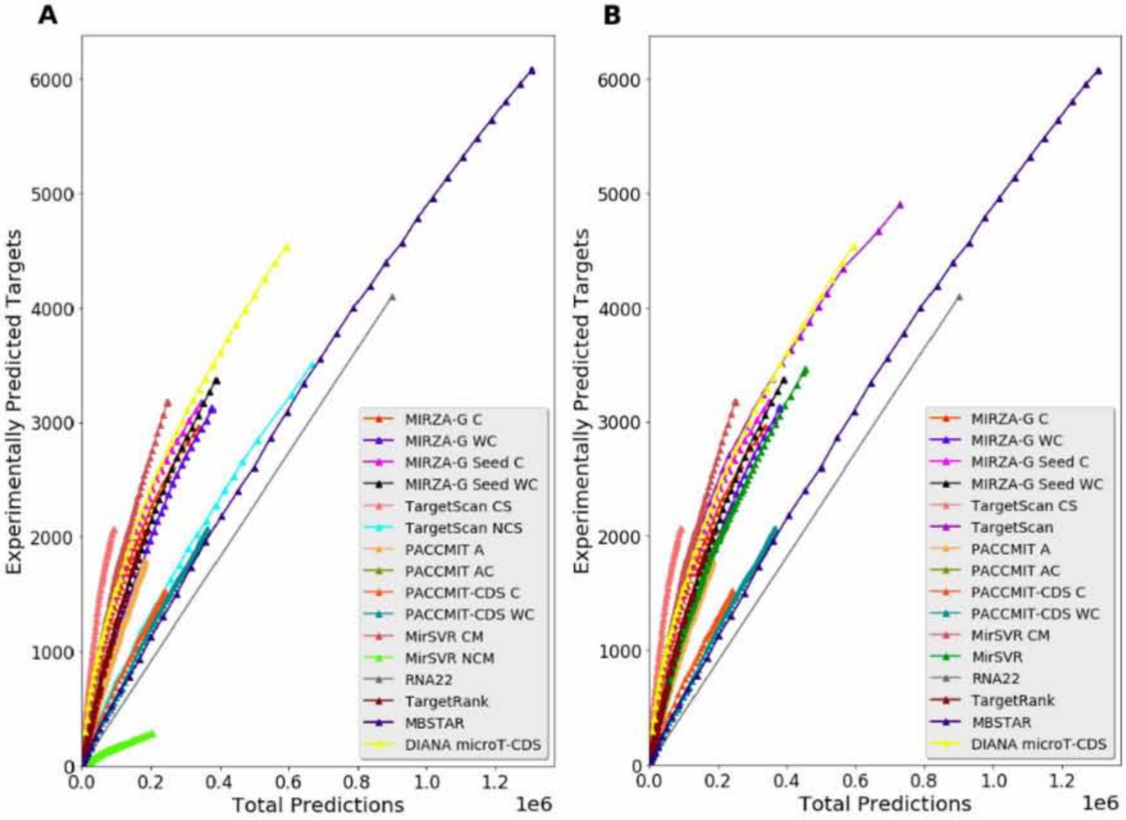


Figure 37. Total Predictions vs Experimentally Predicted Targets of all programs when considering step = 0.01 for score values filtered on the true positive set of miRNAs and the common set of genes among algorithms. A threshold on RNA22 and MBSTAR is set. Between miRNA-gene interactions with the same score, the aggregation of their scoring scheme is selected (Test Case II). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.

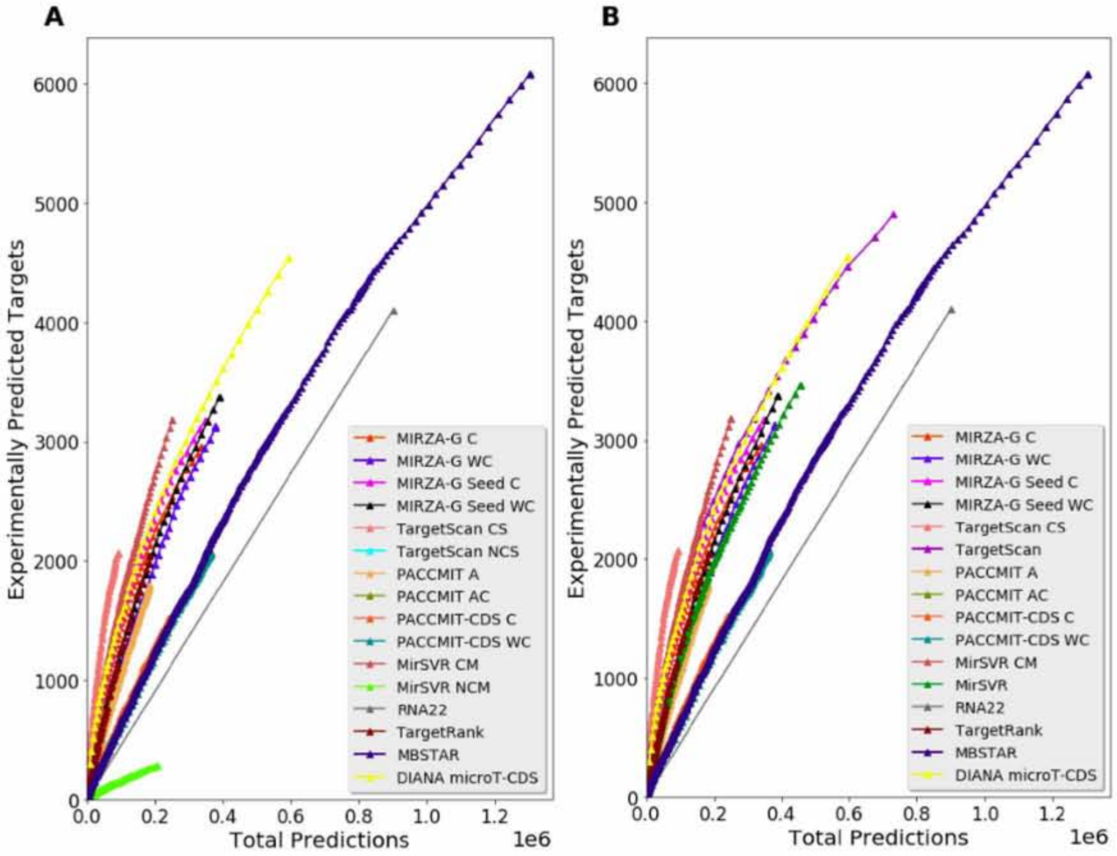
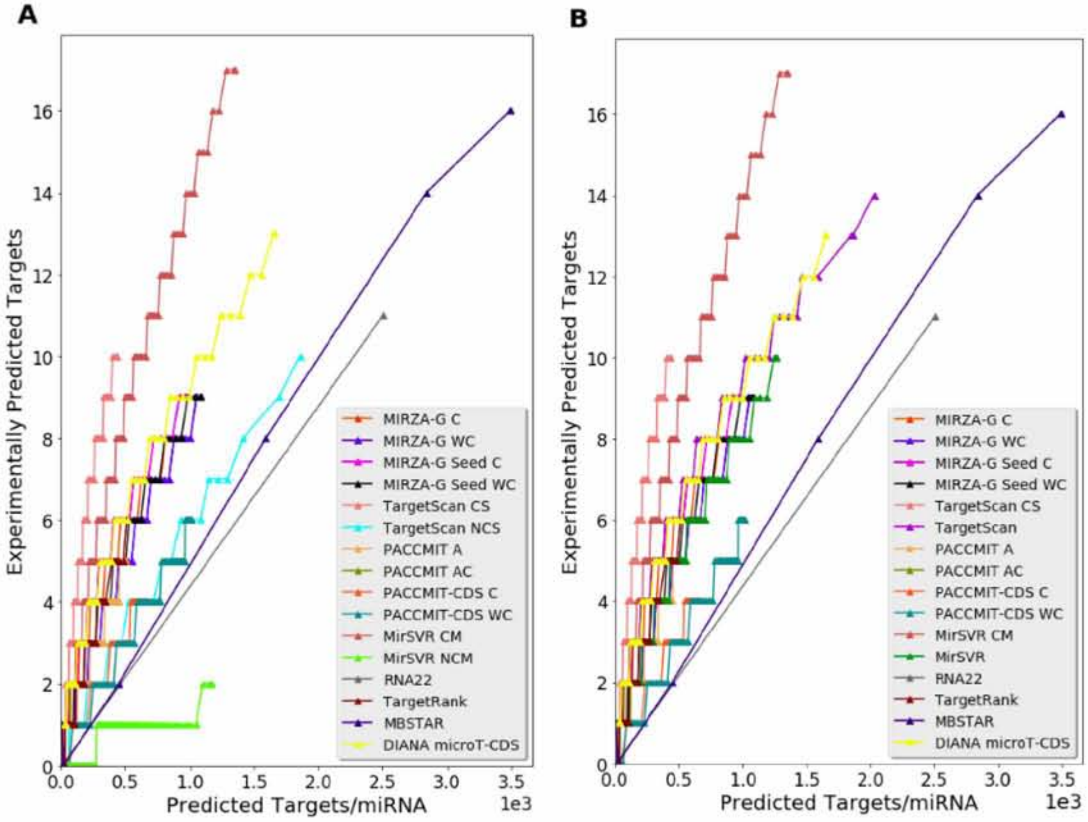


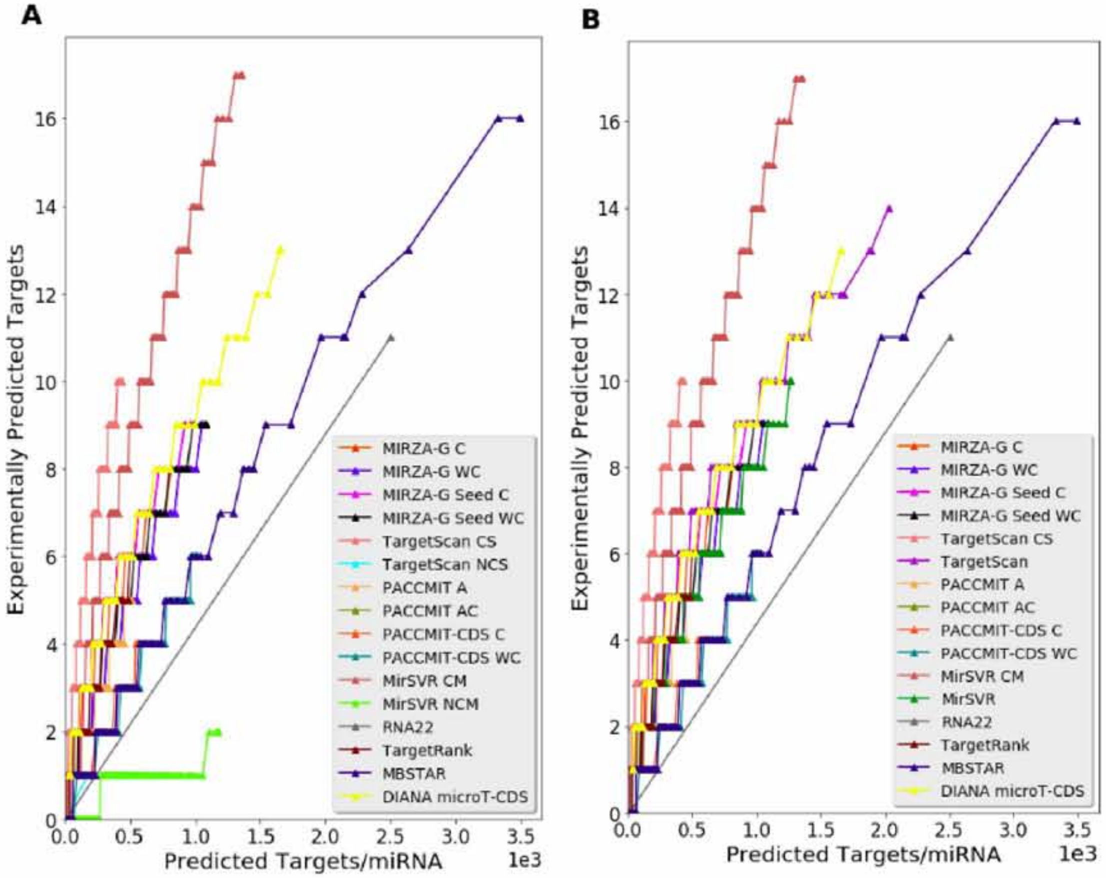
Figure 38 shows the Predicted Targets per miRNA and the corresponding number of correctly predicted targets for different scores, after thresholds for RNA22 and MBSTAR have been employed. Predicted Targets per miRNA constitute the average value of Total predictions of all miRNAs in each program, having been grouped by score and thery indicate the sensitivity of the examined models. Correctly predicted targets are calculated as the average value of experimentally verified targets of all miRNAs in each program, having been grouped by score. **Figures 36 and 38** share the same conclusions as far as the performance of target prediction algorithms is concerned.

Figure 38. Predicted Targets/miRNA vs Experimentally Predicted Targets by setting threshold on RNA22 and MBSTAR. The number of correctly predicted targets is shown by different scores for increasing numbers of predicted targets per miRNA, when considering a step of 0.01. All predictions were filtered on the true positive set of miRNAs and the common set of genes among algorithms. Between miRNA-gene interactions with the same score, those with the maximum scoring scheme are selected (Test Case II). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.



In **Figure 39**, despite the selection of the aggregated score of the corresponding miRNA-gene interactions, all algorithms present the same performance as prior to the application of this filter. The selection of the aggregated score of the corresponding miRNA-gene interactions only alters the performance of TargetScan NCS, due to the existence of site score instead of score for the entire miRNA-gene interaction.

Figure 39. Predicted Targets/miRNA vs Experimentally Predicted Targets by setting threshold on RNA22 and MBSTAR. The number of correctly predicted targets is shown by different scores for increasing numbers of predicted targets per miRNA, when considering a step of 0.01. All predictions were filtered on the true positive set of miRNAs and the common set of genes among algorithms. Between miRNA-gene interactions with the same score, the aggregation of their scoring scheme is selected (Test Case II). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.



In **Figure 40**, it is observed that all algorithms share 91,843 common Total predictions. DIANA microT-CDS contains individually 246,776 total predictions while all algorithms except TargetScan CS share 157,585 predictions. In addition, MIRZA-G C, MIRZA-G Seed C and DIANA microT-CDS share 87,570 total predictions.

In **Figure 41**, it is shown that DIANA microT-CDS, TargetScan and mirSVR share 453,918 common Total predictions. In particular, DIANA microT-CDS and TargetScan contain 140,219 shared initial predictions, while TargetScan has individually 138,935 Total predictions.

Figure 40. Venn Diagram of the total predictions between MirSVR Conserved miRNAs, TargetScan Conserved Sites, MIRZA-G Mirza with conservation, DIANA microT-CDS and MIRZA-G Seed with conservation. These algorithms present the best performance according to previous comparisons (Test Case II).

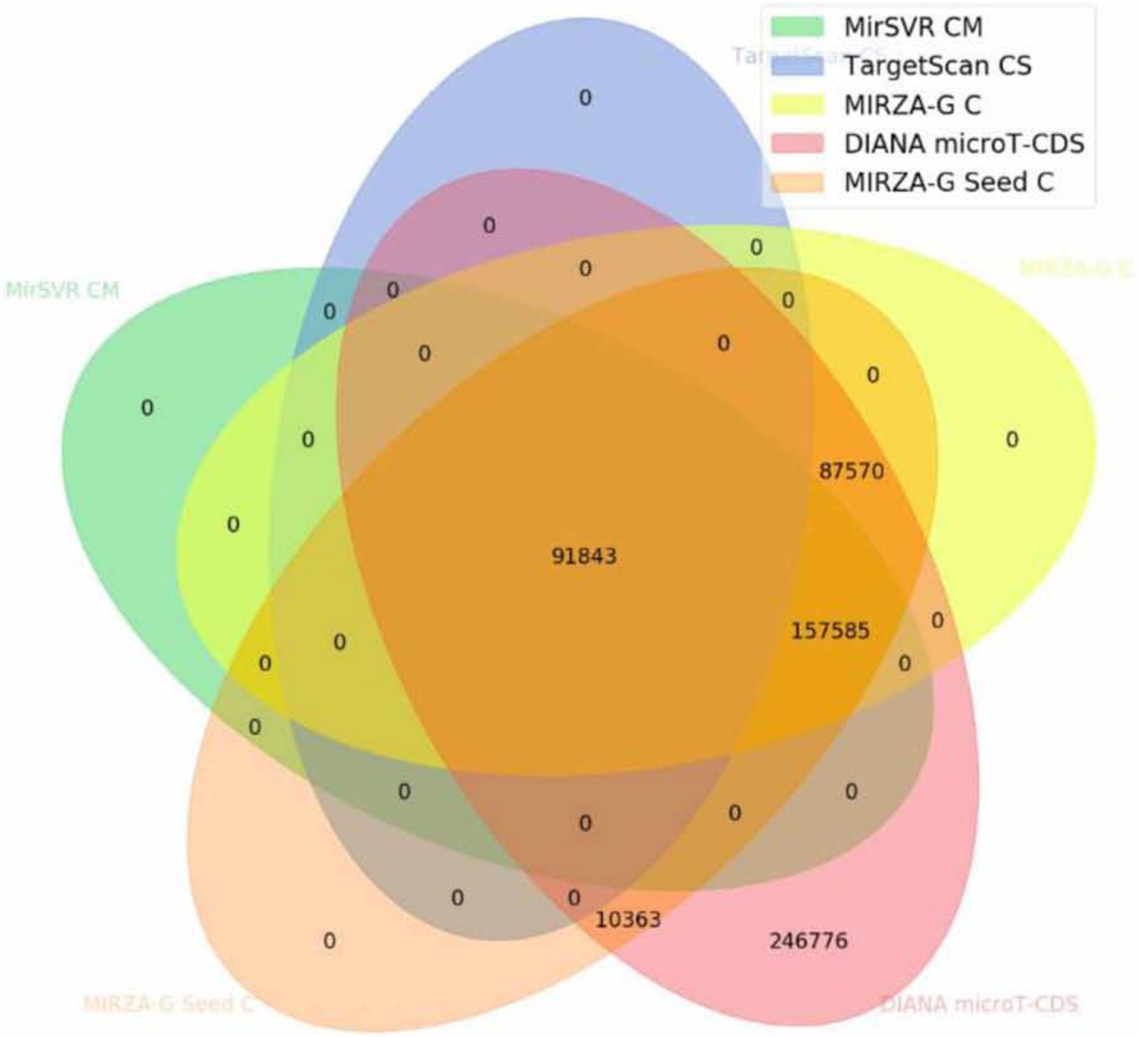
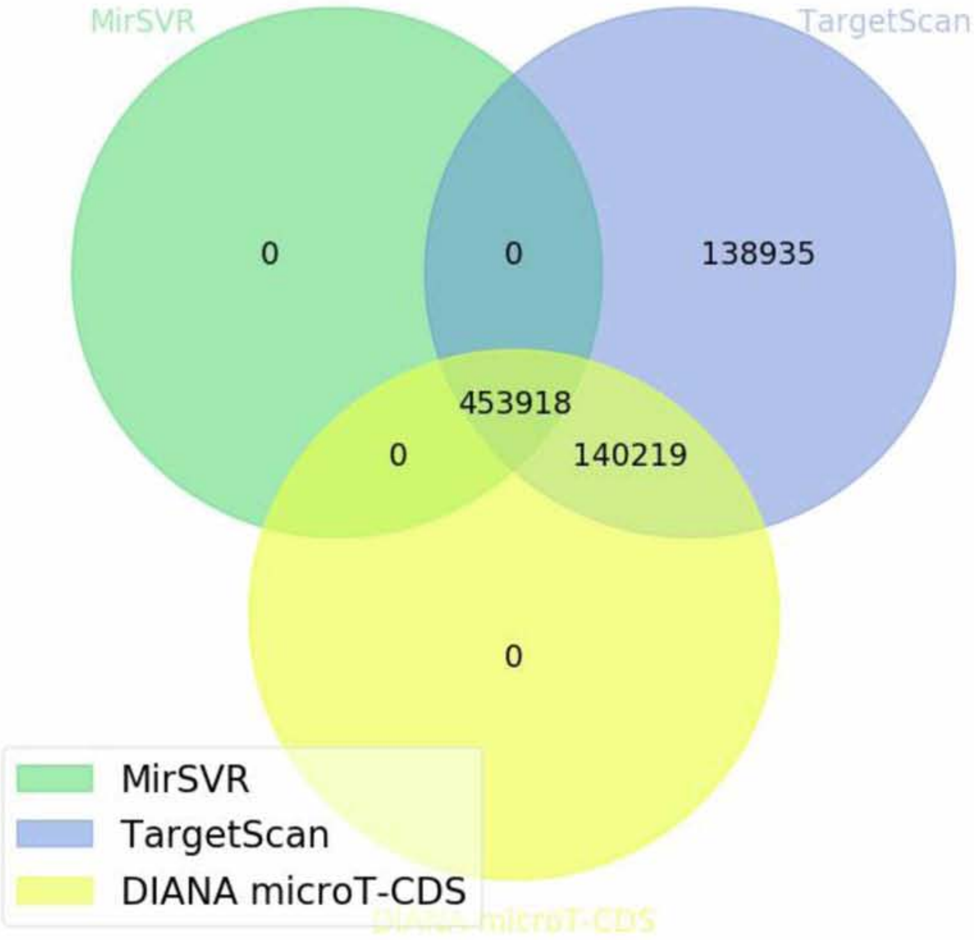


Figure 41. Venn Diagram of the total predictions between MirSVR, TargetScan and DIANA microT-CDS. These algorithms do not differentiate conserved and non conserved predictions (Test Case II).



In **Figure 42**, all algorithms share 2,067 correctly verified miRNA-gene interactions, which is a subset of Test Dataset 1. DIANA microT-CDS predicts individually 1,354 miRNA-gene interactions while all algorithms except TargetScan CS predict 885 common interactions.

In **Figure 43**, it is shown that DIANA microT-CDS, TargetScan and mirSVR share 3,457 common experimentally validated miRNA-gene interactions. Indeed, DIANA microT-CDS and TargetScan intersect in 1,074 interactions, while TargetScan predicts individually 377 miRNA-gene targets. Consequently, it is obvious that TargetScan can forecast correctly both its and other algorithms's interactions, concluding that the combination of itself with other miRNA target prediction programs would not enhance its performance.

Figure 42. Venn Diagram of the experimentally verified predictions between MirSVR Conserved miRNAs, TargetScan Conserved Sites, MIRZA-G Mirza with conservation, DIANA microT-CDS and MIRZA-G Seed with conservation. These algorithms present the best performance according to previous comparisons (Test Case II).

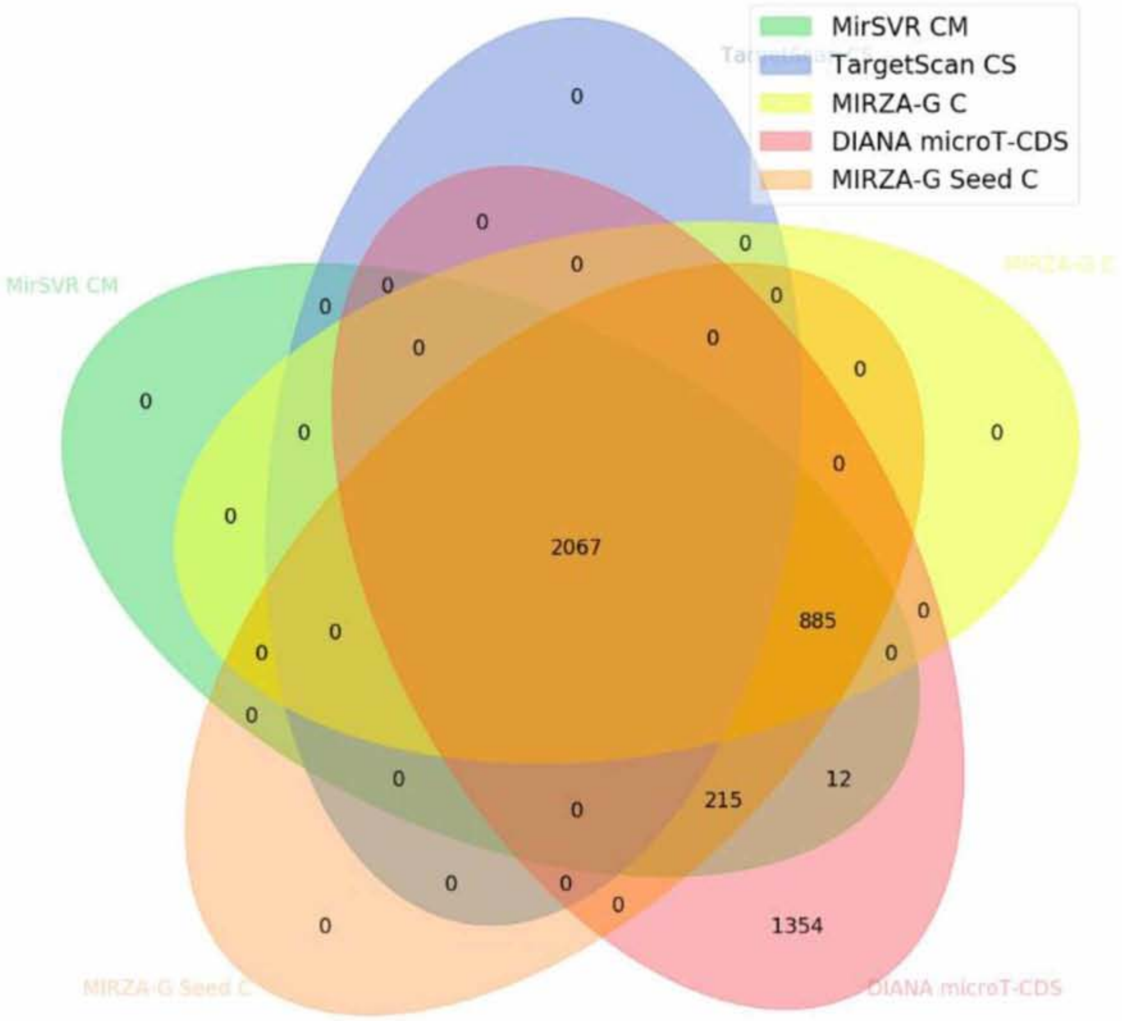
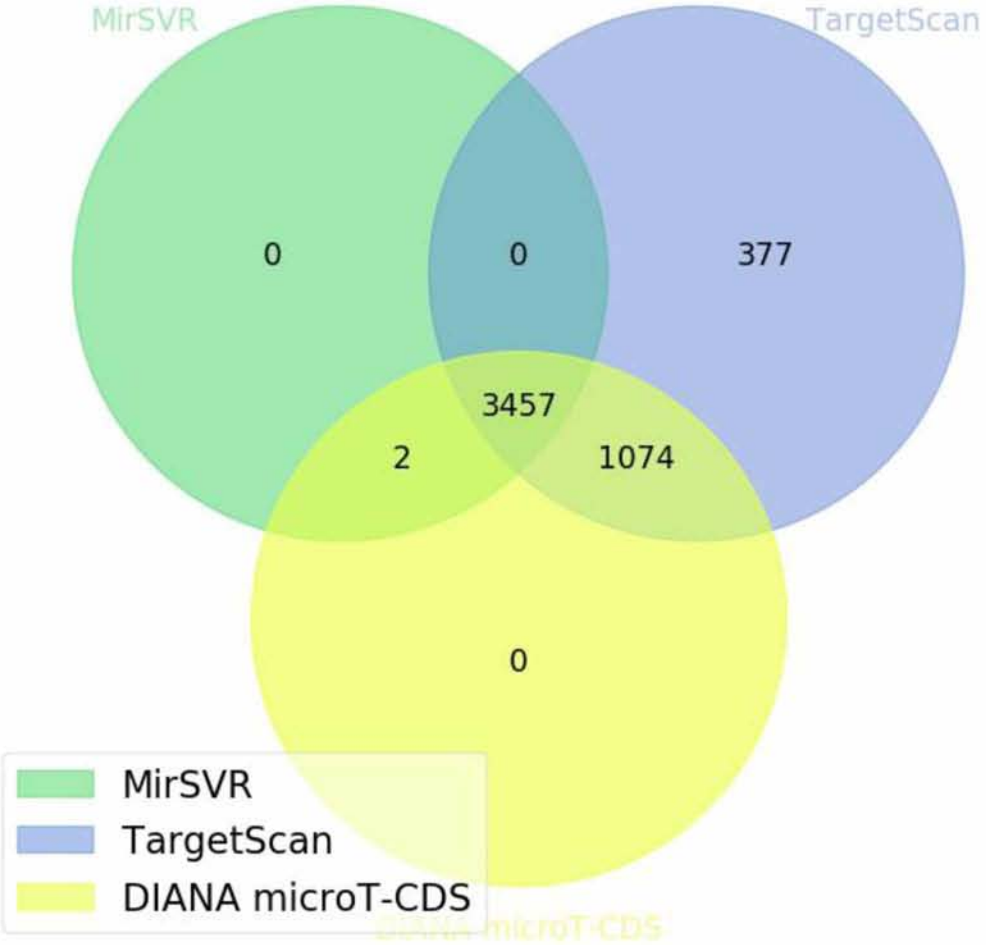


Figure 43. Venn Diagram of the experimentally verified predictions between MirSVR, TargetScan and DIANA microT-CDS. These algorithms do not differentiate conserved and non conserved predictions (Test Case II).



Figures 41, **43** are implemented with the purpose of examining the performance of the most optimal target prediction tools.

4.3 Test Case III

The predictions of each program were filtered at the common set of miRNAs among all programs that is subset of the positive set as well as the common set of Genes among all programs that is subset of the positive set. In addition, a threshold of P-value < 0.05 for RNA22, P-value ≥ 0.5 for MBSTAR and P-value ≥ 0.5 for DIANA microT-CDS is considered.

In **Figure 44**, Total predictions and True positive experimentally validated predictions (shared miRNA-Genes interactions with Test Dataset 1) of all programs are shown in blue and yellow respectively. The results are calculated without taking into account the score of each miRNA-Genes interaction**. Figure 45** illustrates the same results as **Figure 44** with the exception that thresholds for RNA22 and MBSTAR have been applied. The threshold for DIANA microT-CDS has been applied to all test cases. From both **Figures 44, 45** it is evident that TargetScan outperforms the other target prediction programs, while RNA22 has a vast number of predictions and therefore is highly sensitive.

Figure 44. Total and True positive set for Target Prediction Algorithms without considering the score according to Test Dataset 1 (Test Case III).

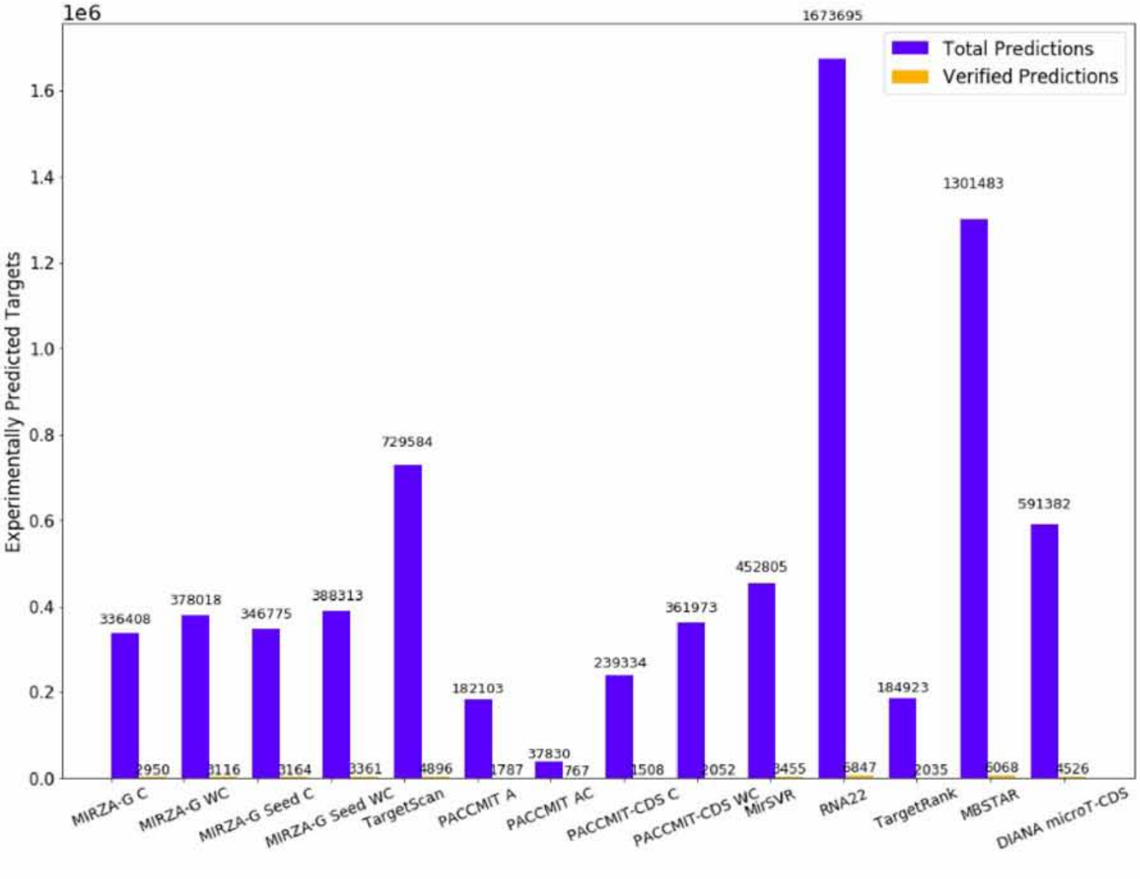


Figure 45. Total and True positive set for Target Prediction Algorithms without considering the score and setting threshold on RNA22 and MBSTAR according to Test Dataset 1 (Test Case III).

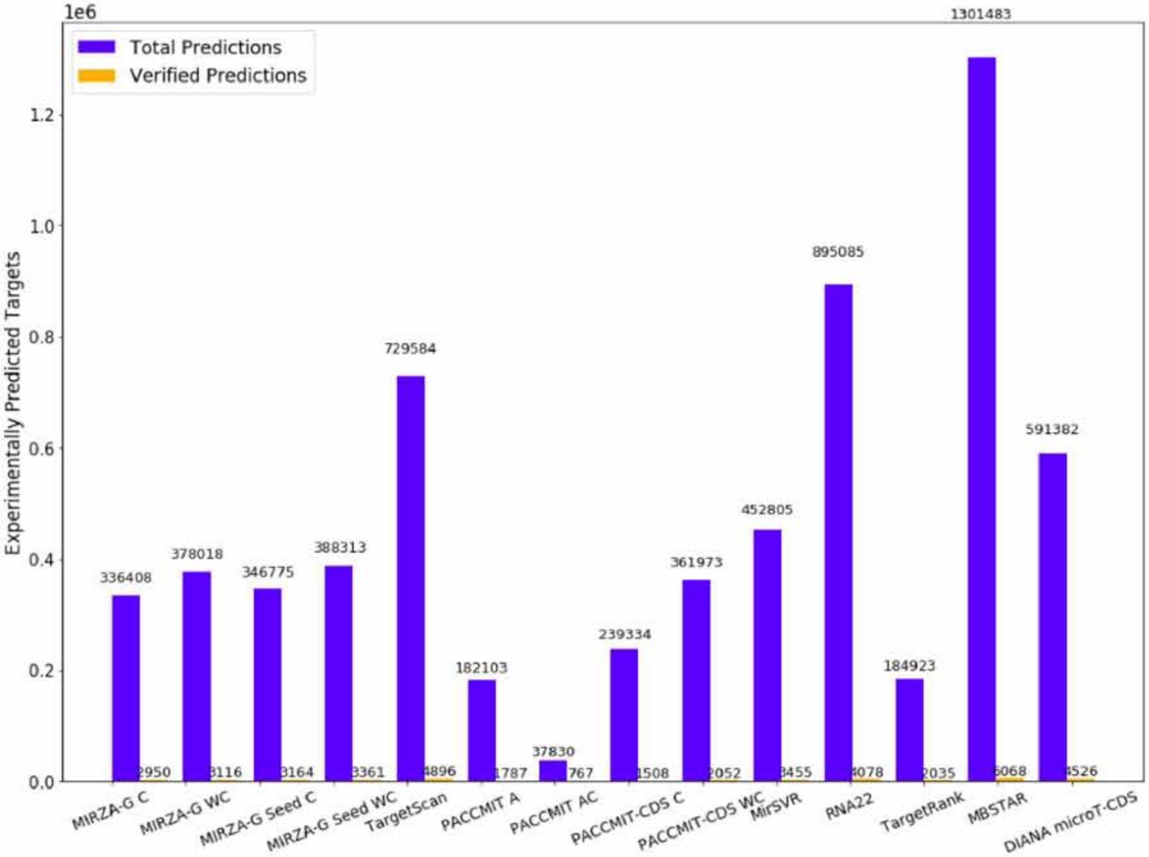
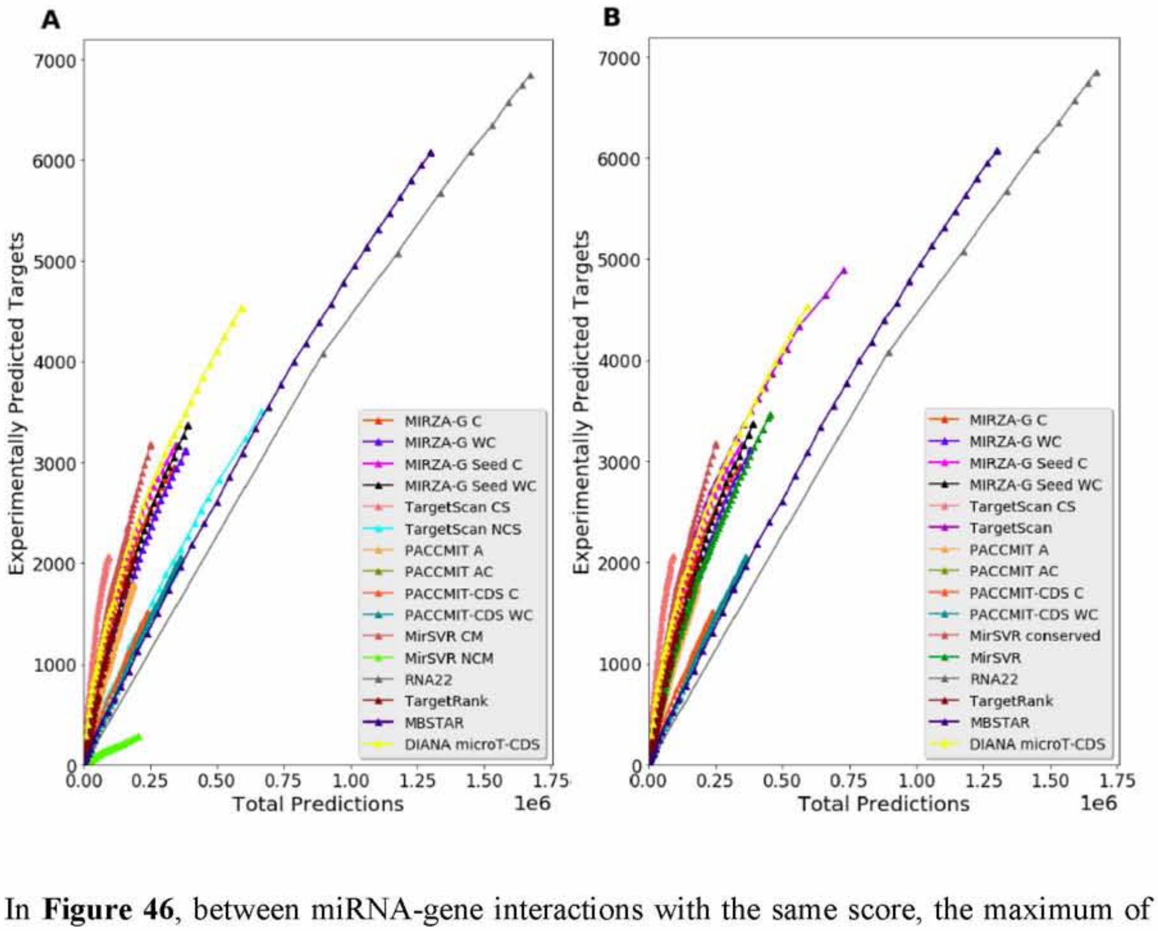


Figure 46 illustrates Total Predictions and Experimentally Predicted Targets of all programs for corresponding score values. The number of correctly predicted targets is shown by different scores for increasing numbers of total predictions. All predictions, whether conserved, non-conserved or total are filtered on the true positive set of miRNAs and the common set of genes among all programs and the positive set.

In **Figure 46A**, it is evident that conserved predictions present an optimized performance compared to non-conserved predictions. In addition, it is observed that TargetScan CS has the optimal performance when compared to the other target prediction algorithms due to the fact it achieves high accuracy and its sensitivity is almost nonexistent. Indeed, the curve that corresponds to Targetscan CS initially starts with a relatively small number of Total predictions and map them to a large number of experimentally verified predictions. Hierarchically, following TargetScan CS, MirSVR CM, DIANA-microT-CDS, MIRZA-G Seed C and MIRZA-G C present the greatest performance among target prediction tools. MBSTAR and RNA22 appear to be very sensitive due to the fact that although they find a larger number of experimentally supported targets, compared to other programs, at the beginning their Total predictions are proportionally very high. PACCMIT AC, PACCMIT-CDS C, PACCMIT A and PACCMIT-CDS WC seem to hold a small number of Total predictions leading to a tiny number of verified targets.

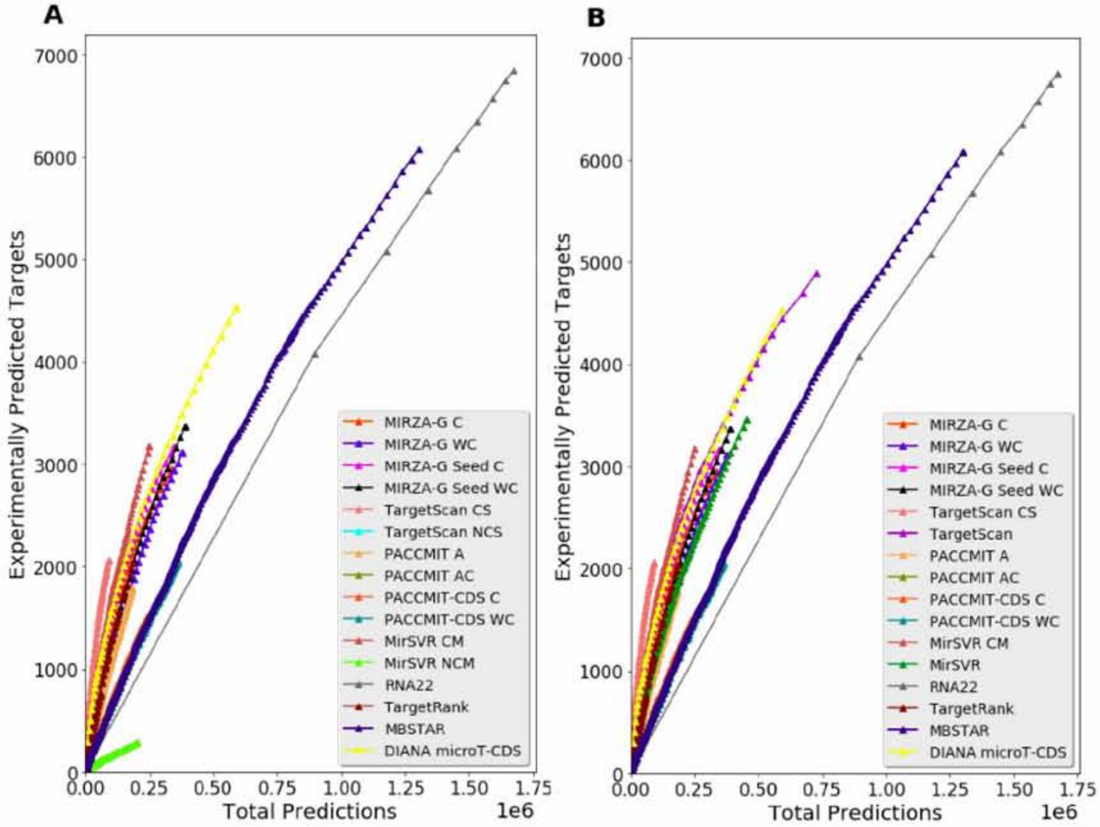
From **Figure 46B**, conserved miRNA:gene human interactions outperform the Total interactions due to the fact that Total predictions constitute the aggregation of conserved and non-conserved interactions. For instance, TargetScan CS presents better performance compared to TargetScan. Moreover, conserved miRNA predictions of MirSVR (MirSVR CM) are less sensitive than the one of MirSVR.

Figure 46. Total Predictions vs Experimentally Predicted Targets of all programs when considering step = 0.01 for score values filtered on the true positive set of miRNAs and the common set of genes among algorithms and Test Dataset 1. Between miRNA-gene interactions with the same score, those with the maximum scoring scheme are selected (Test Case III). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.



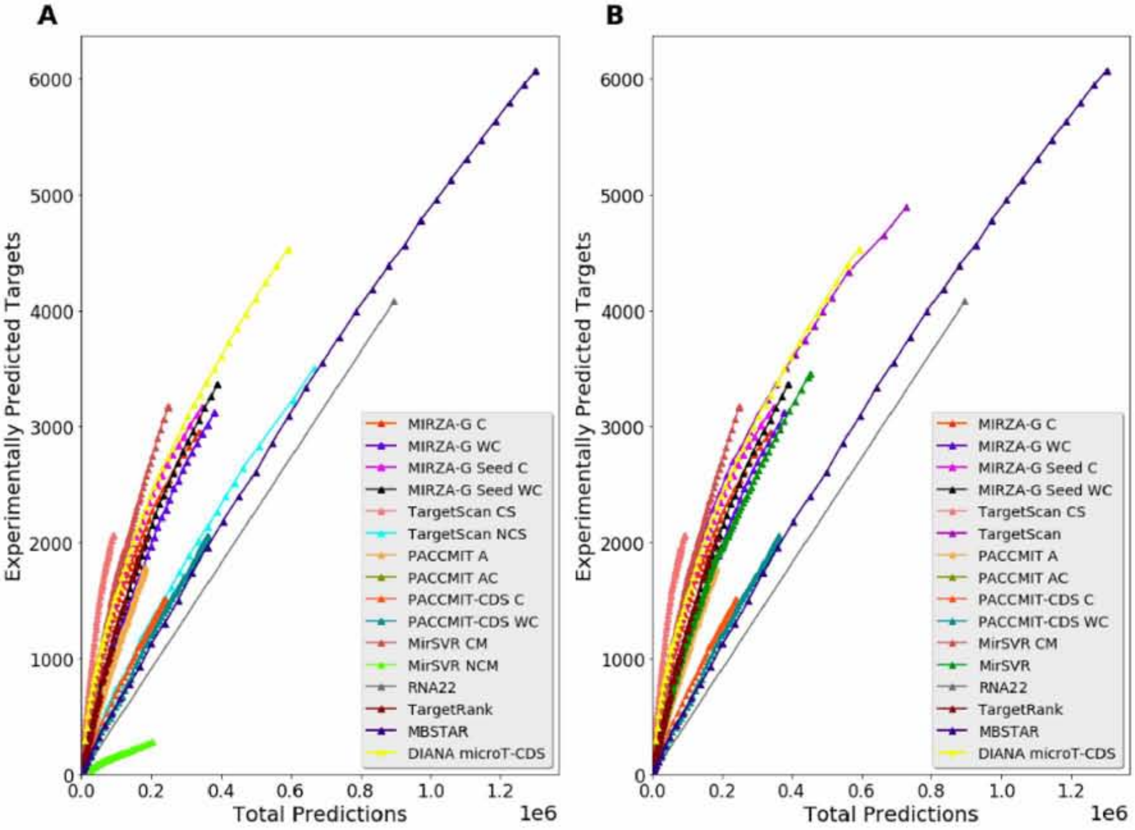
In **Figure 46**, between miRNA-gene interactions with the same score, the maximum of their scoring scheme is selected. On the other hand, in **Figure 47**, the aggregated score of the corresponding interactions is considered. After careful examination of the plots in the latter case, it is observed that the performance of the algorithms remains intact despite the application of the aggregation filter, apart from TargetScan NCS. The performance of TargetScan NCS is altered because TargetScan assigns a score in each site and not in the entire miRNA-gene interaction as other target prediction programs (e.g DIANA microT-CDS).

Figure 47. Total Predictions vs Experimentally Predicted Targets of all programs when considering step = 0.01 for score values filtered on the true positive set of miRNAs and the common set of genes among algorithms and Test Dataset 1. Between miRNA-gene interactions with the same score, the aggregation of their scoring scheme is selected (Test Case III). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.



In **Figure 48** applies the same graphical analysis as in **Figure 47**, with the sole difference that the predictions of RNA22 with P-value ≥ 0.05 have been cut off. As a result, for a relatively large number of Total predictions, RNA22 achieves a larger number of experimentally verified predictions compared to TargetScan CS. As far as MBSTAR is concerned, the application of the threshold of 0.5 do not alter its predictions, as all miRNA-Gene interactions have scores greater or equal to 0.5 prior reaching the stage of implementing the aforementioned threshold.

Figure 48. Total Predictions vs Experimentally Predicted Targets of all programs when considering step = 0.01 for score values filtered on the true positive set of miRNAs and the common set of genes among algorithms and Test Dataset 1. A threshold on RNA22 and MBSTAR has been set. Between miRNA-gene interactions with the same score, those with the maximum scoring scheme are selected (Test Case III). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.



In **Figure 48**, between miRNA-gene interactions with the same score, the max of their scoring scheme is selected. On the other hand, in **Figure 49**, the aggregated score of the corresponding interactions is considered. After careful examination of the plots in the latter case, it is observed that the performance of the algorithms remains intact despite the application of the aggregation filter, without considering TargetScan NCS algorithm.

Figure 49. Total Predictions vs Experimentally Predicted Targets of all programs when considering step = 0.01 for score values filtered on the true positive set of miRNAs and the common set of genes among algorithms and Test Dataset 1. A threshold on RNA22 and MBSTAR has been set. Between miRNA-gene interactions with the same score, the aggregation of their scoring scheme is selected (Test Case III). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.

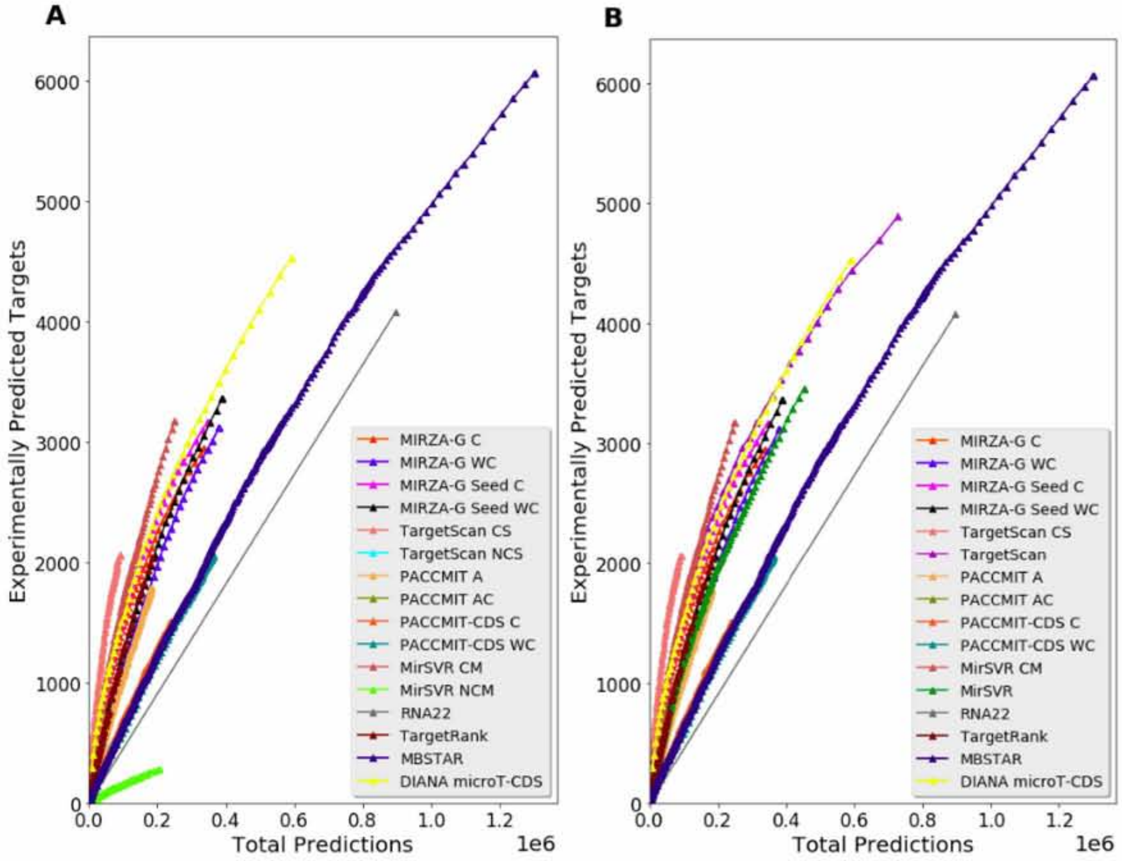
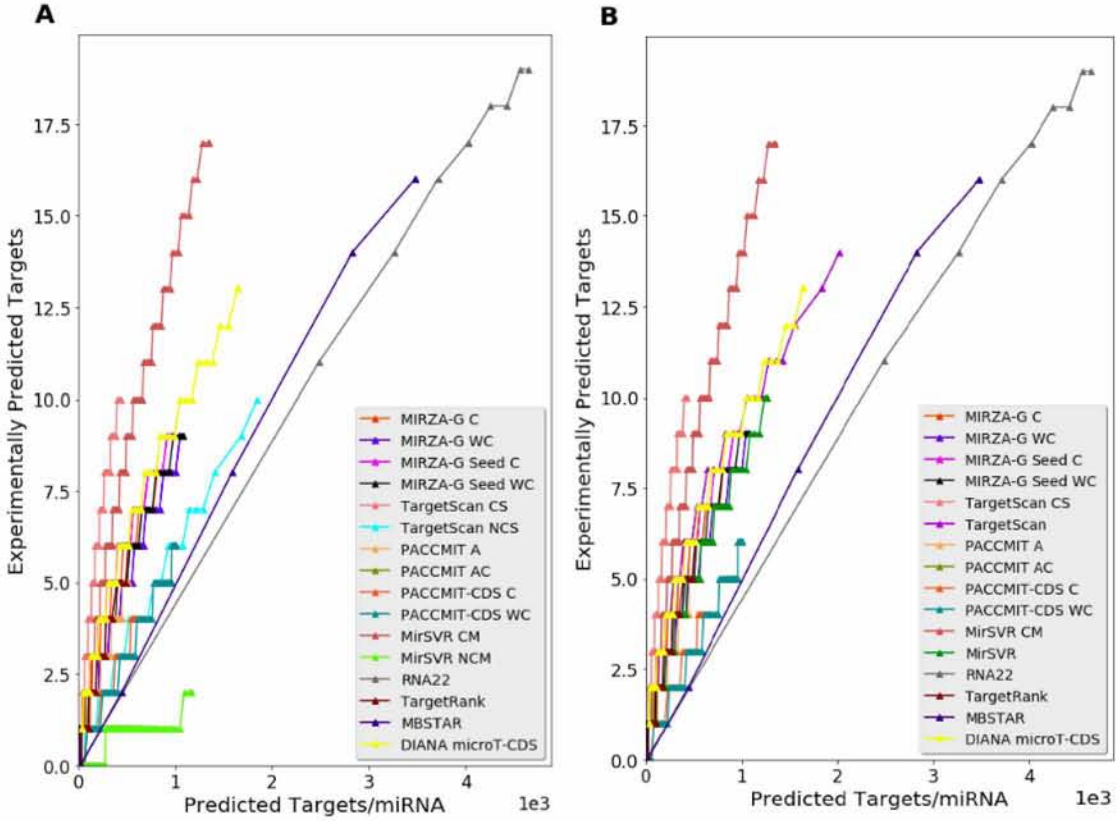


Figure 50 shows the Predicted Targets per miRNA and the corresponding number of correctly predicted targets for different scores. Predicted Targets per miRNA constitute the average value of Total predictions of all miRNAs in each program, having been grouped by score and they indicate the sensitivity of the examined models. Correctly predicted targets are calculated as the average value of experimentally verified targets of all miRNAs in each program, having been grouped by score. **Figures 46 and 50** share the same conclusions as far as the performance of target prediction algorithms is concerned.

Figure 50. Predicted Targets/miRNA vs Experimentally Predicted Targets. The number of correctly predicted targets is shown by different scores for increasing numbers of predicted targets per miRNA, when considering a step of 0.01. All predictions were filtered on the true positive set of miRNAs and the common set of genes among algorithms and Test Dataset 1. Between miRNA-gene interactions with the same score, those with the maximum scoring scheme are selected (Test Case III). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.



In **Figure 51**, despite the selection of the aggregated score of the corresponding miRNA-gene interactions, all algorithms, except TargetScan NCS, present the same performance as prior to the application of this filter.

Figure 52 demonstrates the Predicted Targets per miRNA and the corresponding number of correctly predicted targets for different scores after thresholds for RNA22 and MBSTAR have been employed. In **Figure 52** applies the same graphical analysis as in **Figure 50**, with the sole difference that the predictions of RNA22 with P-value ≥ 0.05 have been cut off. Conclusions are compatible with those in **Figure 46**.

Figure 51. Predicted Targets/miRNA vs Experimentally Predicted Targets. The number of correctly predicted targets is shown by different scores for increasing numbers of predicted targets per miRNA, when considering a step of 0.01. All predictions were filtered on the true positive set of miRNAs and the common set of genes among algorithms and Test Dataset 1. Between miRNA-gene interactions with the same score, the aggregation of their scoring scheme is selected (Test Case III). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.

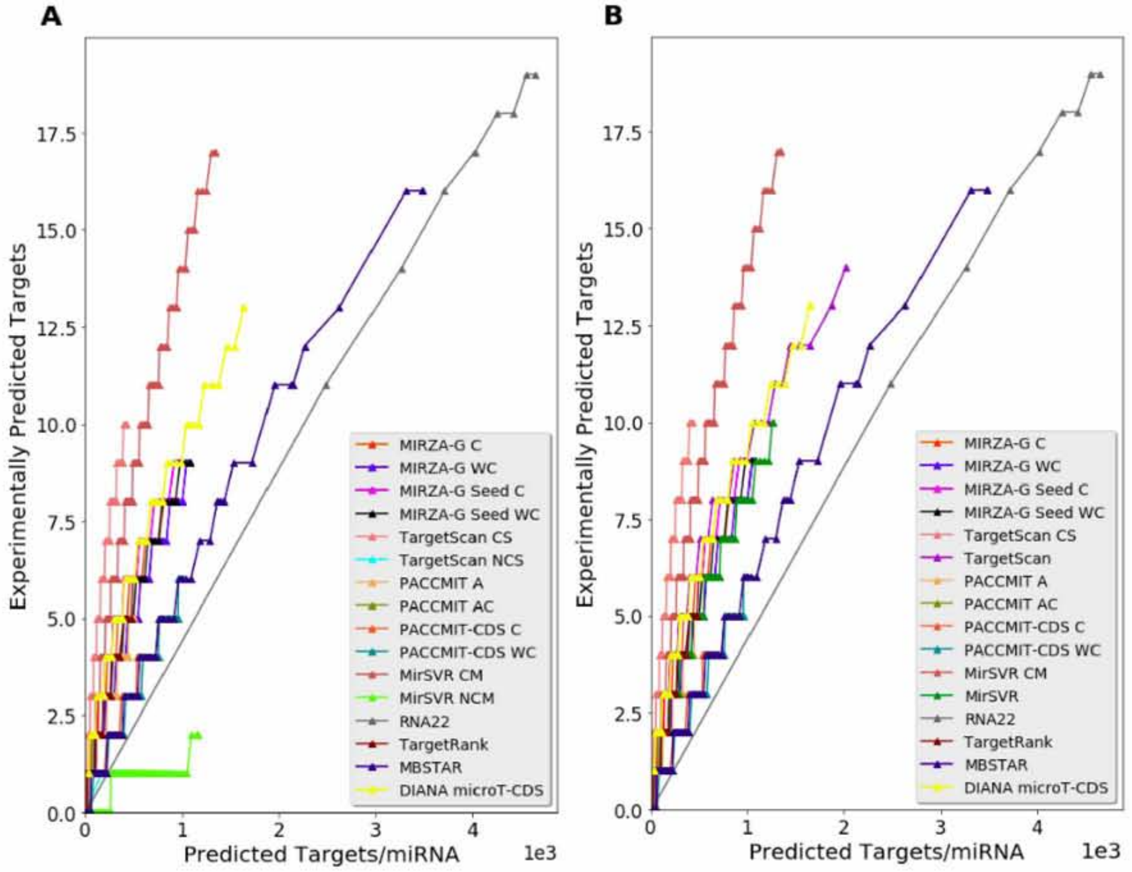
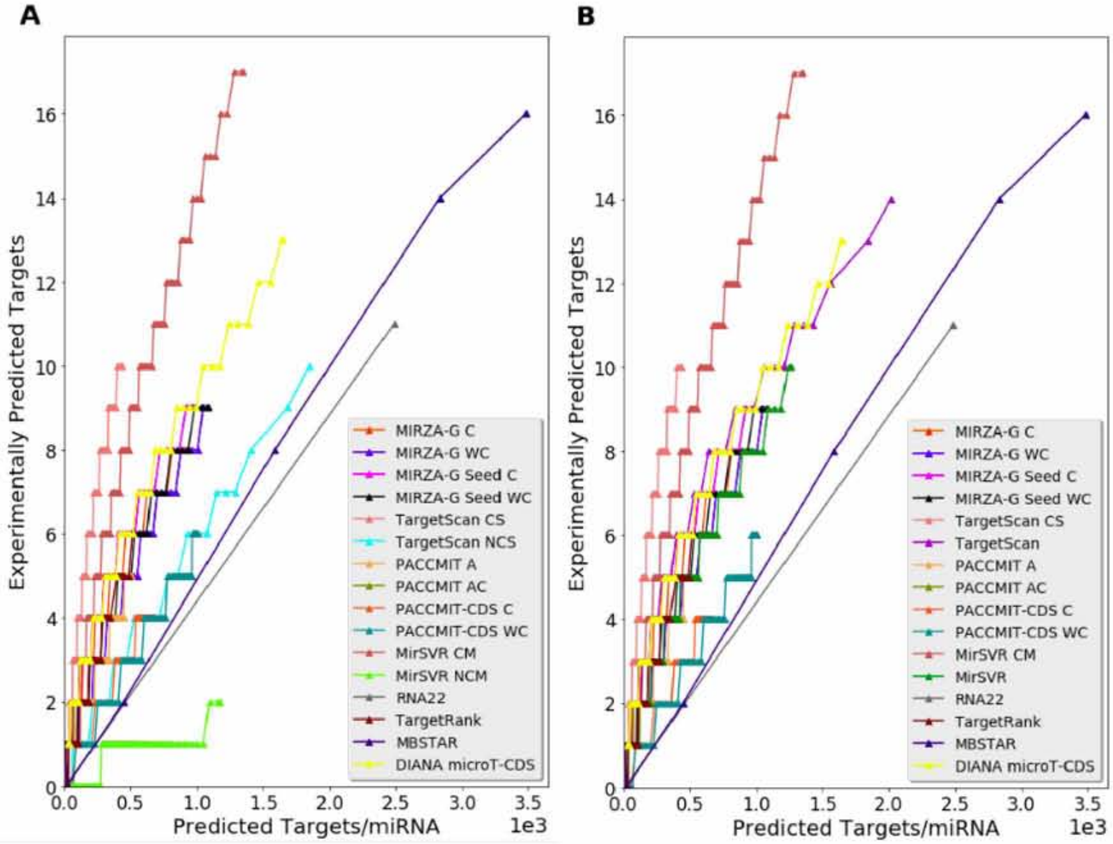
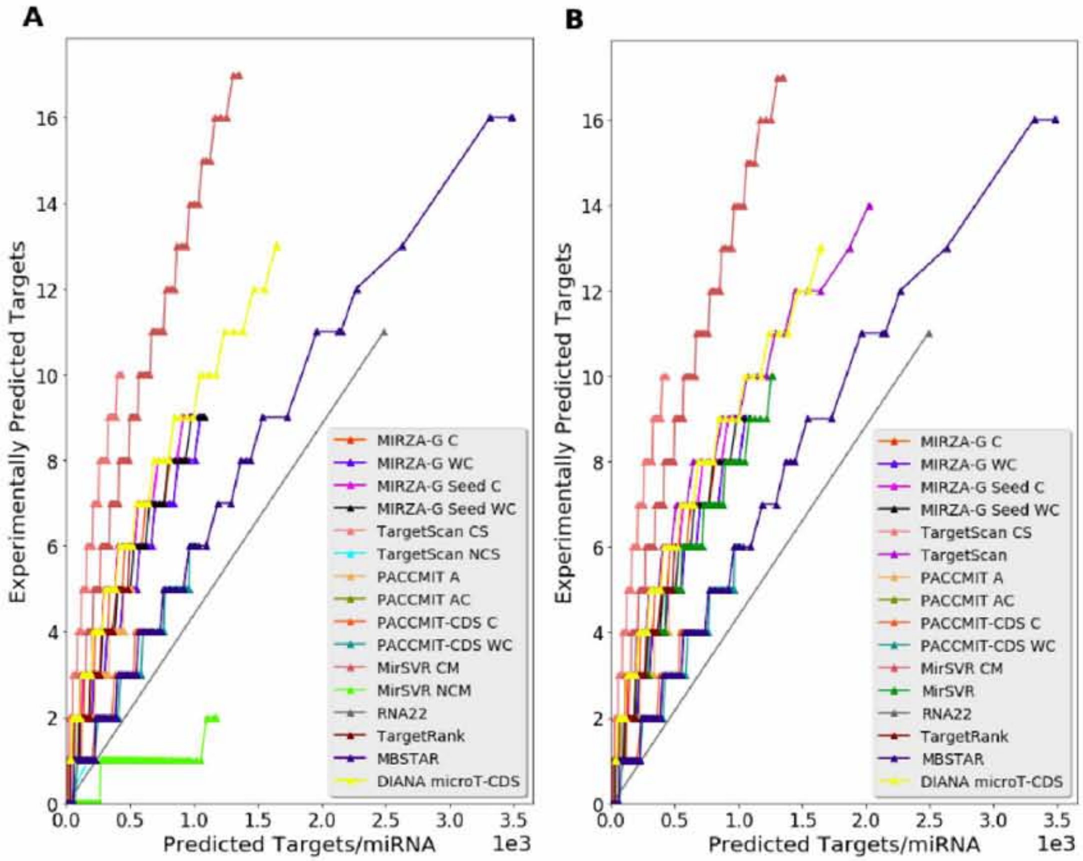


Figure 52. Predicted Targets/miRNA vs Experimentally Predicted Targets by setting threshold on RNA22 and MBSTAR. The number of correctly predicted targets is shown by different scores for increasing numbers of predicted targets per miRNA, when considering a step of 0.01. All predictions were filtered on the true positive set of miRNAs and the common set of genes among algorithms and Test Dataset 1. Between miRNA-gene interactions with the same score, those with the maximum scoring scheme are selected (Test Case III). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.



In **Figure 53**, the selection of the aggregated score of the corresponding miRNA-gene interactions does not alter the performance of the algorithms, with the exception of TargetScan NCS.

Figure 53. Predicted Targets/miRNA vs Experimentally Predicted Targets by setting threshold on RNA22 and MBSTAR. The number of correctly predicted targets is shown by different scores for increasing numbers of predicted targets per miRNA, when considering a step of 0.01. All predictions were filtered on the true positive set of miRNAs and the common set of genes among algorithms and Test Dataset 1. Between miRNA-gene interactions with the same score, the aggregation of their scoring scheme is selected (Test Case III). **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.



In **Figure 54**, it is observed that all algorithms share 91,636 common Total predictions. DIANA microT-CDS contains individually 244,607 total predictions while all algorithms except TargetScan CS share 157,191 predictions. In addition, MIRZA-G C, MIRZA-G Seed C and DIANA microT-CDS share 87,581 total predictions.

In **Figure 55**, it is shown that DIANA microT-CDS, TargetScan and mirSVR share 452,805 common Total predictions. In particular, DIANA microT-CDS and TargetScan contain 138,577 shared initial predictions, while TargetScan has individually 138,202 Total predictions.

Figure 54. Venn Diagram of the total predictions between MirSVR Conserved miRNAs, TargetScan Conserved Sites, MIRZA-G Mirza with conservation, DIANA microT-CDS and MIRZA-G Seed with conservation. These algorithms present the best performance according to previous comparisons (Test Case III).

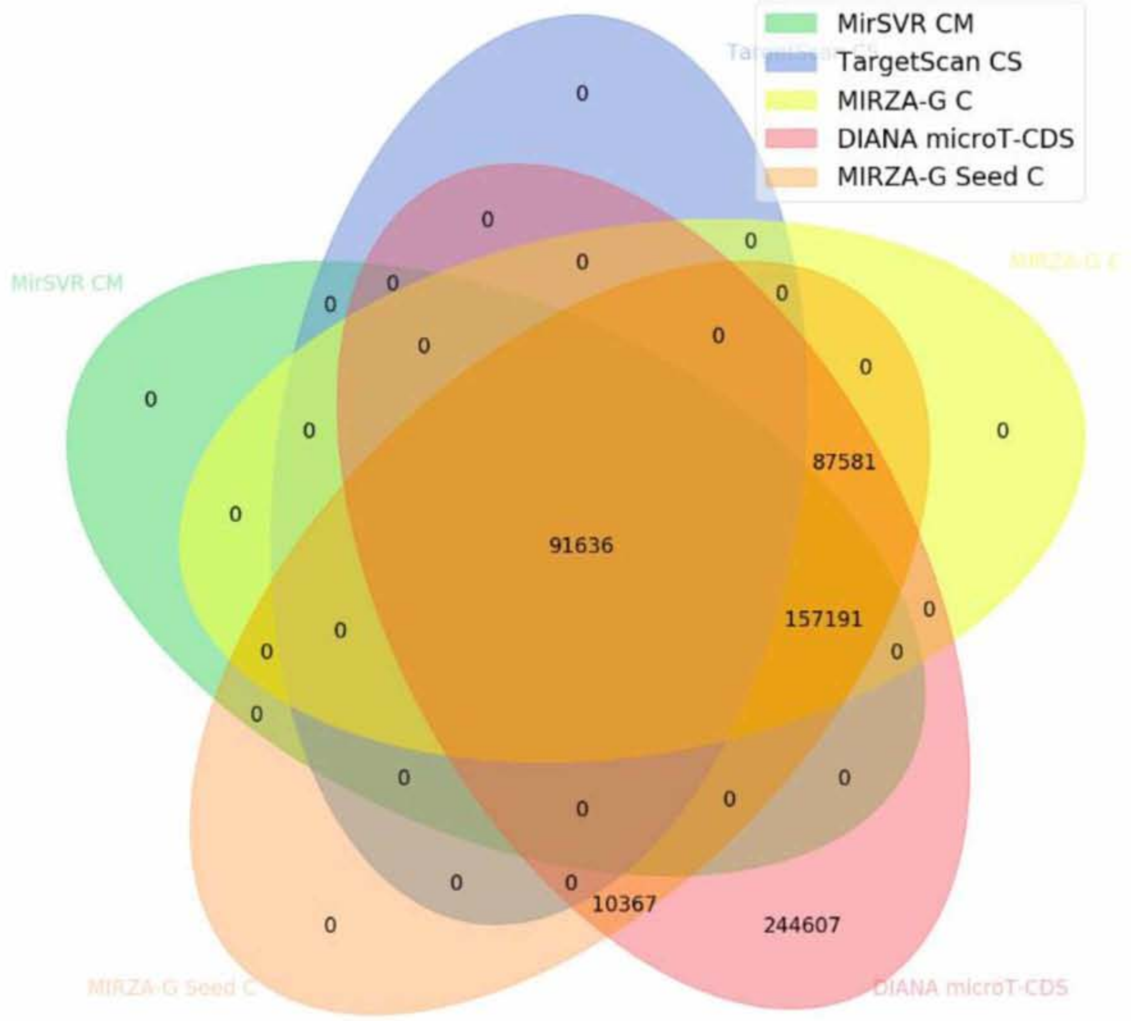
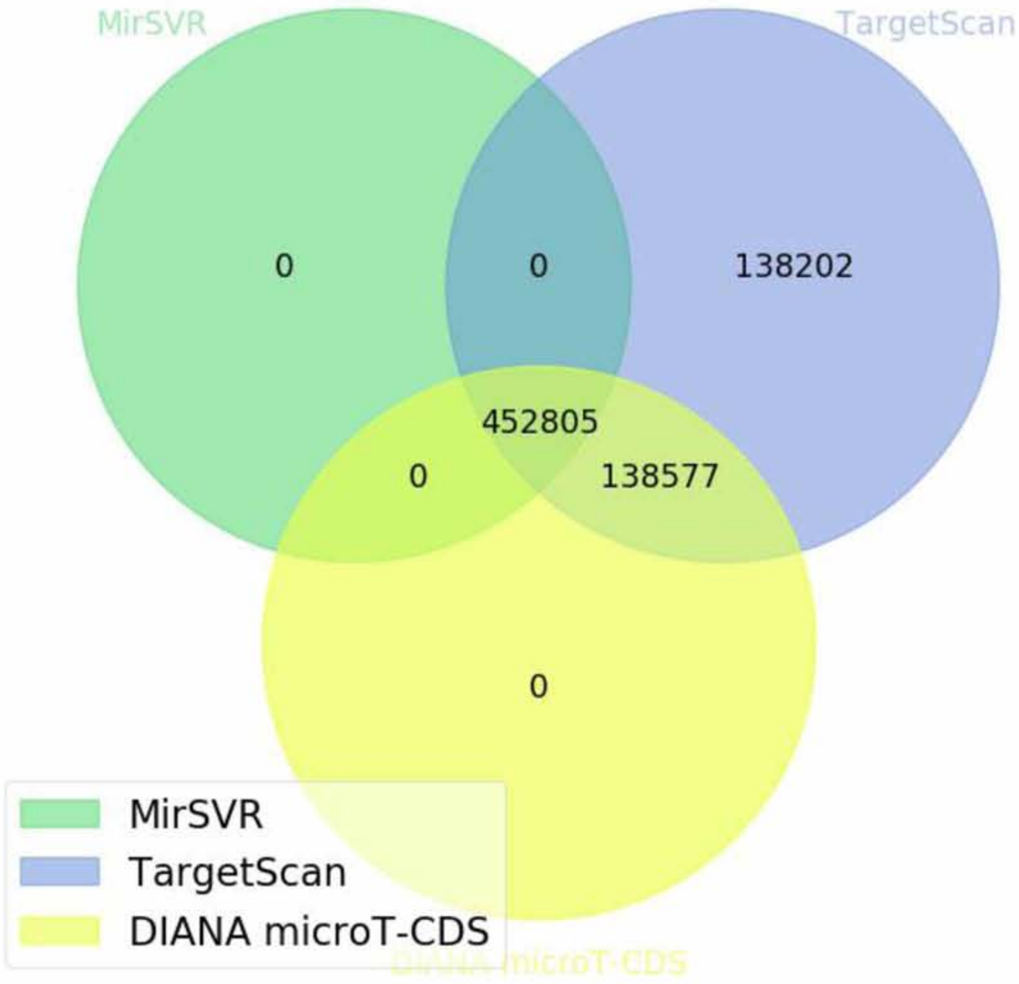


Figure 55. Venn Diagram of the total predictions between MirSVR, TargetScan and DIANA microT-CDS. These algorithms do not differentiate conserved and non conserved predictions (Test Case III).



In **Figure 56**, all algorithms share 2,064 correctly verified miRNA-gene interactions, which is a subset of Test Dataset 1. DIANA microT-CDS predicts individually 1,352 miRNA-gene interactions while all algorithms except TargetScan CS predict 886 common interactions.

In **Figure 57,** it is shown that DIANA microT-CDS, TargetScan and mirSVR share 3,453 common experimentally validated miRNA-gene interactions. Indeed, DIANA microT-CDS and TargetScan intersect in 1,071 interactions, while TargetScan predicts individually 372 miRNA-gene targets. Consequently, it is obvious that TargetScan can forecast correctly both its and other algorithms's interactions, concluding that the combination of itself with other miRNA target prediction programs, would not enhance its performance.

Figure 56. Venn Diagram of the experimentally verified predictions between MirSVR Conserved miRNAs, TargetScan Conserved Sites, MIRZA-G Mirza with conservation, DIANA-microT-CDS and MIRZA-G Seed with conservation. These algorithms present the best performance according to previous comparisons (Test Case III).

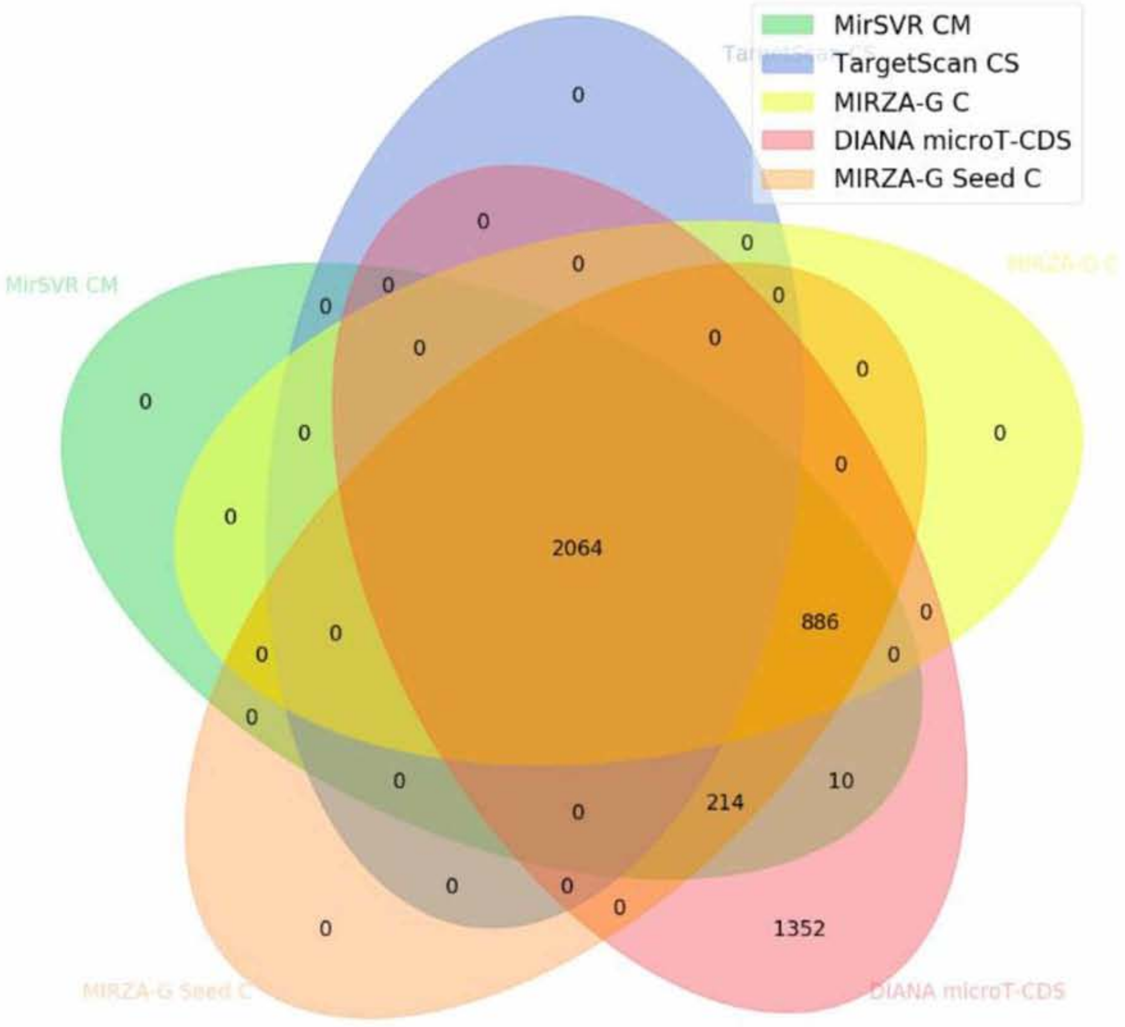
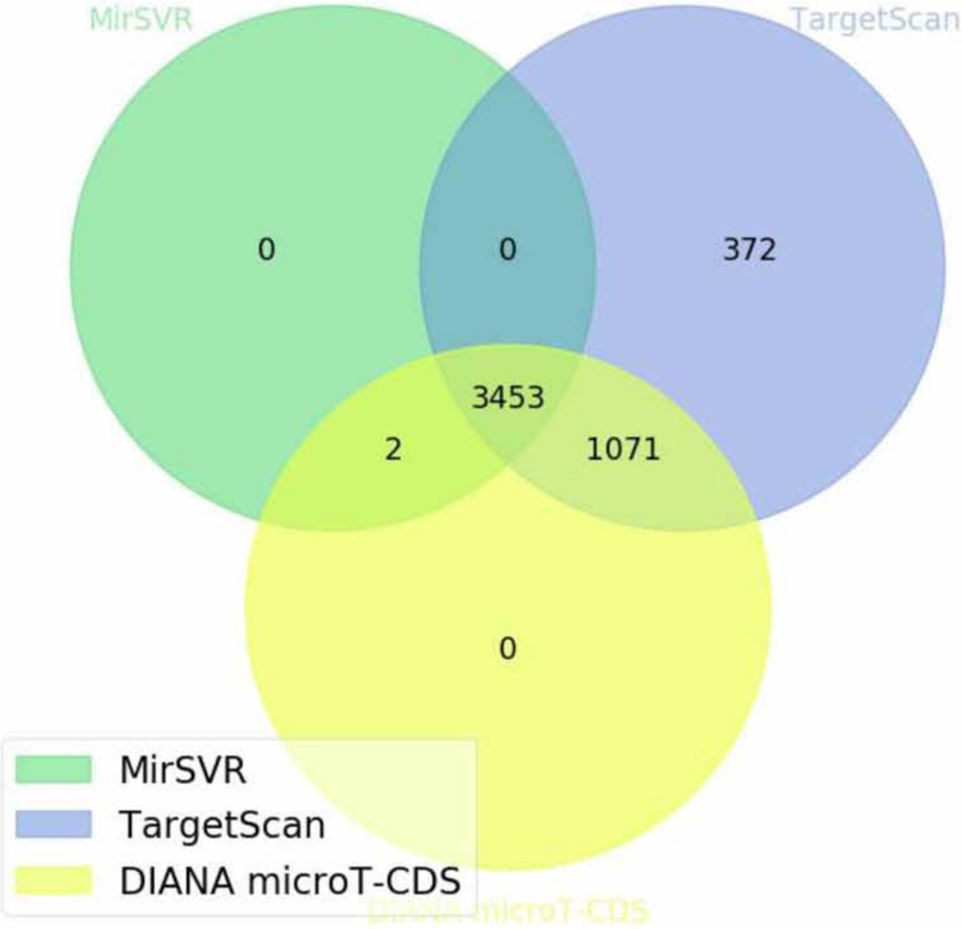


Figure 57. Venn Diagram of the experimentally verified predictions between MirSVR, TargetScan and DIANA microT-CDS. These algorithms do not differentiate conserved and non conserved predictions (Test Case III).



Figures 55, **57** are implemented with the purpose of examining the performance of the most optimal target prediction tools.

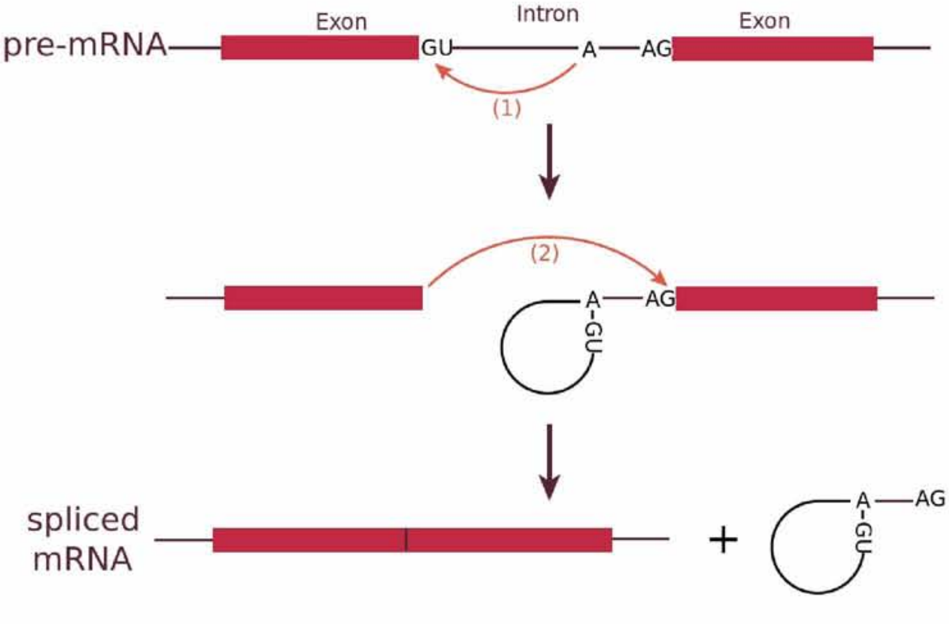
4.4 miRNA-Site Interactions in Test Case II

The goal is to find the common shared miRNA - site interactions between each program and Test Dataset 2. When searching for the shared miRNA - site interactions, at least one nucleotide (binding site) of each program should exist, overlapping with the experimentally verified binding region, for instance of a luciferase or chimeric site. From the set of programs studied, only DIANA microT-CDS, TargetScan, mirSVR, RNA22 and MBSTAR contain supplementary information regarding the coordinates of the binding sites on human genome.

What is more, DIANA microT-CDS contains interactions, which fall into splice junctions. As a result, two exons, with a continued MRE, can be contained in a record. For the analysis, each exon has been split into a single row. Due to the fact that the region of an MRE, shared in two exons, can be extensive enough, if a site in Test Dataset 2 overlaps with both regions of MRE in two exons, then the shared region is taken into account once.

In the splicing of RNA, the site of a former intron in a mature mRNA embraces a splice junction. In molecular biology, splicing is the editing of the nascent precursor messenger RNA (pre-mRNA) transcript into a mature messenger RNA (mRNA). After splicing, introns are removed and exons are joined together. **Figure 58** illustrates the two steps of canonical RNA processing, from pre-mRNA to spliced RNA and the intron lariat.

Figure 58. Diagram illustrating the two-step biochemistry of splicing [149].



Most programs, as presented in **Table 19**, contain relative coordinates to the start or end of a gene, which corresponds to a specific mRNA. For more accurate comparisons, these coordinates are converted to absolute ones, indicating the precise location of the miRNA on the human genome. In addition, the consequent genome coordinates are converted to the appropriate assembly. In particular, all programs hold coordinates in hg19 assembly. Thus, this assembly is transformed to hg38 via LiftOver tool, which satisfies the aforementioned purpose.

The predictions of each program were filtered at the common set of miRNAs among all programs that is subset of the positive set and the common set of Genes among all programs. A threshold of P-value < 0.05 for RNA22, P-value ≥ 0.5 for MBSTAR and P-value ≥ 0.5 for DIANA microT-CDS is considered. **Table 26** demonstrates the sizes of the common sets of miRNAs and Test Dataset 2 and genes among all programs respectively.

Table 26. Common set of miRNAs and Genes across all programs in case of miRNA-Site interactions.

	Total
miRNAs Subset	248
Genes Subset	12,059

In **Figure 59**, Total predictions and True positive experimentally validated predictions (shared miRNA-site interactions with Test Dataset 2) of all programs are shown in blue and yellow respectively, after having applied threshold for DIANA microT-CDS, RNA22 and MBSTAR algorithms. The results are calculated without taking into account the score of each miRNA-Site interaction. It is evident that TargetScan outperforms the other target prediction programs, while DIANA microT-CDS has a vast number of predictions and therefore is highly sensitive.

Figure 60 demonstrates the shared miRNA-Site interactions with Test Dataset 2, as **Figure 59**, with the exception that the number of initial total predictions has been decreased due to the additional filtering of algorithms at the overlapping regions of the positive set. Subsequently, it is observed that TargetScan maintains optimal performance compared to other programs.

Figure 59. Total and True positive set of overlapping sites for Target Prediction Algorithms without considering the score and setting threshold on DIANA microT-CDS according to Test Dataset 2. Between miRNA-site interactions with the same score, those with the maximum scoring scheme are selected.

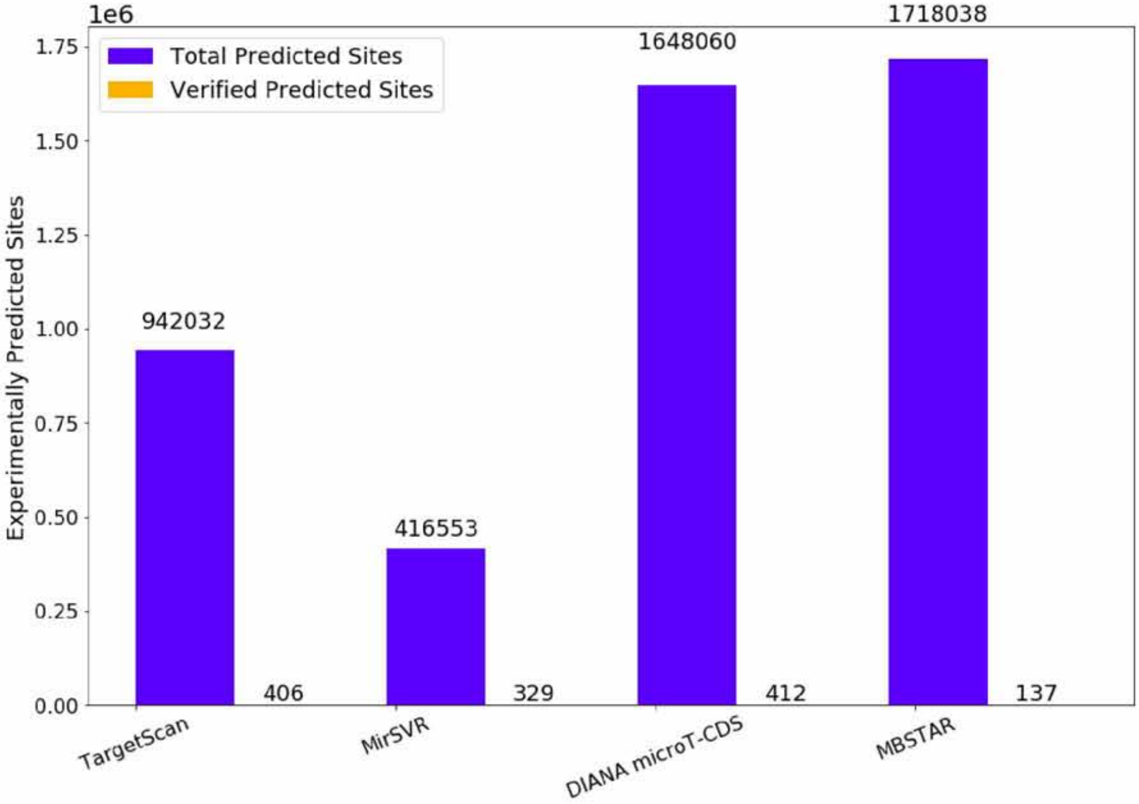


Figure 60. Total and True positive set of overlapping sites of Target Prediction Algorithms with the positive set without considering the score and setting threshold on DIANA-microT-CDS according to Test Dataset 2. Between miRNA-site interactions with the same score, those with the maximum scoring scheme are selected. All algorithms have been filtered at the overlapping regions of the positive set.

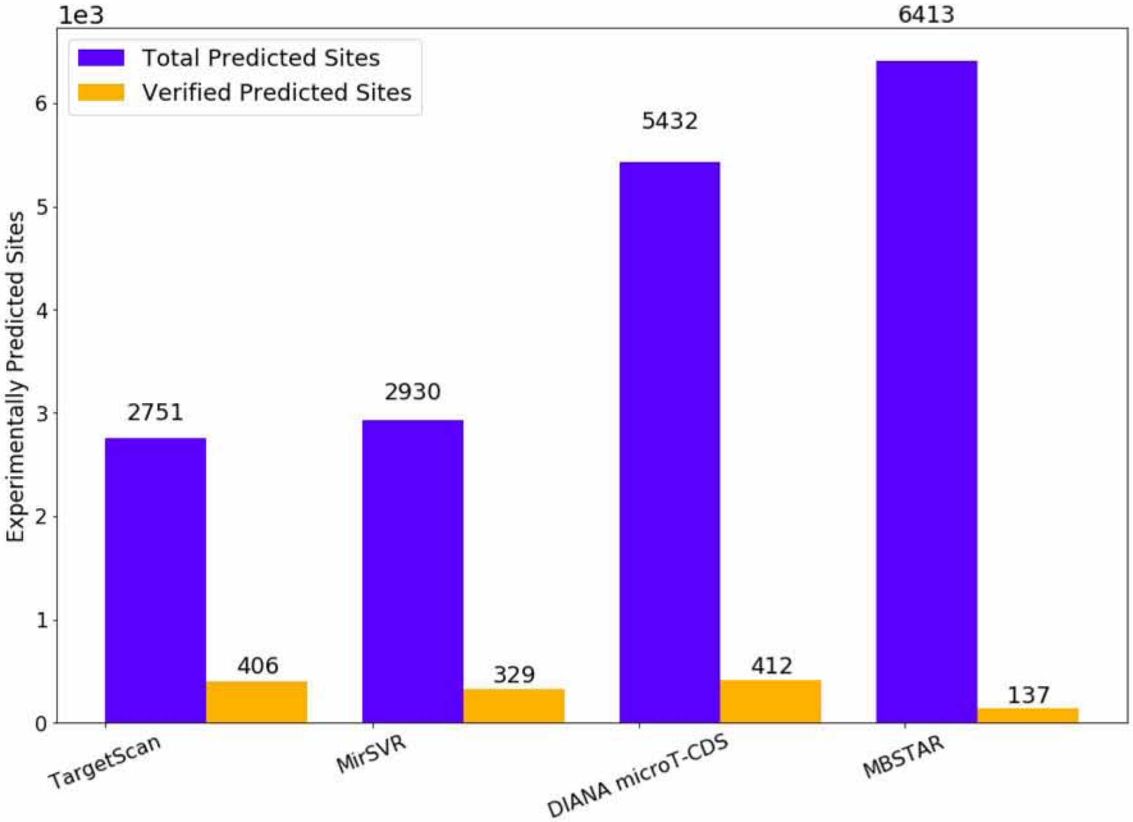
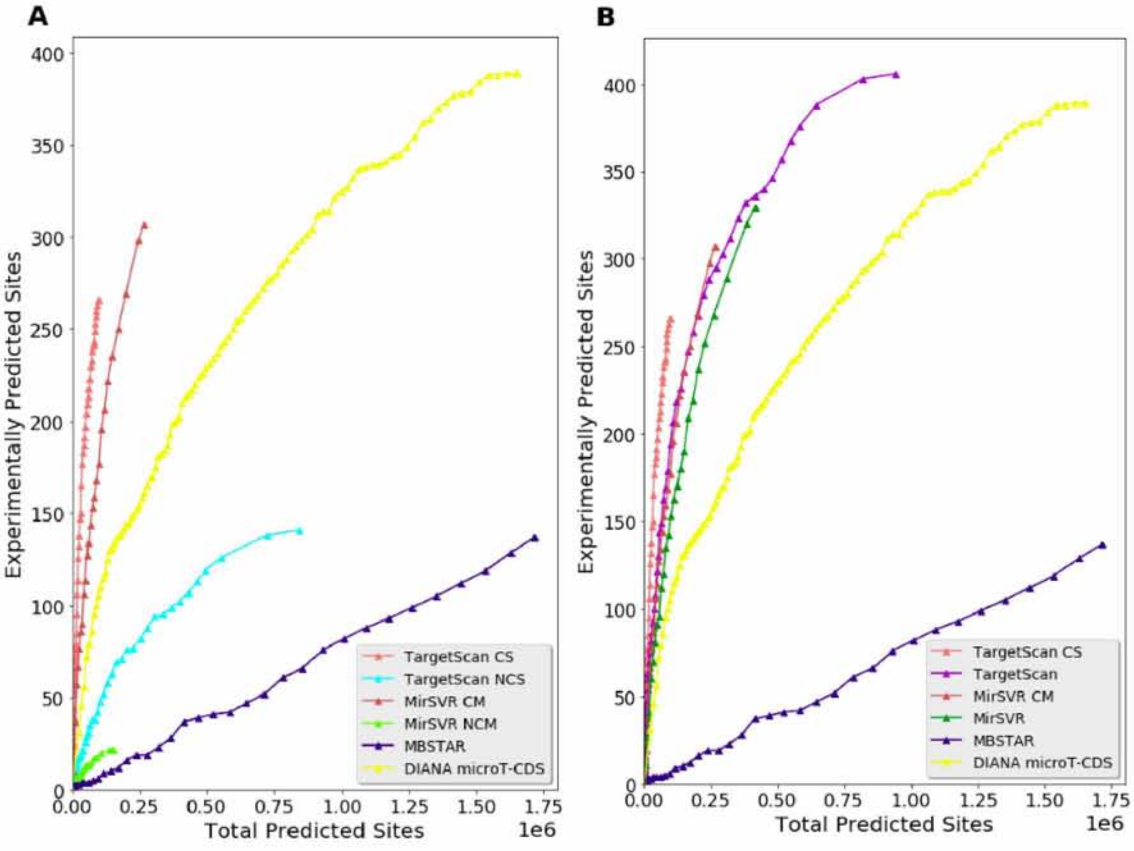


Figure 61 illustrates Total Predicted Sites and Experimentally Predicted Sites of all programs for corresponding score values. The number of correctly predicted sites is shown by different scores for increasing numbers of total predicted sites. Between miRNA-site interactions with the same score, the maximum of their scoring scheme is selected. As a result, for a relatively large number of Total predicted sites, DIANA microT-CDS achieves a larger number of experimentally verified predicted sites (high sensitivity).

In **Figure 61A**, it is evident that conserved predicted sites present an optimized performance compared to non-conserved predicted sites. In addition, it is observed that TargetScan CS has the optimal performance when compared to the other target prediction algorithms due to the fact it achieves high accuracy and its sensitivity is almost nonexistent. Indeed, the curve that corresponds to Targetscan CS initially starts with a relatively small number of Total predicted sites and map them to a large number of experimentally verified predicted sites. Hierarchically, following TargetScan CS, MirSVR CM and DIANA microT-CDS present the greatest performance among target prediction tools. On the other hand, MBSTAR's performance appears to fall drastically compared to miRNA-genes and miRNA-genes/miRNA interactions as far as specificity is concerned.

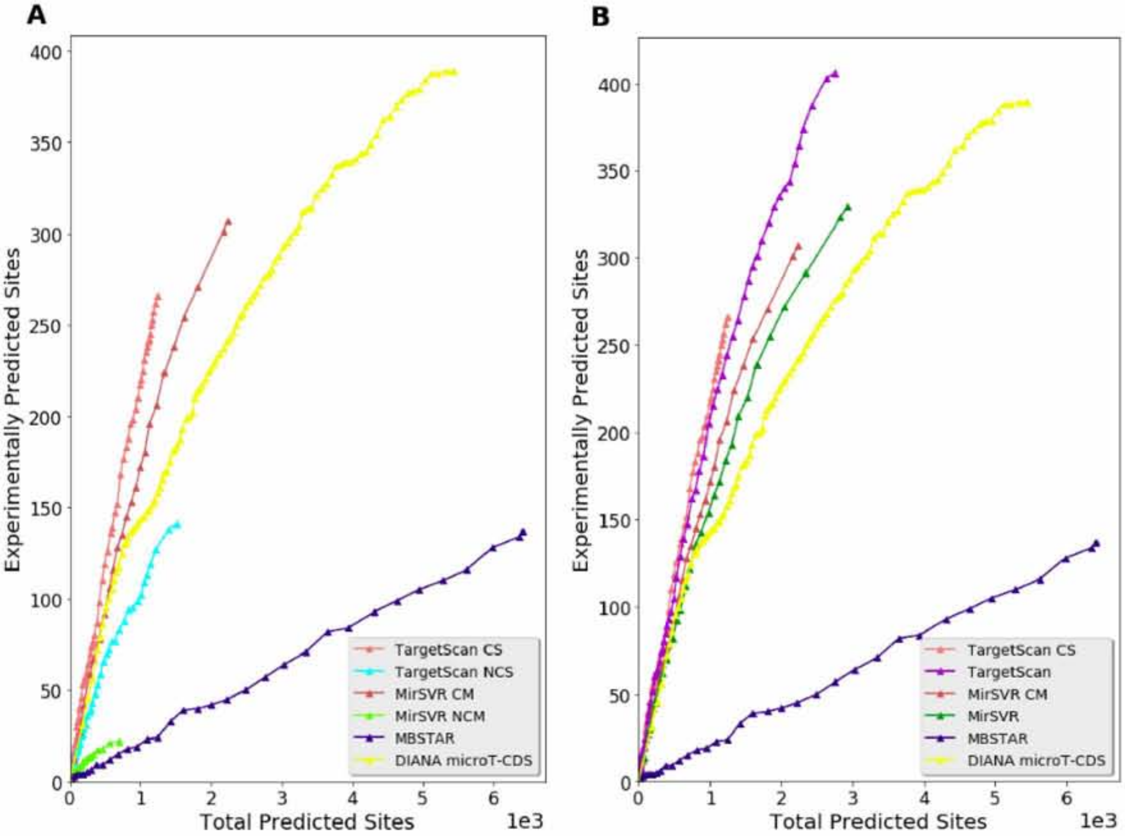
From **Figure 61B**, conserved miRNA:site human interactions outperform the Total interactions due to the fact that Total predictions constitute the aggregation of conserved and non-conserved interactions. For instance, TargetScan CS presents better performance compared to TargetScan. Moreover, conserved miRNA - site interactions of MirSVR (MirSVR CM) are less sensitive than the one of MirSVR.

Figure 61. Total Predicted Sites vs Experimentally Predicted Sites of all programs when considering step = 0.01 for Test Dataset 2. Threshold on DIANA microT-CDS is set. Between miRNA-site interactions with the same score, those with the maximum scoring scheme are selected. **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions.



In **Figure 62** applies the same graphical analysis as in **Figure 61**, with the sole difference that the total predicted sites have been declined due to the restrain of the total predicted sites of all algorithms at the overlapping regions of the positive set. Without discriminating the conserved from non conserved predicted sites, TargetScan contains the optimum performance, following by mirSVR and DIANA microT-CDS algorithms. In addition, DIANA microT-CDS outperforms mirSVR in miRNA-gene interactions' level, while mirSVR presents a slight enhancement in performance in miRNA-site interactions, when compared with DIANA microT-CDS.

Figure 62. Total Predicted Sites vs Experimentally Predicted Sites of all programs when considering step = 0.01for Test Dataset 2. Threshold on DIANA microT-CDS is set. Between miRNA-site interactions with the same score, those with the maximum scoring scheme are selected. **A)** Conserved and Non-conserved predictions. **B)** Total and Conserved predictions. All algorithms have been filtered at the overlapping regions of the positive set.



Chapter V

5. Code Architecture

Due to the complexity and the number of scripts devised in order to study and manipulate the precomputed predictions of all target prediction algorithms, a functional environment has been developed, depicting the entire architecture of the code. All the scripts of the code have been written in python programming language. Below, follows a detailed description of the functionality of all scripts and graphical representations of them for more accurate and circumstantial study.

Test Case I

For PACCMIT Accessibility and PACCMIT Accessibility & Conservation in script: ***Map_miRNA_to_MIMAT.py***

1. Read the data into a dataframe *(data_PACCIMIT*)
2. Clear data_PACCIMIT dataframe from entries containing CDR in Gene ID column and anything else apart from hsa in miRNA column
3. Reformat values for Gene ID column by splitting in the '.' character and holding only the left part. This procedure is performed using an unnamed lambda function.
4. Read the mirBase 18 conversion file into a dataframe (*MirBase_version18*)
5. Merge the *data_PACCIMIT* and *MirBase_version18* dataframes in order each miRNA to be mapped to a MIMAT in the appropriate MirBase version. In case of records that cannot be mapped to a MIMAT, due to the fact that their miRNA does not exist in the newer version of MirBase or is updated in another name, they are removed. The aforementioned are performed by *merge()* and *pandas.notnull()* functions respectively.

Ouput files (Accessibility):
data_PACCIMIT_access_miRNA_18_all_non_cons.csv
Ouput files (Accessibility &Conservation:
data_PACCIMIT_access_cons_miRNA_18_all.csv

For PACCMIT-CDS with and without conservation in scripts:

Map_miRNA_to_MIMAT.py

1. Read the data into a dataframe *(data_PACCIMIT_CDS*)
2. In column "miRNAs_with_the_same_seed_sequence" some records contain more than one values split by the "," character. These records are isolated in a dataframe (*data_PACCIMIT_CDS_mult*) and they are split to separate rows by holding the other columns intact.
3. *data_PACCIMIT_CDS_mult dataframe* is concatenated with the

data_PACCIMIT_CDS dataframe, without the "special" records, into data_PACCIMIT_CDS_new dataframe.

4. Read the mirBase 18 conversion file into a dataframe (*MirBase_version18*)
5. Merge the *data_PACCIMIT_CDS_cons_new* and *MirBase_version18* dataframes in order each miRNA to be mapped to a MIMAT in the appropriate MirBase version. In case of records that cannot be mapped to a MIMAT, due to the fact that their miRNA does not exist in the newer version of MirBase or is updated in another name, they are removed. The aforementioned are performed by *merge()* and *pandas.notnull()* functions respectively.

Ouput files (conservation): **data_PACCIMIT_CDS_cons_miRNA_18_all.csv**
Ouput files (without conservation):
data_PACCIMIT_CDS_non_cons_miRNA_18_all.csv

Map_ENST_to_ENSG.py

1. Read the data (data_PACCIMIT_CDS_miRNA_18_all.csv) into a dataframe *(data_PACCIMIT_CDS_new*).
2. Read the file of mapping (mart_export.txt) between Transcript Stabe ID and Gene Stable ID into a dataframe (enst_to_ensg).
3. Merge the *data_PACCIMIT_CDS_new* and *enst_to_ensg* dataframes in order each transcript to be mapped to an ensembl gene ID. In case of records that cannot be mapped to an ensembl gene ID, they are removed. The aforementioned are performed by *merge()* and *pandas.notnull()* functions respectively.

Ouput files (conservation): **PACCMIT_CDS_Gene_stable_ID_ready**
Ouput files (without conservation): **PACCMIT_CDS_non_Gene_stable_ID_ready**

For TargetScan Conserved and Non Conserved predictions, in script:
Map_miRNA_to_MIMAT.py

1. Read the data into a dataframe *(data_TargetScan*)
2. Clear data_TargetScan dataframe from entries containing CDR in Gene ID column and anything else apart from hsa in miRNA column
3. Reformat values for Gene ID column by splitting in the '.' character and holding only the left part. This procedure is performed using an unnamed lambda function.
4. Read the mirBase 21 conversion file into a dataframe (*MirBase_version21*)
5. Merge the *data_TargetScan dataframe* and *MirBase_version21* dataframes in order each miRNA to be mapped to a MIMAT in the appropriate MirBase version. In case of records that cannot be mapped to a MIMAT, due to the fact that their miRNA does not exist in the newer version of MirBase or is updated in another name, they are removed. The aforementioned are performed by *merge()* and *pandas.notnull()* functions respectively.

Ouput files: **data_cons_TargetScan_MIMAT.csv**, **data_non_cons_TargetScan_MIMAT.csv**

For TargetScan Total predictions, in script: ***TargetScan_Total_Prep_for_common_comp.py***

1. Read data_cons_TargetScan_MIMAT.csv
2. Read data_non_cons_TargetScan_MIMAT.csv
3. Concatenate data_cons_TargetScan_MIMAT.csv and data_non_cons_TargetScan_MIMAT.csv into a dataframe

Ouput files: **data_Total_TargetScan_MIMAT.csv**

For MIRSVR Conserved and Non Conserved predictions, in script: ***Conserved_Map_NM_to_ENSG.py***

1. Read the data into a dataframe *(MirSVR*).
2. Read the file of mapping **(NM_TO_ENG.txt**) between RefSeq mRNA ID and Gene Stable ID into a dataframe (NM_to_ENG).
3. Merge the *MirSVR* and *NM_to_ENG* dataframes in order each RefSeq to be mapped to an ensembl gene ID. In case of records that cannot be mapped to an ensembl gene ID, they are removed. The aforementioned are performed by *merge()* and *pandas.notnull()* functions respectively.

Ouput files**: MIRSVR_MIMAT_Gene_stable_ID_**, **MIRSVR_MIMAT_Gene_stable_ID_non**

For MIRSVR Total predictions, in script: ***MirSVR_TOTAL_Prep_for_common_comp.py***

1. Read MIRSVR_MIMAT_Gene_stable_ID_
2. Read MIRSVR_MIMAT_Gene_stable_ID_non
3. Concatenate MIRSVR_MIMAT_Gene_stable_ID_ and MIRSVR_MIMAT_Gene_stable_ID_non into a dataframe

Ouput files: **MIRSVR_MIMAT_Gene_stable_ID_TOTAL**

For MIRZA-G Mirza/Seed Analysis with and without conservation, in scripts: ***Mirza_Map_miRNA_to_MIMAT.py/Seed_Map_miRNA_to_MIMAT.py***

1. Read the data into a dataframe *(MIRZAG_data*)
2. Read the mirBase 20 conversion file into a dataframe (*MirBase_version20*)
3. Merge the *MIRZAG_data* and *MirBase_version20* dataframes in order each miRNA to be mapped to a MIMAT in the appropriate MirBase version. In case of records that cannot be mapped to a MIMAT, due to the fact that their miRNA does not exist in the newer version of MirBase or is updated in another name,

they are removed. The aforementioned are performed by *merge()* and *pandas.notnull()* functions respectively.

Ouput files: **MIRZAG_data_MIMAT.csv (4)** [1]

Mirza_Map_gene_id_to_ENSG.py / Seed_Map_gene_id_to_ENSG.py

1. Read the data into a dataframe *(MIRZAG_data*).
2. Read the file of mapping (gene_to_ensembl.txt) between NCBI Gene ID and Gene Stable ID into a dataframe (gene_id_to_ensg).
3. Merge the *MIRZAG_data* and *gene_id_to_ensg* dataframes in order each NCBI Gene ID to be mapped to an ensembl gene ID. In case of records that cannot be mapped to an ensembl gene ID, they are removed. The aforementioned are performed by *merge()* and *pandas.notnull()* functions respectively.

Ouput files: **MIRZA_G_MIMAT_Gene_stable_ID_(2)**,
MIRZA_G_Seed_Gene_stable_ID_ (2) [1]

For programs RNA22 and TargetRank, in scripts
Filtering_miRNAs_on_experiments_data.py

1. Read the mirBase conversion file into a dataframe (*MirBase_version*)
2. Read the positive set file (Test Dataset 1) into a dataframe (experiment_data).
3. Merge *MirBase_version and experiment_data dataframes* in order each MIMAT of Test Dataset 1 to be mapped to a miRNA in the appropriate MirBase version. In case of records that cannot be mapped to a MIMAT they are removed. The aforementioned are performed by *merge()* and *pandas.notnull()* functions respectively.
4. Iterate all files in data_dir[2] directory and remove the one whose miRNA does not exist in Test Dataset 1. The successful remaining files are hold in list (data_list).
5. data_list list is converted to dataframe.
6. Merge *MirBase_version and* data_list dataframes in order each miRNA to be mapped to a MIMAT in the appropriate MirBase version. In case of records that cannot be mapped to a MIMAT, due to the fact that their miRNA does not exist in the newer version of MirBase or is updated in another name, they are removed. The aforementioned are performed by *merge()* and *pandas.notnull()* functions respectively.

Ouput files: **RNA22_df_MIMAT, HomoSapiens_mRNA_ENSEMBL78_dir_df** (data_list), **TargetRank_df_MIMAT**

For MBSTAR, in script:
MBSTAR_Prep_for_common_comp.py

1. Read the mirBase 20 conversion file into a dataframe (*MirBase_version20*)

2. Read the positive set file (Test Dataset 1) into a dataframe (experiment_data).
3. Merge *MirBase_version20 and experiment_data dataframes* in order each MIMAT of Test Dataset 1 to be mapped to a miRNA in the appropriate MirBase version(20). In case of records that cannot be mapped to a MIMAT they are removed. The aforementioned are performed by *merge()* and *pandas.notnull()* functions respectively.
4. Iterate all files in all folders[3] of MBSTAR_Genome_Wide_pred_res directory and create a folder (UPDATE_Output) with the files whose miRNA only exist in Test Dataset 1. Also, these files are included in a list (sum_mbstar_files).
5. sum_mbstar_filelist is converted to dataframe.
6. Merge *MirBase_version20 and* sum_mbstar_file dataframes in order each miRNA to be mapped to a MIMAT in the appropriate MirBase version. In case of records that cannot be mapped to a MIMAT, due to the fact that their miRNA does not exist in the newer version of MirBase or is updated in another name, they are removed. The aforementioned are performed by *merge()* and *pandas.notnull()* functions respectively.

Ouput files: **MBSTAR_df_MIMAT**

For DIANA-microT-CDS, in script:
Map_miRNA_to_MIMAT.py

1. Read the mirBase 18 conversion file into a dataframe (*MirBase_version18*)
2. Read the data into a dataframe *(Micro_T_CDS*) in chunks, with chunk size = 10^6.
3. Clear *Micro_T_CDS* dataframe from entries containing CDR in Gene ID column and anything else apart from hsa in miRNA column
4. Reformat values for Gene ID column by splitting in the '.' character and holding only the left part. This procedure is performed using an unnamed lambda function.
5. Merge *MirBase_version18 andMicro_T_CDS* dataframes in order each miRNA to be mapped to a MIMAT in the appropriate MirBase version. In case of records that cannot be mapped to a MIMAT, due to the fact that their miRNA does not exist in the newer version of MirBase or is updated in another name, they are removed. The aforementioned are performed by *merge()* and *pandas.notnull()* functions respectively.

Output files: **Micro_T_CDS_MIMAT.csv**

As far as the final running of all algorithms for the production of plots is concerned, all programs follow the subsequent generic steps.

In script **data_Total_Verified_Predictions.py:**

1. Read Common_df_alforithms_total file into a dataframe.
2. Read algorithm_input into data_TOTAL dataframe

3. Merge Common_df_alforithms_total and data_TOTAL dataframes in order to filter the records of each algorithm on the MIMATs of Test Dataset 1 and all algorithms. The result is stored in data_TOTAL dataframe.

4. Read the positive set file (Test Dataset 1) into a dataframe (experiment_data).

5. Create 2 flows of code.

5.1. Flow of maximum scores.

5.1.1. Between records with the same MIMAT and Gene_ID,the ones with the maximum score[4] are chosen.

5.1.2. Export of data_TOTAL into data_TOTAL_ready_for_2_comp file.

5.1.2.1. Merge data_TOTAL and experiment_data on Ensembl_Gene_id- MIMAT interactions. The resultis saved in data_TOTAL_first_comp file.

5.1.2.2. Export data_TOTAL_first_comp file along with the number of total predictions and experimentally verified predictions into data_TOTAL_first_comp_ready file.

Steps 5.1.2.1, 5.1.2.2 concern the case of finding the common miRNA-gene interactions with Test Dataset 1 without taking into account the score.

5.1.3. For each score between the maximun value of data and 0 the following steps (5.1.3.1, 5.1.3.2) are repeated.

5.1.3.1. Merge data_TOTAL and experiment_data on Ensembl_Gene_id- MIMAT interactions. The resultsare saved in Verified_pred_per_loop_total file.

5.1.3.2. Total predictions and experimentally verified predictions are appended into(*Total_pred_verified*) dataframe.

5.1.4. *Total_pred_verified* dataframe is saved into data_Total_pred_verified_total file.

5.2. Flow of aggregated scores

5.2.1. Between records with the same MIMAT and Gene_ID, a sole record is chosen with score the aggregation of all intermediate scores.

5.2.2. Export of data_TOTAL into data_TOTAL_ready_for_2_comp_sum file.

5.2.2.1.1. Merge data_TOTAL and experiment_data on Ensembl_Gene_id- MIMAT interactions. Theresult is saved in data_TOTAL_first_comp file.

5.2.2.1.2. Export data_TOTAL_first_comp file along with the number of total predictions and experimentally verified predictions into data_TOTAL_first_comp_ready_sum file.

Steps 5.2.2.1.1, 5.2.2.1.2 concern the case of finding the common miRNA-gene interactions with Test Dataset 1 without taking into account the score.

5.2.3. For each score between the maximum value of data and 0 the following steps (5.2.3.1, 5.2.3.2) are repeated.

5.2.3.1. Merge data_TOTAL and experiment_data on Ensembl_Gene_id- MIMAT interactions. The resultsare saved in Verified_pred_per_loop_total_sum file .

5.2.3.2. Total predictions and experimentally verified predictions are appended into (*Total_pred_verified*) dataframe.

5.2.4. *Total_pred_verified* dataframe is saved into data_*Total_pred_verified_total_sum* file.

Output files:
Verified_pred_per_loop_total,data_Total_pred_verified_total,Verified_pred_per_loop_total_sum ,data_*Total_pred_verified_total_sum*

It is important to mention that thresholds are set to algorithms such as RNA22, MBSTAR and DIANA Micro-T-CDS. This step precedes the creation of two flows inthe procedure (step 5).

In script ***data_Total_Average_Verified_Predictions .py***:

1. Read the positive set file (Test Dataset 1) into a dataframe (experiment_data).
2. Read data_TOTAL_ready_for_2_comp file into data_TOTAL dataframe
3. Create a flow for maximum score: For each score between the maximum value of data and 0 the following steps (3.1, 3.2) are repeated.
   3.1. Merge data_TOTAL and experiment_data on Ensembl_Gene_id- MIMAT interactions.
   3.2. Total predictions and experimentally verified predictions are appended into (*Total_pred_verified*) dataframe.
4. *Total_pred_verified* dataframe is grouped by score and two new columns are created. "Average_Total_Predictions" column contains the average value of "Total_Predictions" column while "Average_Correctly_Verified_Predictions" column contains the average value of "Correctly_Verified_Predictions".
5. T*he new average_df* dataframe is saved into data_Total_pred_verified_total file.
6. .Read data_TOTAL_ready_for_2_comp_sum file into data_TOTAL dataframe
7. Create a flow for aggregated scores: For each score between the maximum value of PACCMIT and 0 the following steps (7.1, 7.2) are repeated.
   7.1. Merge data_TOTAL and experiment_data on Ensembl_Gene_id- MIMATinteractions.
   7.2. Total predictions and experimentally verified predictions are appended into (*Total_pred_verified*)dataframe.
8. *Total_pred_verified* dataframe is grouped by score and two new columns are created. "Average_Total_Predictions" column contains the average value of "Total_Predictions" column while "Average_Correctly_Verified_Predictions" column contains the average value of "Correctly_Verified_Predictions".
9. T*he new average_df* dataframe is saved into data_Total_pred_verified_total_sum file.

Output files: **data_Total_pred_verified_total**, data_Total_pred_verified_total_sum

For all Test Cases, the development of common set of MIMAT among algorithms and the positive set (***Common_TOTAL_Dataset_Extraction_among_algorithms.py***), common set of genes among all programs (***Common_TOTAL_Dataset_Extraction_Genes_among_algorithms.py***) and common set of genes among all programs and the positive set (***Common_TOTAL_Dataset_Extraction_Genes_among_algorithms.py***) is showed below:

1. Read the dataframes of all algorithms
2. Merge them together in order to create the appropriate common set.

Output files: **Common_df_alforithms_total, Common_df_algorithms_genes_total(2)** [1]

The flow chart of the procedure of analyzing the precomputed predictions of each target prediction algorithm, as described above, is shown in **Figure 63**. Furthermore, **Figure 63A, 63B** portray the formation of shared miRNA-genes interactions with or without taking into account the optimal or the aggregated score values respectively. Moreover, they present the shared miRNA-genes interactions per miRNA for each one of the aforementioned score cases.

Test Case II and III retain the same functionality as Test Case I. Certain differences include the production of the scripts which extract only the genes of the algorithms or the one of both the algorithms and Test Dataset 1, aiming in the generation of the common set of genes among algorithms or along with the positive set. Such scripts include the ***Filtering_genes_on_RNA22.py***, ***Filtering_genes_on_TargetRank.py*** and python ***Filtering_genes_on_MBSTAR.py***. Furthermore, in ***data_Total_Verified_Predictions.py*** script, useful for the generation of plots, each algorithm is filtered additionally on the common set of genes among algorithms for Test Case II and the common set of genes among all programs and the positive set for Test Case III. As a result, prior to step 4, Common_df_algorithms_genes_total file should be read and filtered with the algorithm.

The same flow charts presented in Test Case I are also designed for Test Cases II, III and are shown in **Figure 64**, **Figure 64A, 64B**.

In addition, the overall environment, depicting the entire architecture of the code for Test CasesI, II, III is presented in **Figures 65-71**.

[1] The number inside the parenthesis symbolizes the number of repetitions of the specific name in all output files.

[2] Data directory for RNA22: HomoSapiens_mRNA_ENSEMBL78_dir, for TargetRank: TargetRank_output. TargetRank_output directory is created by hand. In particular, each

human miRBase miRNA gene name has been extracted from *hsa_miRBase_miR_ranked_targets.txt* file (TargetRank precomputed predictions) and entered into the web application due to the fact that the aforementioned file did not contained the score value. The output of the web server is saved into TargetRank_output directory.

[3] The precomputed predictions of MBSTAR contain 20 directories of predictions.

[4] For each program, maximum score refers to the optimized score for each program. TargetScan, PACCMIT and MirSVR contain negatives score values and their minimum ones constitute their optimized score. On the other hand, PACCMIT-CDS and RNA22 consist of positive score values, with the best being the one close to 0. Furthermore, DIANA-microT-CDS, MIRZA-G, TargetRank and MBSTAR hold positive score values, having their maximum as their top score.

Figure 63. Flow chart of analyzing and processing the precomputed predictions of each target prediction algorithm in Test Case I.

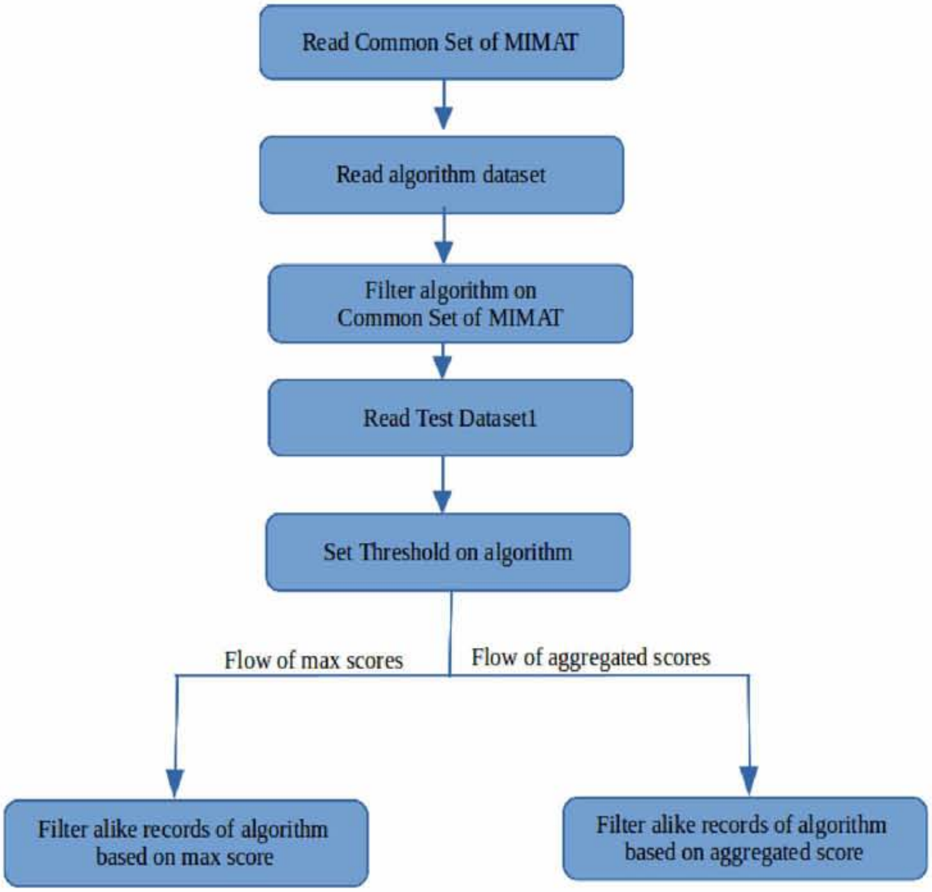

Read Common Set of MIMAT
Read algorithm dataset
Filter algorithm on Common Set of MIMAT
Read Test Dataset1
Set Threshold on algorithm
Flow of max scores
Flow of aggregated scores
Filter alike records of algorithm based on max score
Filter alike records of algorithm based on aggregated score

Figure 63A. Generation of shared miRNA-genes interactions with or without taking into account the optimal score values. In addition, shared miRNA-genes interactions per miRNA are illustrated in Test Case I.

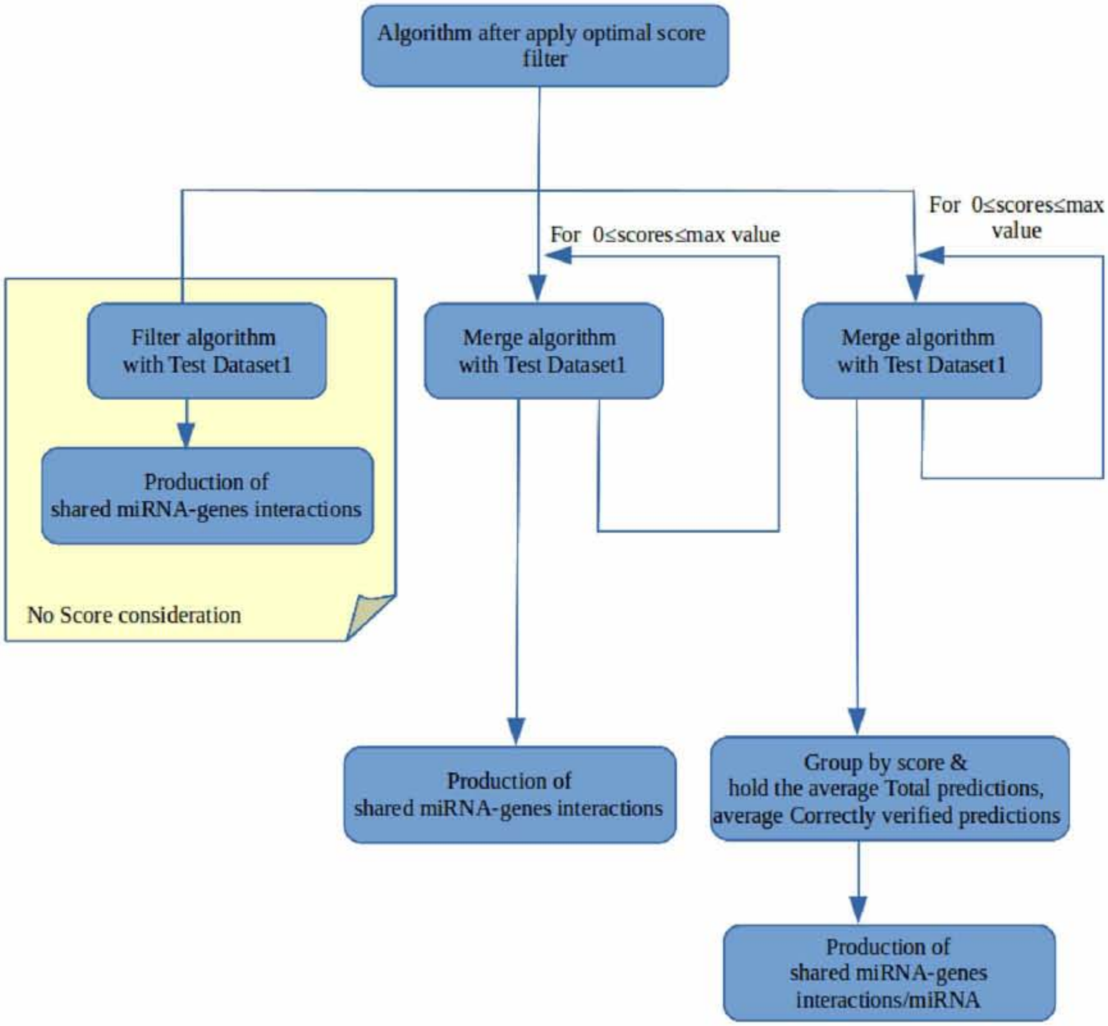

Algorithm after apply optimal score filter
For 0≤scores≤max value
For 0≤scores≤max value
Filter algorithm with Test Dataset1
Merge algorithm with Test Dataset1
Merge algorithm with Test Dataset1
Production of shared miRNA-genes interactions
No Score consideration
Production of shared miRNA-genes interactions
Group by score & hold the average Total predictions, average Correctly verified predictions
Production of shared miRNA-genes interactions/miRNA

Figure 63B. Generation of shared miRNA-genes interactions with or without taking into account the aggregated score values. In addition, shared miRNA-genes interactions per miRNA are illustrated in Test Case I.

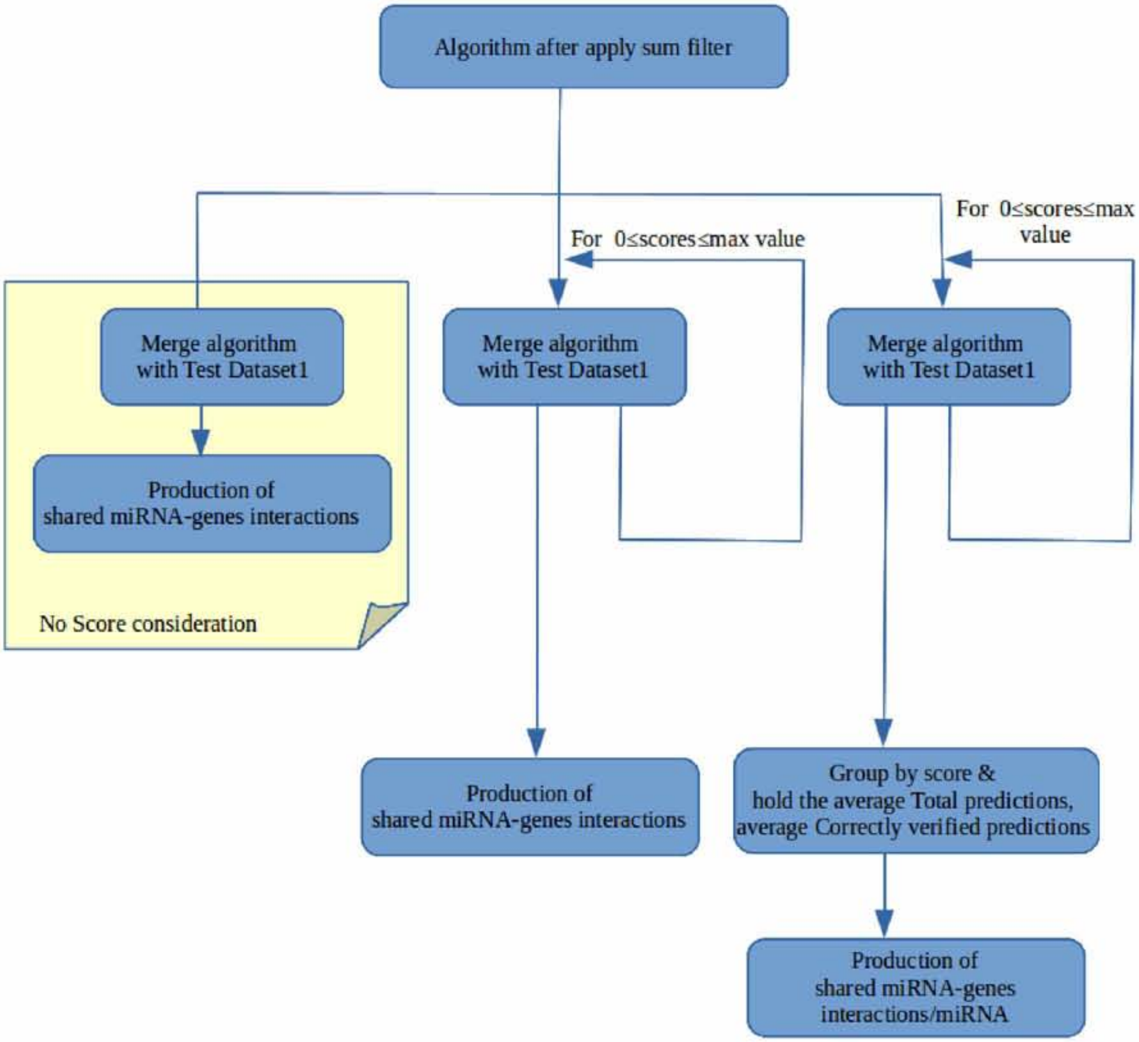

Algorithm after apply sum filter
For 0≤scores≤max value
For 0≤scores≤max value
Merge algorithm with Test Dataset1
Production of shared miRNA-genes interactions
No Score consideration
Merge algorithm with Test Dataset1
Merge algorithm with Test Dataset1
Production of shared miRNA-genes interactions
Group by score & hold the average Total predictions, average Correctly verified predictions
Production of shared miRNA-genes interactions/miRNA

Figure 64. Flow chart of analyzing and processing the precomputed predictions of each target prediction algorithm in Test Cases II, III.

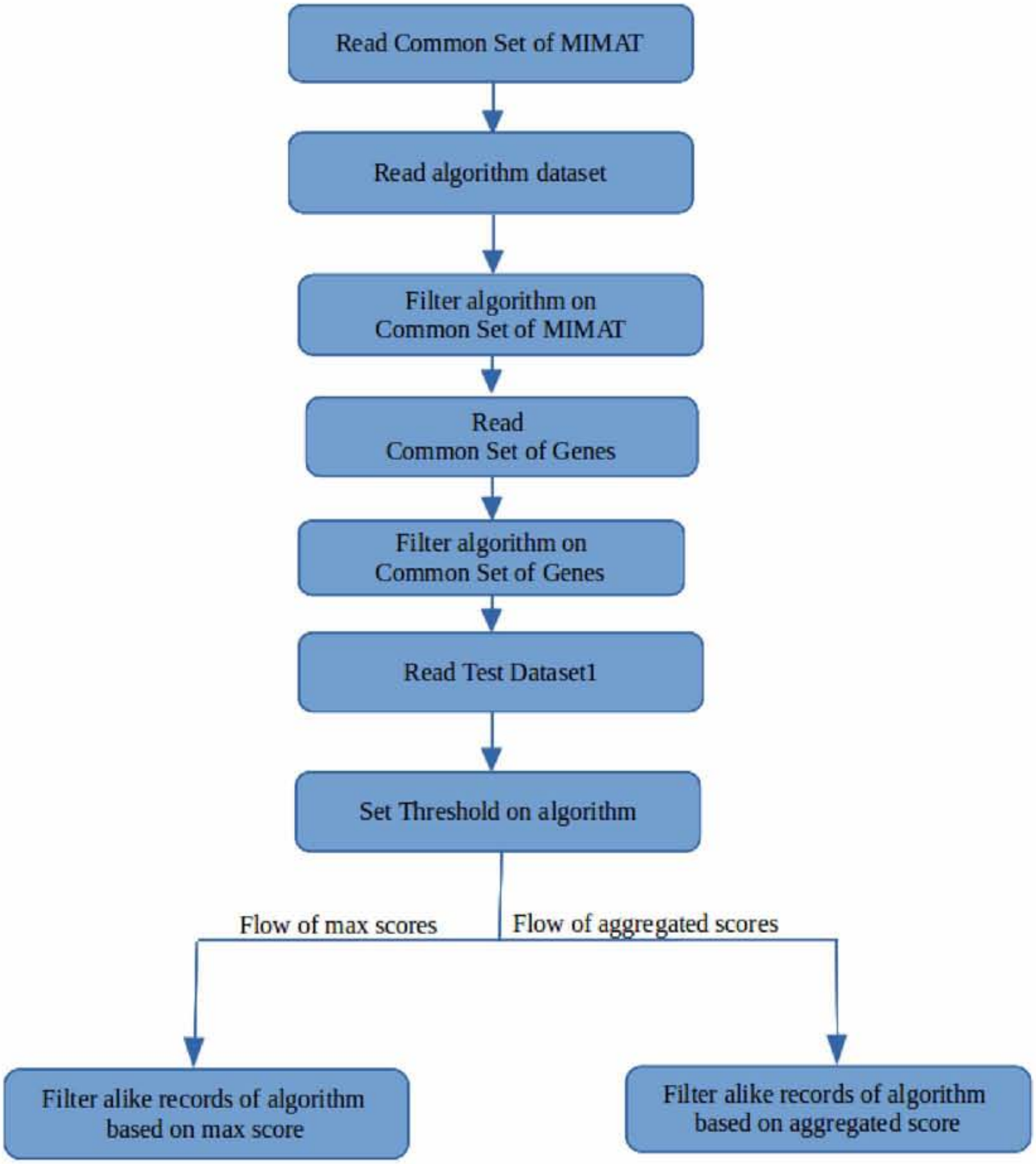

Read Common Set of MIMAT
Read algorithm dataset
Filter algorithm on Common Set of MIMAT
Read Common Set of Genes
Filter algorithm on Common Set of Genes
Read Test Dataset1
Set Threshold on algorithm
Flow of max scores
Flow of aggregated scores
Filter alike records of algorithm based on max score
Filter alike records of algorithm based on aggregated score

Figure 64A. Generation of shared miRNA-genes interactions with or without taking into account the optimal score values. In addition, shared miRNA-genes interactions per miRNA are illustrated in Test Case II, III.

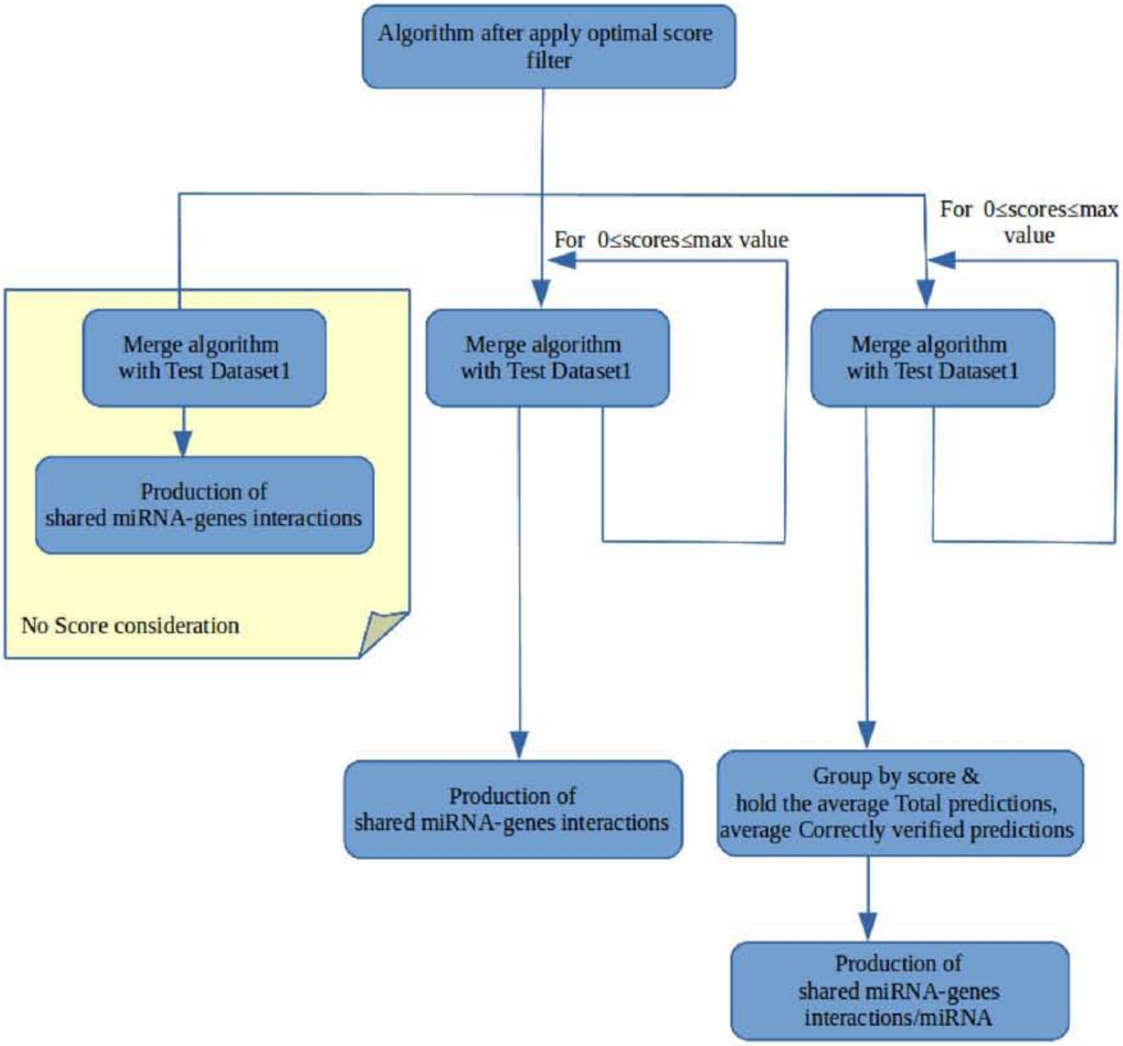

Algorithm after apply optimal score filter
For 0≤scores≤max value
For 0≤scores≤max value
Merge algorithm with Test Dataset1
Production of shared miRNA-genes interactions
No Score consideration
Merge algorithm with Test Dataset1
Merge algorithm with Test Dataset1
Production of shared miRNA-genes interactions
Group by score & hold the average Total predictions, average Correctly verified predictions
Production of shared miRNA-genes interactions/miRNA

Figure 64B. Generation of shared miRNA-genes interactions with or without taking into account the aggregated score values. In addition, shared miRNA-genes interactions per miRNA are illustrated in Test Cases II, III.

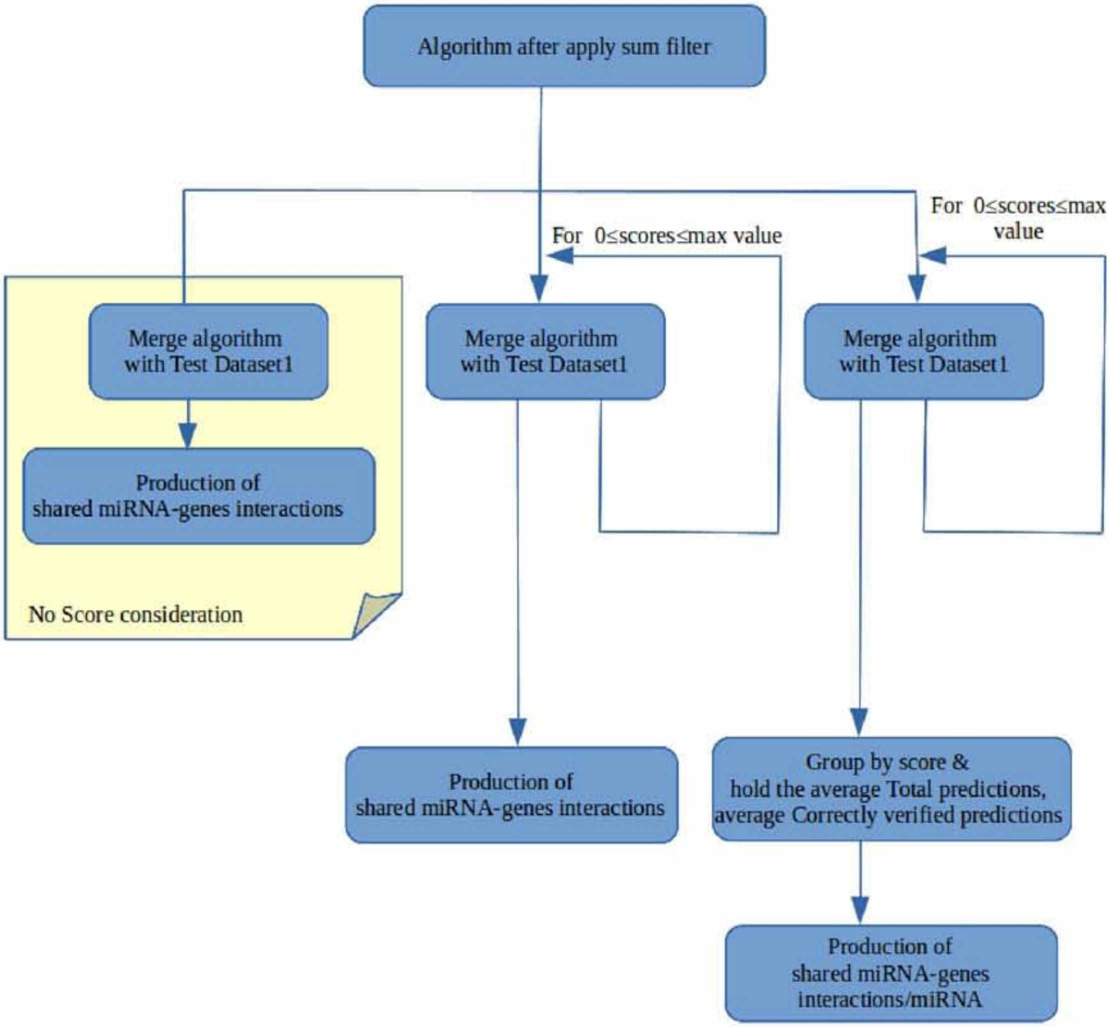

Algorithm after apply sum filter
For 0≤scores≤max value
For 0≤scores≤max value
Merge algorithm with Test Dataset1
Production of shared miRNA-genes interactions
No Score consideration
Merge algorithm with Test Dataset1
Merge algorithm with Test Dataset1
Production of shared miRNA-genes interactions
Group by score & hold the average Total predictions, average Correctly verified predictions
Production of shared miRNA-genes interactions/miRNA

Figure 65. Flow chart of Test Case I

TEST CASE I

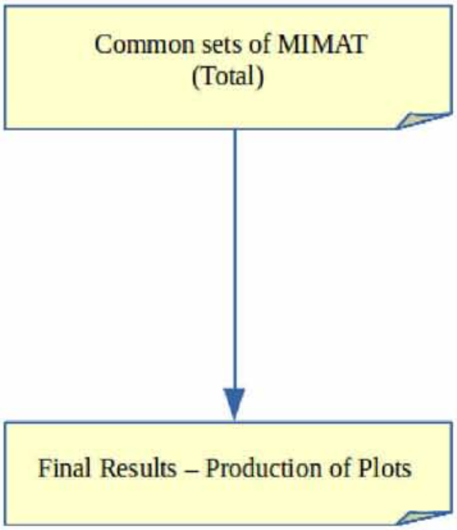


Figure 66. Flow chart of Test Case II

TEST CASE II

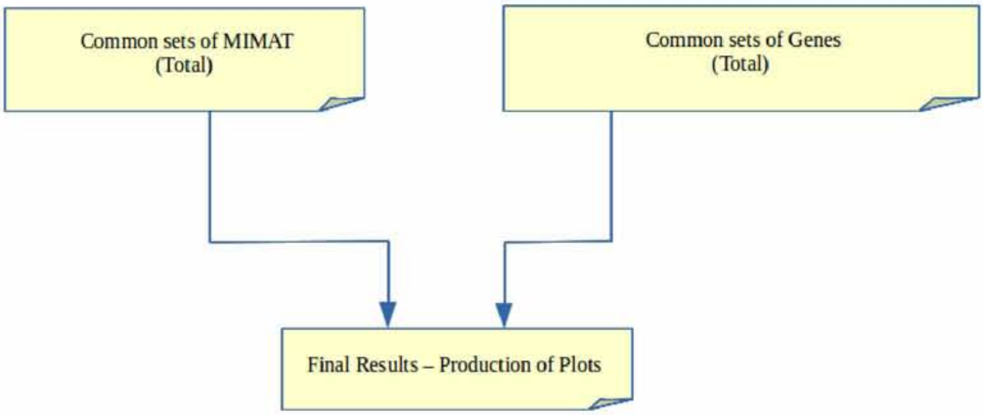


Figure 67. Flow chart of Test Case III

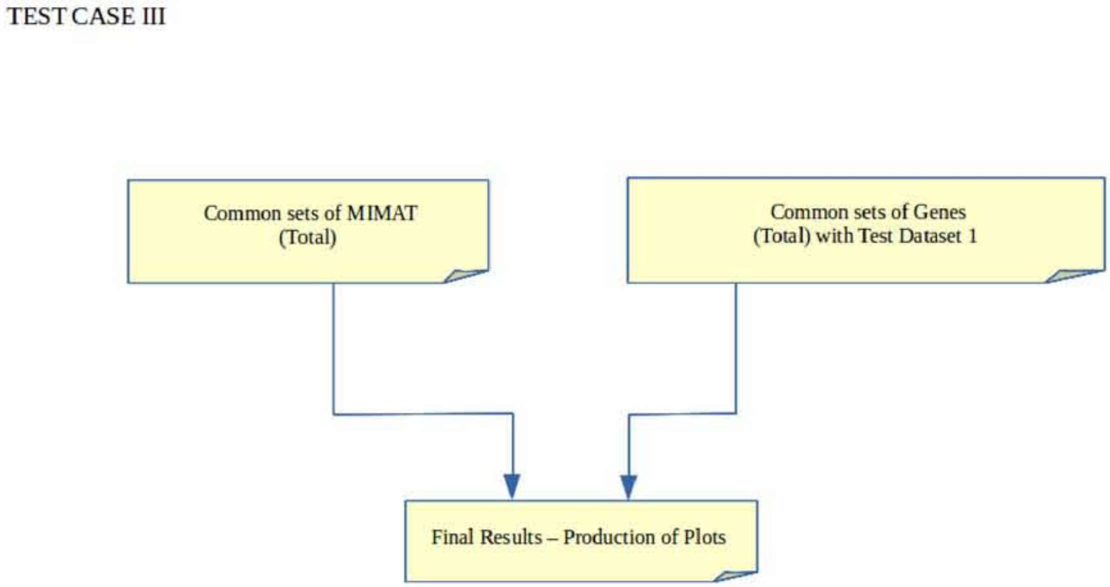

TEST CASE III
Common sets of MIMAT (Total)
Common sets of Genes (Total) with Test Dataset 1
Final Results – Production of Plots

Figure 68. Production of the Common set of MIMAT among all algorithms and Test Dataset 1 (Test Cases I, II, III).

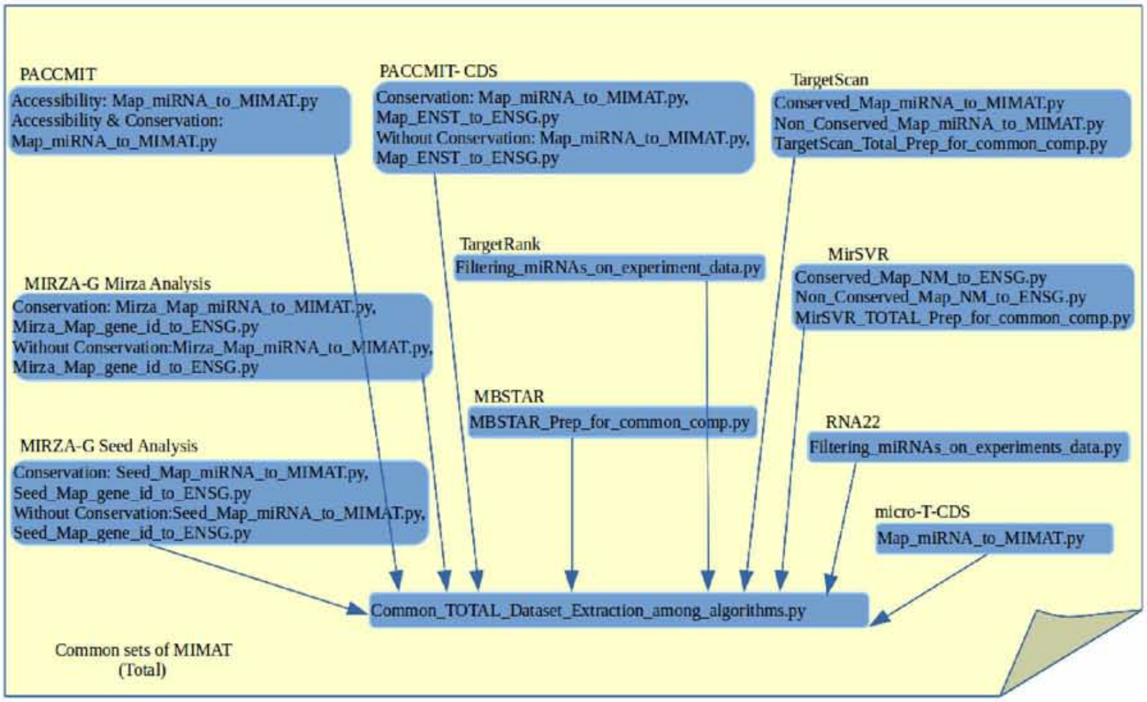

PACCMIT
Accessibility: Map_miRNA_to_MIMAT.py
Accessibility & Conservation:
Map_miRNA_to_MIMAT.py
PACCMIT- CDS
Conservation: Map_miRNA_to_MIMAT.py,
Map_ENST_to_ENSG.py
Without Conservation: Map_miRNA_to_MIMAT.py,
Map_ENST_to_ENSG.py
TargetScan
Conserved_Map_miRNA_to_MIMAT.py
Non_Conserved_Map_miRNA_to_MIMAT.py
TargetScan_Total_Prep_for_common_comp.py
TargetRank
Filtering_miRNAs_on_experiment_data.py
MirSVR
Conserved_Map_NM_to_ENSG.py
Non_Conserved_Map_NM_to_ENSG.py
MirSVR_TOTAL_Prep_for_common_comp.py
MIRZA-G Mirza Analysis
Conservation: Mirza_Map_miRNA_to_MIMAT.py,
Mirza_Map_gene_id_to_ENSG.py
Without Conservation:Mirza_Map_miRNA_to_MIMAT.py,
Mirza_Map_gene_id_to_ENSG.py
MBSTAR
MBSTAR_Prep_for_common_comp.py
RNA22
Filtering_miRNAs_on_experiments_data.py
MIRZA-G Seed Analysis
Conservation: Seed_Map_miRNA_to_MIMAT.py,
Seed_Map_gene_id_to_ENSG.py
Without Conservation:Seed_Map_miRNA_to_MIMAT.py,
Seed_Map_gene_id_to_ENSG.py
micro-T-CDS
Map_miRNA_to_MIMAT.py
Common_TOTAL_Dataset_Extraction_among_algorithms.py
Common sets of MIMAT
(Total)

Figure 69. Production of the Common set of Genes among all algorithms (Test Cases I, II, III).

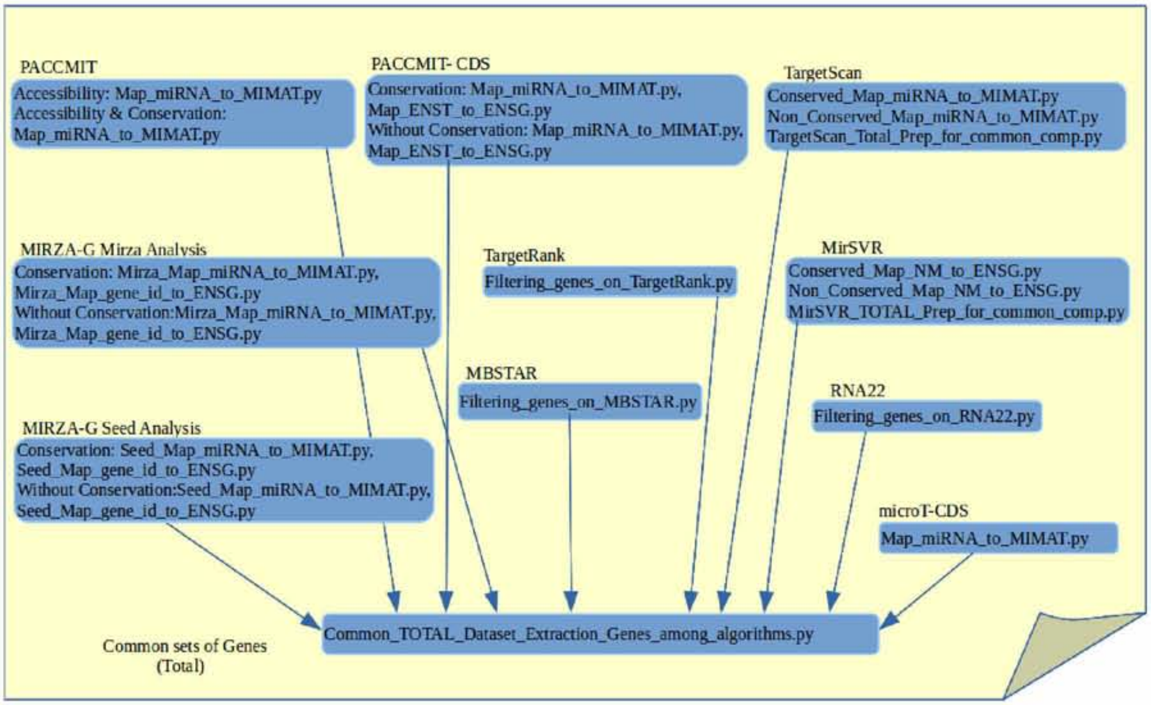

PACCMIT
Accessibility: Map_miRNA_to_MIMAT.py
Accessibility & Conservation:
Map_miRNA_to_MIMAT.py
PACCMIT- CDS
Conservation: Map_miRNA_to_MIMAT.py,
Map_ENST_to_ENSG.py
Without Conservation: Map_miRNA_to_MIMAT.py,
Map_ENST_to_ENSG.py
TargetScan
Conserved_Map_miRNA_to_MIMAT.py
Non_Conserved_Map_miRNA_to_MIMAT.py
TargetScan_Total_Prep_for_common_comp.py
MIRZA-G Mirza Analysis
Conservation: Mirza_Map_miRNA_to_MIMAT.py,
Mirza_Map_gene_id_to_ENSG.py
Without Conservation:Mirza_Map_miRNA_to_MIMAT.py,
Mirza_Map_gene_id_to_ENSG.py
TargetRank
Filtering_genes_on_TargetRank.py
MirSVR
Conserved_Map_NM_to_ENSG.py
Non_Conserved_Map_NM_to_ENSG.py
MirSVR_TOTAL_Prep_for_common_comp.py
MBSTAR
Filtering_genes_on_MBSTAR.py
RNA22
Filtering_genes_on_RNA22.py
MIRZA-G Seed Analysis
Conservation: Seed_Map_miRNA_to_MIMAT.py,
Seed_Map_gene_id_to_ENSG.py
Without Conservation:Seed_Map_miRNA_to_MIMAT.py,
Seed_Map_gene_id_to_ENSG.py
microT-CDS
Map_miRNA_to_MIMAT.py
Common_TOTAL_Dataset_Extraction_Genes_among_algorithms.py
Common sets of Genes
(Total)

Figure 70. Production of the Common set of Genes among all algorithms and Test Dataset 1(Test Cases I, II, III).

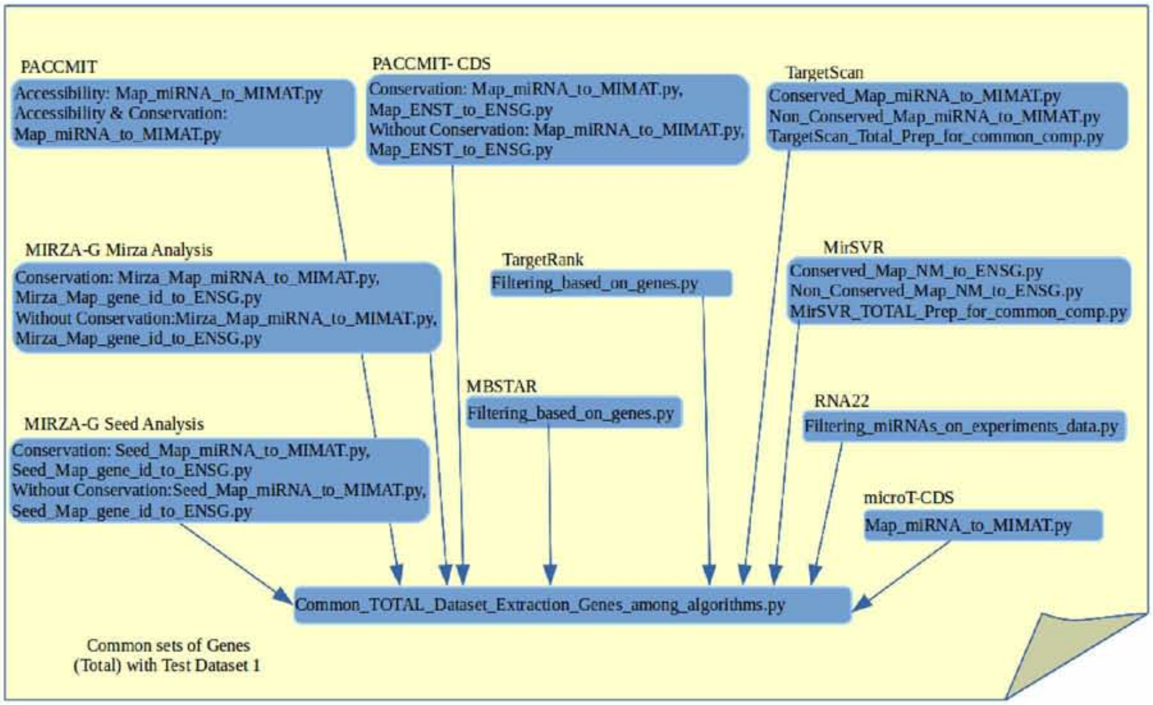
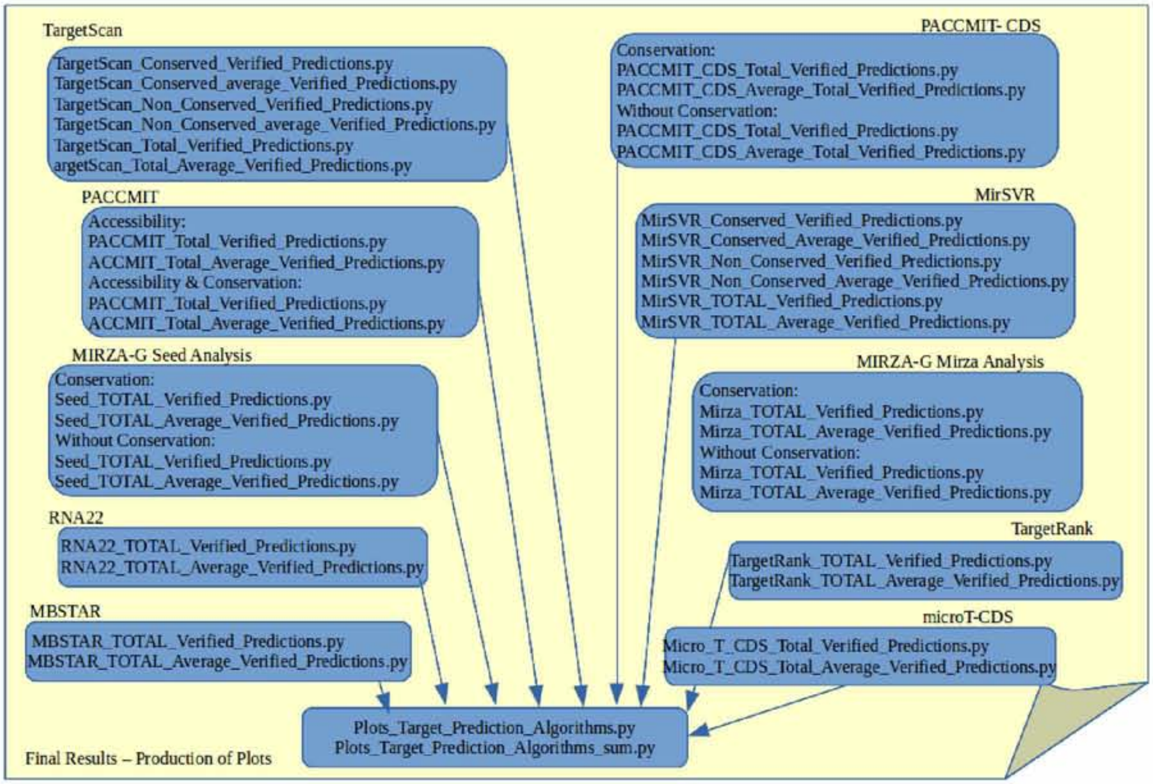

PACCMIT
Accessibility: Map_miRNA_to_MIMAT.py
Accessibility & Conservation:
Map_miRNA_to_MIMAT.py
PACCMIT- CDS
Conservation: Map_miRNA_to_MIMAT.py,
Map_ENST_to_ENSG.py
Without Conservation: Map_miRNA_to_MIMAT.py,
Map_ENST_to_ENSG.py
TargetScan
Conserved_Map_miRNA_to_MIMAT.py
Non_Conserved_Map_miRNA_to_MIMAT.py
TargetScan_Total_Prep_for_common_comp.py
MIRZA-G Mirza Analysis
Conservation: Mirza_Map_miRNA_to_MIMAT.py,
Mirza_Map_gene_id_to_ENSG.py
Without Conservation:Mirza_Map_miRNA_to_MIMAT.py,
Mirza_Map_gene_id_to_ENSG.py
TargetRank
Filtering_based_on_genes.py
MirSVR
Conserved_Map_NM_to_ENSG.py
Non_Conserved_Map_NM_to_ENSG.py
MirSVR_TOTAL_Prep_for_common_comp.py
MBSTAR
Filtering_based_on_genes.py
RNA22
Filtering_miRNAs_on_experiments_data.py
MIRZA-G Seed Analysis
Conservation: Seed_Map_miRNA_to_MIMAT.py,
Seed_Map_gene_id_to_ENSG.py
Without Conservation:Seed_Map_miRNA_to_MIMAT.py,
Seed_Map_gene_id_to_ENSG.py
microT-CDS
Map_miRNA_to_MIMAT.py
Common_TOTAL_Dataset_Extraction_Genes_among_algorithms.py
Common sets of Genes
(Total) with Test Dataset 1

Figure 71. Production of Plots for Test Cases I, II, III


TargetScan
TargetScan_Conserved_Verified_Predictions.py
TargetScan_Conserved_average_Verified_Predictions.py
TargetScan_Non_Conserved_Verified_Predictions.py
TargetScan_Non_Conserved_average_Verified_Predictions.py
TargetScan_Total_Verified_Predictions.py
argetScan_Total_Average_Verified_Predictions.py
PACCMIT
Accessibility:
PACCMIT_Total_Verified_Predictions.py
ACCMIT_Total_Average_Verified_Predictions.py
Accessibility & Conservation:
PACCMIT_Total_Verified_Predictions.py
ACCMIT_Total_Average_Verified_Predictions.py
MIRZA-G Seed Analysis
Conservation:
Seed_TOTAL_Verified_Predictions.py
Seed_TOTAL_Average_Verified_Predictions.py
Without Conservation:
Seed_TOTAL_Verified_Predictions.py
Seed_TOTAL_Average_Verified_Predictions.py
RNA22
RNA22_TOTAL_Verified_Predictions.py
RNA22_TOTAL_Average_Verified_Predictions.py
MBSTAR
MBSTAR_TOTAL_Verified_Predictions.py
MBSTAR_TOTAL_Average_Verified_Predictions.py
PACCMIT- CDS
Conservation:
PACCMIT_CDS_Total_Verified_Predictions.py
PACCMIT_CDS_Average_Total_Verified_Predictions.py
Without Conservation:
PACCMIT_CDS_Total_Verified_Predictions.py
PACCMIT_CDS_Average_Total_Verified_Predictions.py
MirSVR
MirSVR_Conserved_Verified_Predictions.py
MirSVR_Conserved_Average_Verified_Predictions.py
MirSVR_Non_Conserved_Verified_Predictions.py
MirSVR_Non_Conserved_Average_Verified_Predictions.py
MirSVR_TOTAL_Verified_Predictions.py
MirSVR_TOTAL_Average_Verified_Predictions.py
MIRZA-G Mirza Analysis
Conservation:
Mirza_TOTAL_Verified_Predictions.py
Mirza_TOTAL_Average_Verified_Predictions.py
Without Conservation:
Mirza_TOTAL_Verified_Predictions.py
Mirza_TOTAL_Average_Verified_Predictions.py
TargetRank
TargetRank_TOTAL_Verified_Predictions.py
TargetRank_TOTAL_Average_Verified_Predictions.py
microT-CDS
Micro_T_CDS_Total_Verified_Predictions.py
Micro_T_CDS_Total_Average_Verified_Predictions.py
Plots_Target_Prediction_Algorithms.py
Plots_Target_Prediction_Algorithms_sum.py
Final Results – Production of Plots

Binding Sites – Test Case II

As far as the analysis of binding sites is concerned, for TargetScan, MirSVR, RNA22, MBSTAR and DIANA Micro-T-CDS, the following steps are pursued:

In script ***Map_hg19_to_hg38_coord.py***:

1. Read the data, including the coordinates, into a dataframe. In case the size of data files is very large and cannot be loaded entirely in memory, it is read in chunks of 10^6.
2. Remove the records, which involve chromosomes different than 1-22, X, Y, M, MT.
3. Convert the coordinates in the form chr: start position – end position.
4. Export hg19_to_hg38_liftover file.

Ouput files: **hg19_to_hg38_liftover**

Script **Map_miRNA_to_MIMAT.py** holds alike functionality as described in the previous section.

In script **data_Total_Verified_Predictions.py** the generic steps below are implemented:

1. Read Common_df_alforithms_total file into a dataframe.
2. Read algorithm_input into data_TOTAL dataframe
3. Merge (inner) Common_df_alforithms_total and data_TOTAL dataframes in order to filter the records of each algorithm on the MIMATs of Test Dataset 2 and all algorithms. The result is stored in data_TOTAL dataframe.
4. Read Common_TOTAL_Dataset_Extraction_Genes_among_algorithms.py into a dataframe.
5. Merge Common_TOTAL_Dataset_Extraction_Genes_among_algorithms and data_TOTAL dataframes in order to filter the records of each algorithm on the set of Gene Stable ID of all algorithms. The result is stored in data_TOTAL dataframe.
6. In case of algorithms, which are read in chunks, due to their large size, duplicate records are removed.
7. Read hglft_genome_bed file into a dataframe. This file is derived from LiftOver tool through the conversion of genome coordinates from hg19 assembly to hg38 assembly.
8. Map old genome coordinates to the new ones (in assembly hg38).
9. Set threshold on data_TOTAL dataframe.
10. Read the positive set file (Test Dataset 2) into a dataframe (experiment_data).
11. Filter and overlapping sites/miRNA and retain the site with the optimal score.
12. Create 2 flows of code.
    - 12.1. Filter solely on the overlapping regions of algorithm with Test Dataset 2, not on MIMAT. In this case, the overlapping regions of the algorithm are chosen.
        - 12.1.1.1. Merge (inner) data_TOTAL and experiment_data on MIMAT, chr, strand.
        - 12.1.1.2. Find the percentage overlap between binding sites of algorithm and Test Dataset 2.
        - 12.1.1.3. Remove records that percentage overlap equals to 0. The overlapping regions of Test Dataset 2are chosen.
        - 12.1.1.4. Filter overlapping sites of the algorithm with the correctly verified sites and retain the ones containing the optimal score. The results are saved in Verified_pred_per_loop_total file.

12.1.1.5. Export data_TOTAL_first_comp file along with the number of total predicted sites andexperimentally verified sites into data_TOTAL_first_comp_ready file.

Steps 12.1.1.1, 12.1.1.2, 12.1.1.3, 12.1.1.4 concern the case of finding the common miRNA-sites interactions with Test Dataset 2 without taking into account the score.

12.1.2. For each score between the optimal value of data and 0 the following steps (12.1.2.1, 12.1.2.2, 12.1.2.3 and 12.1.2.4) are repeated.

12.1.2.1. Merge (inner) data_TOTAL and experiment_data on MIMAT, chr, strand.

12.1.2.2. Find the percentage overlap between binding sites of algorithm and Test Dataset 2.

12.1.2.3. Remove records that percentage overlap equals to 0. The overlapping regions of Test Dataset 2are chosen.

12.1.2.4. Filter overlapping sites of the algorithm with the correctly verified sites and retain theones containing the optimal score. The results are saved in Verified_pred_per_loop_total file.

12.1.2.5. Total predicted sites and experimentally verified sites are appended into(Total_pred_verified) dataframe.

12.1.2.6. Total_pred_verified dataframe is saved into data_Total_pred_verified_total file.

12.2. No filter applied to the overlapping regions of algorithm with Test Dataset 2.

12.2.1.1. Merge (inner) data_TOTAL and experiment_data on MIMAT, chr, strand.

12.2.1.2. Find the percentage overlap between binding sites of algorithm and Test Dataset 2.

12.2.1.3. Remove records that percentage overlap equals to 0. The overlapping regions of Test Dataset 2 are chosen.

12.2.1.4. Filter overlapping sites of the algorithm with the correctly verified sites and retain the ones containing the optimal score. The results are saved in Verified_pred_per_loop_total file.

12.2.1.5. Export data_TOTAL_first_comp file along with the number of total predicted sites and experimentally verified sites into data_TOTAL_first_comp_ready file.

Steps 12.2.1.1, 12.2.1.2, 12.2.1.3, 12.2.1.4 concern the case of finding the common miRNA - sites interactions with Test Dataset 2 without taking into account the score.

12.2.2. For each score between the optimal value of data and 0 the following steps (12.2.2.1, 12.2.2.2, 12.2.2.3and 12.2.2.4, 12.2.2.5) are repeated.

12.2.2.1. Merge (inner) data_TOTAL and experiment_data on MIMAT, chr, strand.

12.2.2.2. Find the percentage overlap between binding sites of algorithm and Test Dataset 2.

12.2.2.3. Remove records that percentage overlap equals to 0. The overlapping regions of Test Dataset 2are chosen.

12.2.2.4. Filter overlapping sites of the algorithm with the correctly verified sites and retain theones containing the optimal score. The results are saved in Verified_pred_per_loop_total file.

12.2.2.5. Total predicted sites and experimentally verified sites are appended into(Total_pred_verified) dataframe.

12.2.2.6. Total_pred_verified dataframe is saved into data_Total_pred_verified_total file.

Output files: **Verified_pred_per_loop_total** (2), **data_Total_pred_verified_total** (2)

It is important to mention that thresholds are set to algorithms such as RNA22, MBSTAR and DIANA-microT-CDS (step 9).

The flow chart of the procedure of analyzing the precomputed predictions of each target prediction algorithm, as described above, is shown in **Figure 72**. Furthermore, **Figure 72A, 72B** portray the formation of shared miRNA-site interactions with or without filtering the overlapping sites per miRNA respectively.

In addition, the overall environment, depicting the entire architecture of the code for Test Case II as far as binding sites analysis is concerned, is presented in **Figures 73-76**.

Figure 72. Flow chart of analyzing and processing the precomputed predictions of each target prediction algorithm in Test Case II when binding sites are examined.

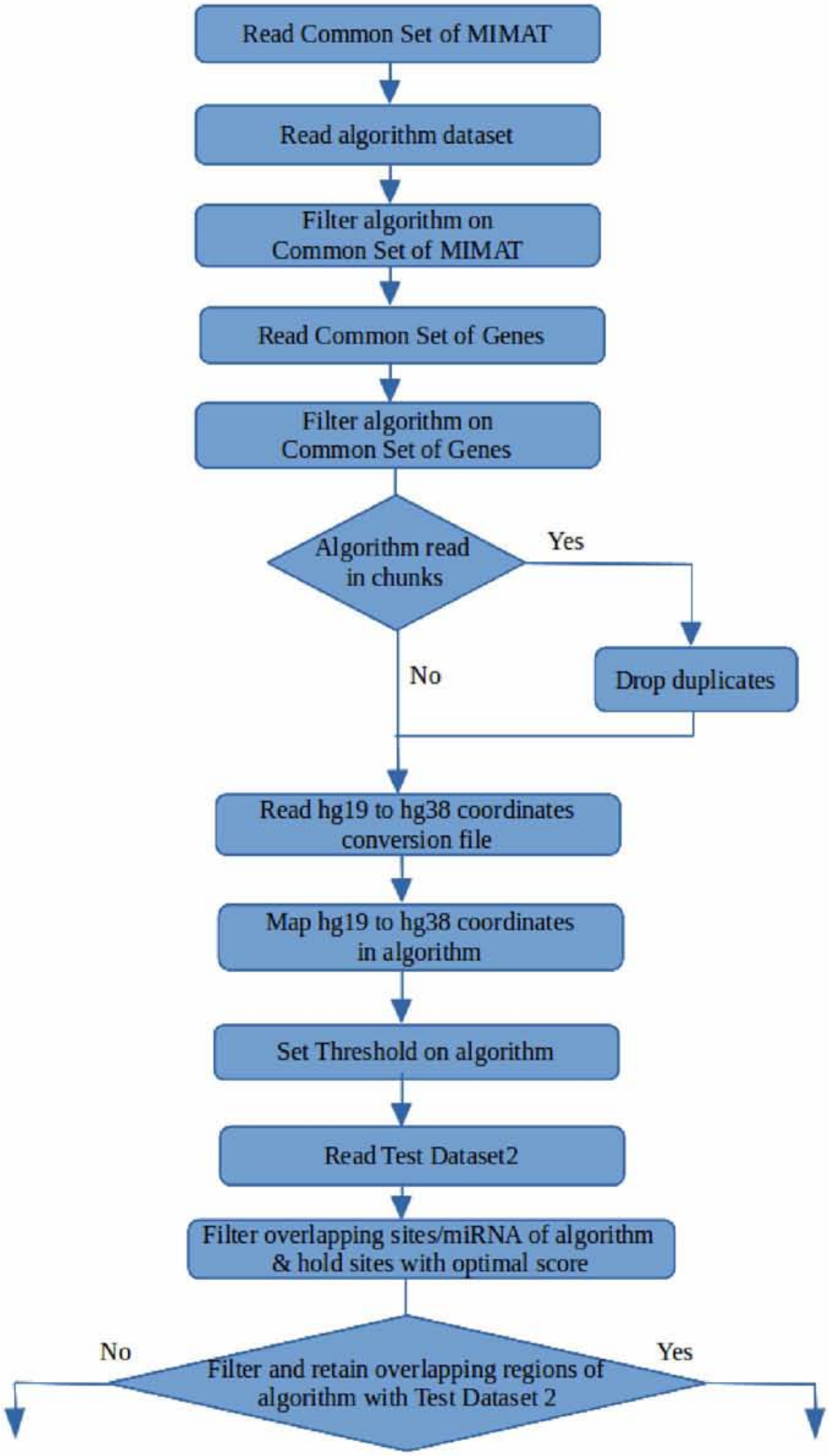

Read Common Set of MIMAT
Read algorithm dataset
Filter algorithm on Common Set of MIMAT
Read Common Set of Genes
Filter algorithm on Common Set of Genes
Algorithm read in chunks
Yes
Drop duplicates
No
Read hg19 to hg38 coordinates conversion file
Map hg19 to hg38 coordinates in algorithm
Set Threshold on algorithm
Read Test Dataset2
Filter overlapping sites/miRNA of algorithm & hold sites with optimal score
No
Filter and retain overlapping regions of algorithm with Test Dataset 2
Yes

Figure 72A. Generation of shared miRNA-site interactions with or without taking into account the optimal score values in Test Case II. All overlapping regions between each program and Test Dataset 2 are retained.

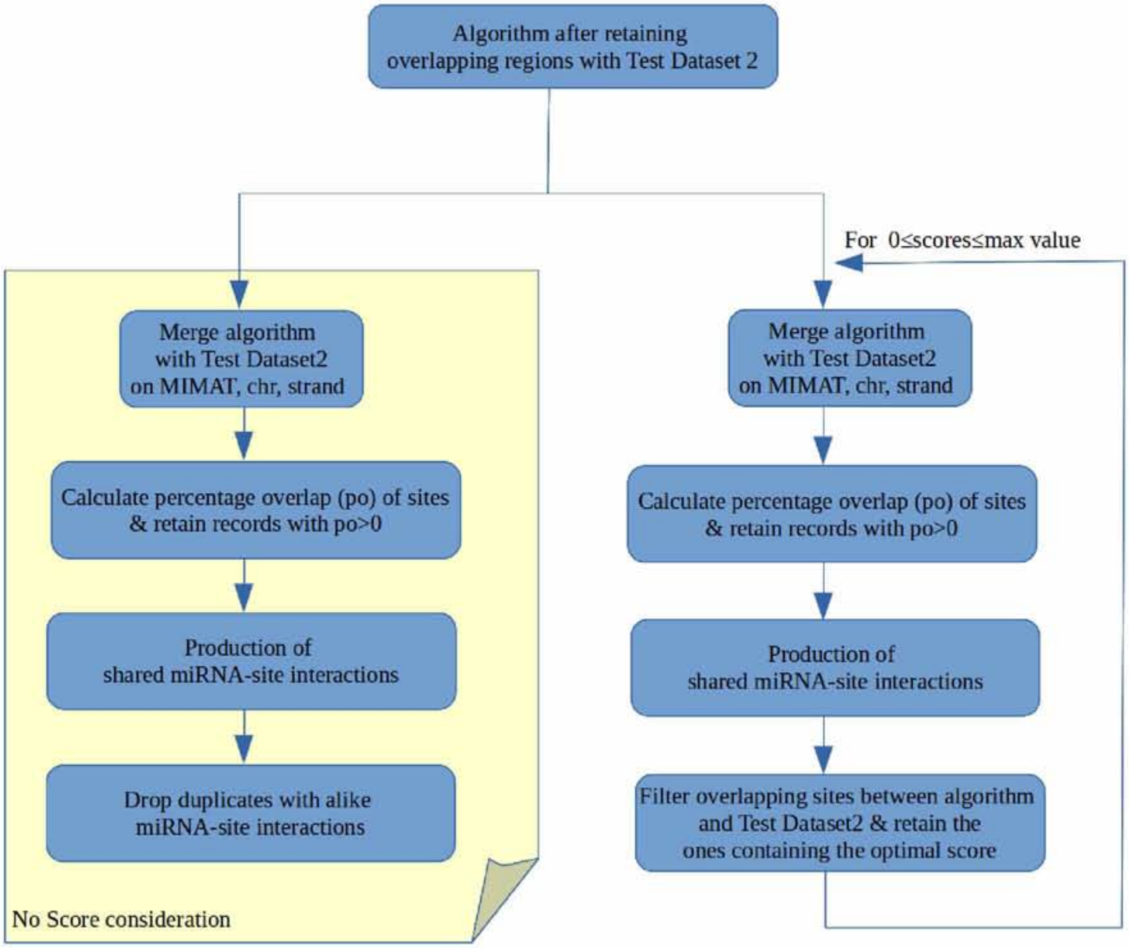

Algorithm after retaining overlapping regions with Test Dataset 2
Merge algorithm with Test Dataset2 on MIMAT, chr, strand
Calculate percentage overlap (po) of sites & retain records with po>0
Production of shared miRNA-site interactions
Drop duplicates with alike miRNA-site interactions
No Score consideration
For 0≤scores≤max value
Merge algorithm with Test Dataset2 on MIMAT, chr, strand
Calculate percentage overlap (po) of sites & retain records with po>0
Production of shared miRNA-site interactions
Filter overlapping sites between algorithm and Test Dataset2 & retain the ones containing the optimal score

Figure 72B. Generation of shared miRNA-site interactions with or without taking into account the optimal score values in Test Case II.

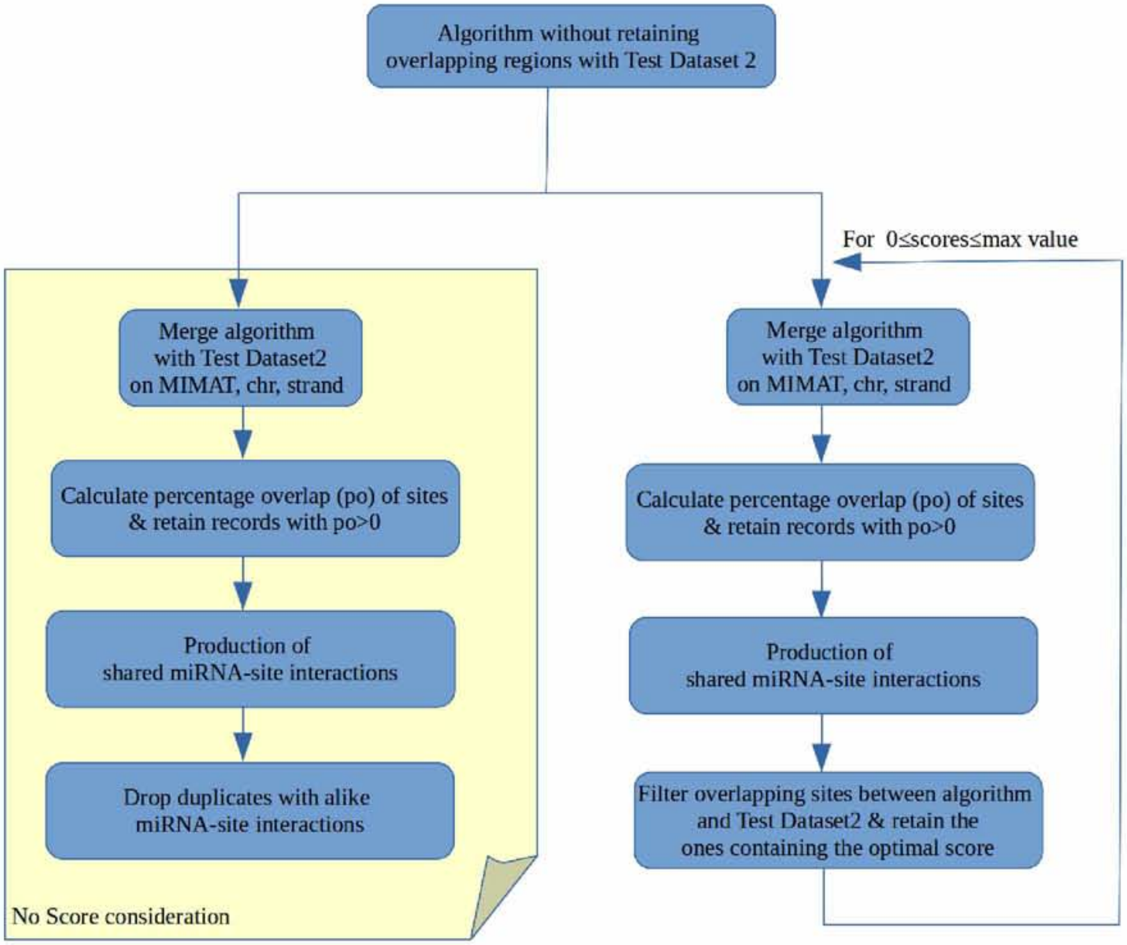

Algorithm without retaining overlapping regions with Test Dataset 2
Merge algorithm with Test Dataset2 on MIMAT, chr, strand
Calculate percentage overlap (po) of sites & retain records with po>0
Production of shared miRNA-site interactions
Drop duplicates with alike miRNA-site interactions
No Score consideration
For 0≤scores≤max value
Merge algorithm with Test Dataset2 on MIMAT, chr, strand
Calculate percentage overlap (po) of sites & retain records with po>0
Production of shared miRNA-site interactions
Filter overlapping sites between algorithm and Test Dataset2 & retain the ones containing the optimal score

Figure 73. Flow chart of Test Case II for binding sites analysis

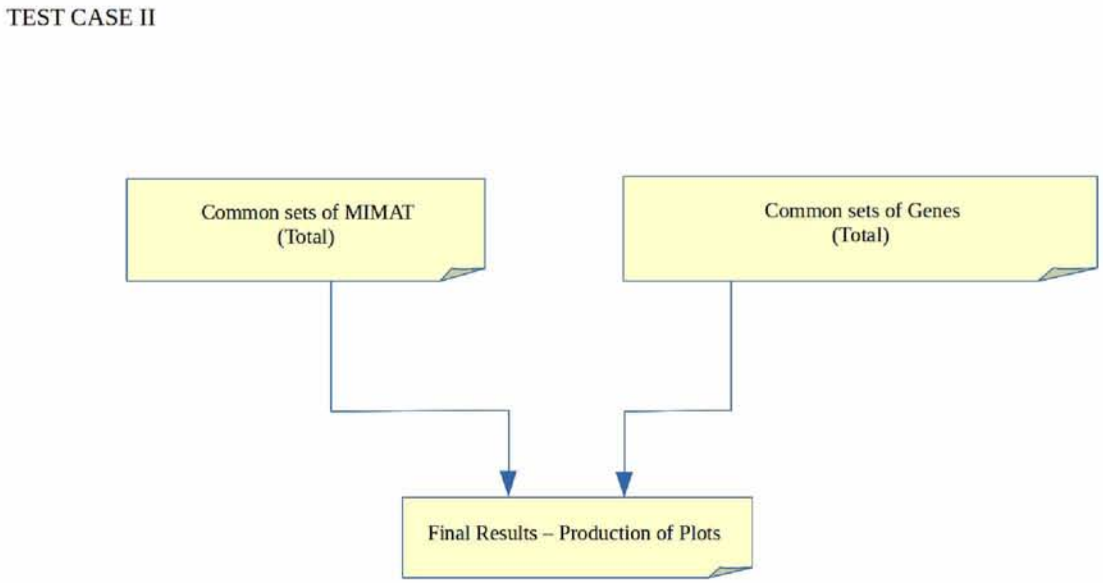

TEST CASE II
Common sets of MIMAT (Total)
Common sets of Genes (Total)
Final Results – Production of Plots

Figure 74. Production of the Common set of MIMAT among all algorithms and Test Dataset 2 in TestCase II for binding sites analysis.

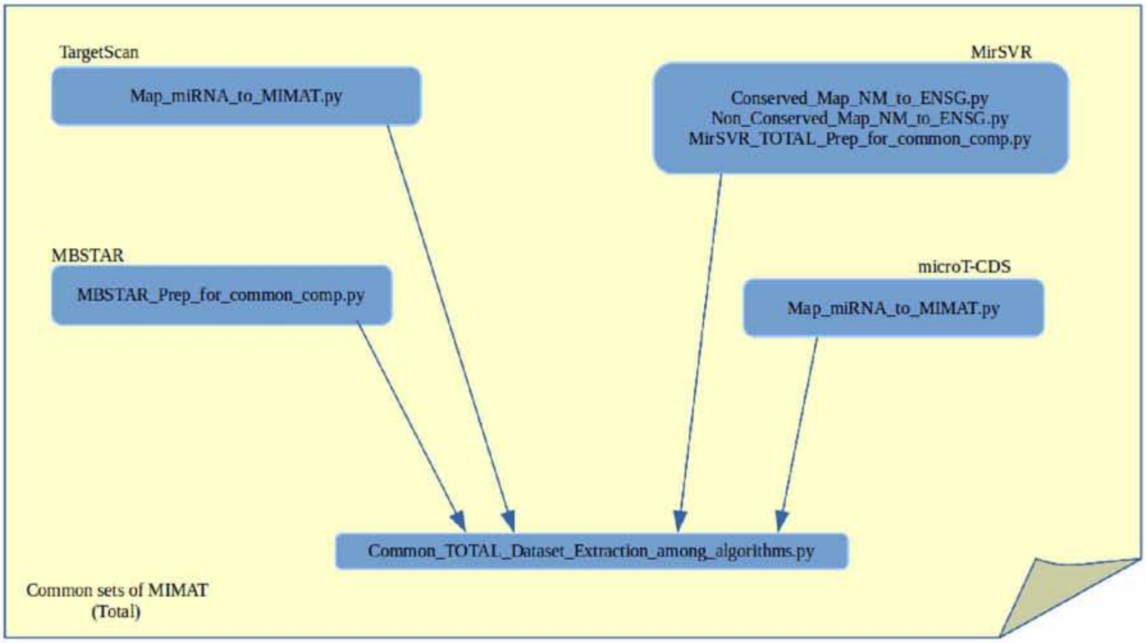

TargetScan
Map_miRNA_to_MIMAT.py
MirSVR
Conserved_Map_NM_to_ENSG.py
Non_Conserved_Map_NM_to_ENSG.py
MirSVR_TOTAL_Prep_for_common_comp.py
MBSTAR
MBSTAR_Prep_for_common_comp.py
microT-CDS
Map_miRNA_to_MIMAT.py
Common_TOTAL_Dataset_Extraction_among_algorithms.py
Common sets of MIMAT (Total)

Figure 75. Production of the Common set of Genes among all algorithms in Test Case II for binding sites analysis.

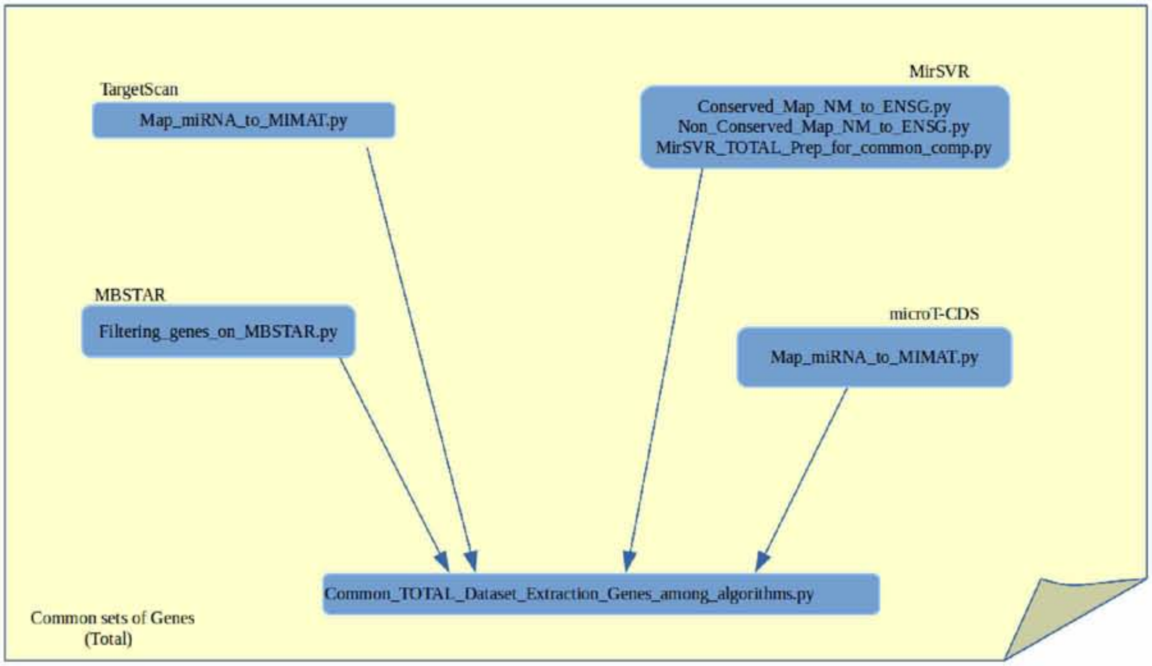
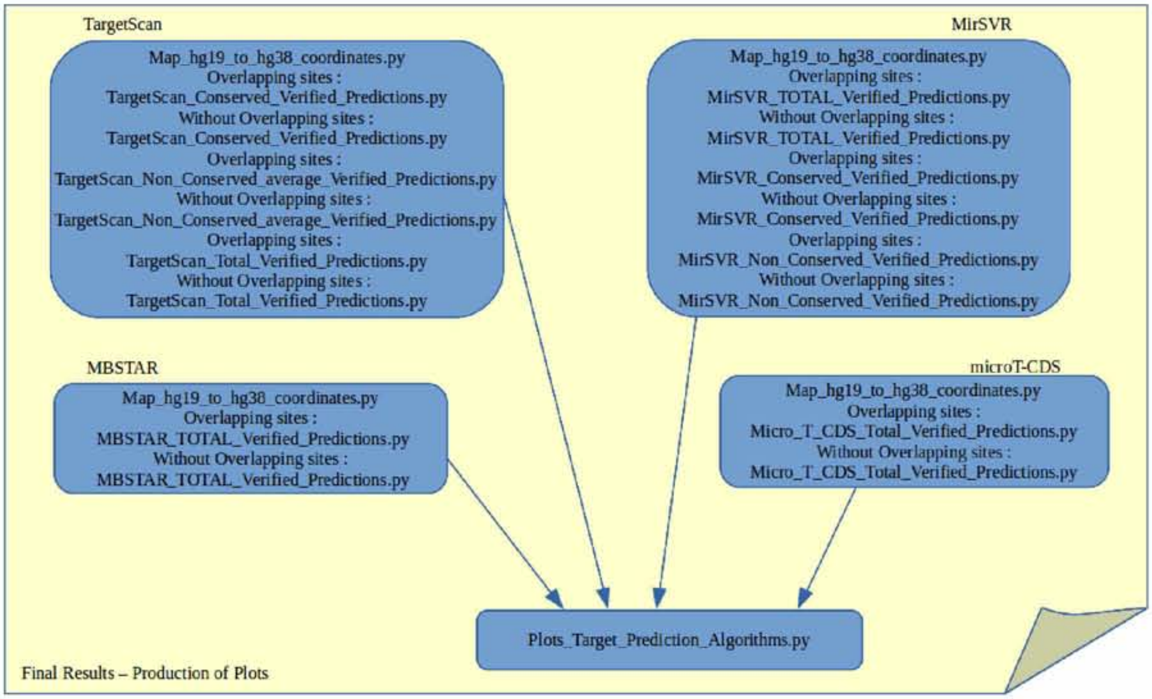

TargetScan
Map_miRNA_to_MIMAT.py
MirSVR
Conserved_Map_NM_to_ENSG.py
Non_Conserved_Map_NM_to_ENSG.py
MirSVR_TOTAL_Prep_for_common_comp.py
MBSTAR
Filtering_genes_on_MBSTAR.py
microT-CDS
Map_miRNA_to_MIMAT.py
Common_TOTAL_Dataset_Extraction_Genes_among_algorithms.py
Common sets of Genes
(Total)

Figure 76. Production of Plots for Test Cases II for binding sites analysis


TargetScan
Map_hg19_to_hg38_coordinates.py
Overlapping sites :
TargetScan_Conserved_Verified_Predictions.py
Without Overlapping sites :
TargetScan_Conserved_Verified_Predictions.py
Overlapping sites :
TargetScan_Non_Conserved_average_Verified_Predictions.py
Without Overlapping sites :
TargetScan_Non_Conserved_average_Verified_Predictions.py
Overlapping sites :
TargetScan_Total_Verified_Predictions.py
Without Overlapping sites :
TargetScan_Total_Verified_Predictions.py
MirSVR
Map_hg19_to_hg38_coordinates.py
Overlapping sites :
MirSVR_TOTAL_Verified_Predictions.py
Without Overlapping sites :
MirSVR_TOTAL_Verified_Predictions.py
Overlapping sites :
MirSVR_Conserved_Verified_Predictions.py
Without Overlapping sites :
MirSVR_Conserved_Verified_Predictions.py
Overlapping sites :
MirSVR_Non_Conserved_Verified_Predictions.py
Without Overlapping sites :
MirSVR_Non_Conserved_Verified_Predictions.py
MBSTAR
Map_hg19_to_hg38_coordinates.py
Overlapping sites :
MBSTAR_TOTAL_Verified_Predictions.py
Without Overlapping sites :
MBSTAR_TOTAL_Verified_Predictions.py
microT-CDS
Map_hg19_to_hg38_coordinates.py
Overlapping sites :
Micro_T_CDS_Total_Verified_Predictions.py
Without Overlapping sites :
Micro_T_CDS_Total_Verified_Predictions.py
Plots_Target_Prediction_Algorithms.py
Final Results – Production of Plots

Chapter VI

Conclusion

miRNA-gene interactions are considered the pillar of most miRNA target prediction studies and significant effort has been employed to deeply understand the underlying mechanisms concerning its function. The development of target prediction programs contributed to the enhancement of in silico miRNA research. The majority of the developed algorithms focused on the detection of miRNA binding sites solely on the 3'UTR of mRNAs. However, current advancement in high-throughput sequencing reveals the presence of abundant target sites in CDS region. Numerous tools have been developed for de novo identification of miRNA-gene interactions. These in silico applications generate diverse outcome results due to the fact that manifold contextual features and experimental data are incorporated in the elaboration of each model. Consequently, the validation of miRNA target prediction tools is multifaceted and it takes into account various parameters. In this study, a variety of such algorithms is examined, considering the shared miRNA-gene and miRNA-site interactions among computational approaches and in vivo experimentally verified predictions.

Ten (10) target prediction algorithms are investigated, namely TargetScan, PACCMIT, PACCMIT-CDS, MIRZA-G (Mirza Analysis), MIRZA-G (Seed Analysis), RNA22, TargetRank, mirSVR, MBSTAR and DIANA microT-CDS. To evaluate their performance, two (2) testing datasets, including experimentally validated positive miRNA targets, were extracted from DIANA - TarBase. Bar plots and scatter plots of common miRNA-gene, miRNA-gene per miRNA and miRNA-site interactions of each algorithm were constructed. The bar plots display the experimentally validated interactions, derived from the analysis, without taking into consideration the score values of algorithms. The scatter plots reveal the relationship between total predictions and correctly verified predictions, either in gene or site interactions, for diverse score values. Furthermore, venn diagrams were devised, depicting the intersection among all programs as far as their shared interactions are concerned.

The major problem of target prediction algorithms constitutes the diversity of their results, which is due to the different positive and negative sets, employed for training and testing the models as well as the various trained features, incorporated in heterogeneous machine learning models. As a result, the aforementioned, followed by the utilization of different databases, assemblies and annotation files from target prediction tools, increases the necessity of achieving a precise assessment of miRNA prediction models under the same base and lead to the development of three (3) test cases. In Test Case I, a set of miRNAs, which contains the common miRNAs between algorithms and the positive set, is devised. In the other two (2) cases, a set of genes, which contains the common genes, either among algorithms or among both algorithms and the positive set, is devised. The goal of these sets is to create a fair set corresponding to all miRNA target prediction programs. Therefore, it is enabled the assessement of the performance of these tools, when shared groups of miRNAs and genes are assumed.

In addition, in the aforementioned cases, between miRNA-gene interactions of the same score, they are selected either the one with the max scoring scheme or the one with an aggregated scoring scheme. Certain algorithms such as TargetScan and TargetRank contain an aggregate score, which is applied to a specific binding site, whereas most algorithms in this analysis such as DIANA-microT-CDS assign this score to the entire miRNA-gene interaction. Consequently, the overall performance of all programs is evaluated in both cases in order to estimate if an alteration in their performance occurs. What is more, in case of miRNA-site interaction analysis, the initial total predicted sites are additionally filtered at the overlapping regions of the positive set.

From the preceding analysis, it is concluded that TargetScan Conserved Sites constitutes the optimal miRNA target prediction program due to its non-existent sensitivity and its accuracy in the predictions. Following TargetScan, mirSVR Conserved MiRNA and DIANA microT-CDS appear to present high performance in terms of miRNA-gene and miRNA-site interactions. In miRNA-gene level, MIRZA-G Seed Conserved and MIRZA-G Conserved present an intermediate sensitivity. MBSTAR and RNA22 are extremely sensitive as they detect a plethora of experimentally supported targets, however, in the basis that they contain a vast number of initial total predictions. Furthermore, conserved algorithms embrace lower sensitivity and thus improved performance compared to non-conserved algorithms. The distinction between optimal and aggregated scores does not bring any changes in the general performance of target prediction programs, with the exception of TargetScan Non Conserved Sites. The performance of the programs remains also intact after the application of filtering at the overlapping regions of the positive set in the case of searching the potential sites in which luciferase or chimeric sites bind. Aggregated scores are independent of binding sites; hence in binding sites analysis, they have not been examined. Overall, summarizing the results from the aforementioned cases, it is found that miRNA target prediction programs that consider conservation, present higher performance such as TargetScan CS, mirSVR CM and MIRZA-G C. Furthermore, without differentiating the conserved and non-conserved sites of programs, TargetScan, mirSVR and DIANA microT-CDS have optimal performance in terms of accuracy and precision in miRNA-gene and miRNA-site levels. In particular, in miRNA-gene level, DIANA microT-CDS performes better than mirSVR, while in miRNA-site level, mirSVR is slightly better than DIANA microT-CDS.

As far as the intersection of shared interactions among programs is concerned, all algorithms shared ~2000 experimantally validated interactions and DIANA microT-CDS embraces separately ~1350 miRNA-gene interactions, when compared to the optimal conserved programs. Moreover, TargetScan reaches ~370 interactions in comparison with high performed target prediction tools, such as mirSVR and DIANA microT-CDS. As a result, it is concluded that TargetScan and DIANA microT-CDS should be examined independently as their intersection with other programs will not enhance their performance. What is more, venn diagrams indicate the diversity of results of miRNA target prediction programs due to the utilization of different machine learning models and trained features.

Even though the currently available aforementioned miRNA target prediction tools identify several correctly predicted targets, it is estimated that even the most sensitive or the most optimal algorithm detects at most 25% of them in both gene and site levels. Consequently, due to the inability of these tools to predict correctly all the experimentally supported targets, it is imperative to devise a novel machine learning model that not only incorporates all information available in experimentally validated repositories (e.g TarBase) but also achieves high performance, presicion and accuracy than previous models, as it would be trained in more rightful data.

Chapter VII

Future Directions

In the future, the analysis of target prediction algorithms can be further expanded and the predictions of Chimiric and MirMark algorithms can be incorporated in the ultimate results. The aforementioned algorithms needed to be executed on servers due to their extremely large set of data, rendering their execution, time consuming and difficult to be incorporated in the present research. Moreover, another direction constitutes the utilization of the data of all programs as training sets in various machine learning models such as SVMs, Decision Trees and Neural Networks (NNs), in order to examine their performance and observe their behavior when many programs intersect with each other. The data of the algorithms, which will feed the classifiers, would have been derived from the process and analysis of target prediction algorithms as far as shared miRNA-gene, miRNA-gene per miRNA as well as miRNA-site interactions are concerned.

It is widely known that the available implementations of algorithms can produce diverse outcomes, as a result of their divergent analysis pipelines. Even though combining different algorithms with set-theoretic operations (e.g. union and intersection) is a common practice, it has been proven not advantageous compared to the use of just one of the best available implementations. Consequently, a future research will involve to what extent two or more programs can be compounded in order to produce results comparable to the ones generated by the most accurate target prediction executions. An in-depth analysis of the comparison of the aforementioned miRNA target prediction programs could be proceded by the utilization of negative sets from TarBase. Since TarBase contains a vast number of negative examples that need to be exploited, valid negative sets would enable the exact computation of sensitivity and specificity metrics of target prediction programs, contributing to an additional assessement of their performance.

Currently, the most significant issue of miRNA target prediction algorithms is the relatively small percentage of experimentally supported targets that they identify. Indeed, the fact that the most sensitive or the most optimal algorithm distinguishs at most 25% of correctly experimental targets in both gene and site levels, indicates that scientific community should mind the gap of miRNA target prediction algorithms. As a result, this could be achieved by the creation of an unprecedented machine learning model, which would dramatically increase the accuracy of the prediction, balancing sensitivity and specificity due to the fact that it would take into consideration the broad unexploited experimental supported interactions, derived from repositories such as TarBase, miRecords and miRTarBase.

References

[1] Ambros, V. (2004). The functions of animal microRNAs. Nature, 431, 350–355. doi: 10.1038/nature02871.

[2] Lee, R. C., Feinbaum, R. L. and Ambros, V. (1993). The C. elegans heterochronic gene lin-4encodes small RNAs with antisense complementarity to lin-14. Cell, 75, 843–854.

[3] Wightman, B., Ha, I. and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell, 75, 855–862.

[4] Pasquinelli, A. E., Reinhart, B. J., Slack, F. et al. (2000). Conservation of the sequence andtemporal expression of let-7 heterochronic regulatory RNA. Nature, 408, 86–89.

[5] Reinhart, B. J., Slack, F. J., Basson, M. et al. (2000). The 21-nucleotide let-7 RNA regulatesdevelopmental timing in Caenorhabditis elegans. Nature, 403, 901–906. doi: 10.1038/35002607.

[6] Slack, F. J., Basson, M., Liu, Z. et al. (2000). The lin-41 RBCC gene acts in the C. elegansheterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. Molecular Cell, 5, 659–669.

[7] Kong Y, Han JH. MicroRNA: biological and computational perspective. Genom.Proteom.Bioinform.2005; 3:62–72. doi: https://doi.org/10.1016/S1672-0229(05)03011-1.

[8] University of Manchester. miRBase. (2006) Available at: http://www.mirbase.org.

[9] Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 116,281–297. doi: https://doi.org/10.1016/S0092-8674(04)00045-5.

[10] Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A. and Tuschl, T. (2003). NewmicroRNAs from mouse and human. RNA, 9, 175–179. doi:10.1261/rna.2146903

[11] Lee, Y., Ahn, C., Han, J. et al. (2003). The nuclear RNase III Drosha initiates microRNAprocessing. Nature, 425, 415–419. doi: 10.1038/nature01957.

[12] Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F. and Hannon, G. J. (2004). Processing of primary microRNAs by the Microprocessor complex. Nature, 432, 231–235. doi: 10.1038/nature03049.

[13] Gregory, R. I., Yan, K. P., Amuthan, G. et al. (2004). The Microprocessor complex mediatesthe genesis of microRNAs. Nature, 432, 235–240. doi: 10.1038/nature03120.

[14] Han, J., Lee, Y., Yeom, K. H. et al. (2004). The Drosha-DGCR8 complex in primary microRNA processing. Genes & Development, 18, 3016–3027.
doi: 10.1101/gad.1262504.

[15] Landthaler, M., Yalcin, A. and Tuschl, T. (2004). The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. Curr.Biol. 14, 2162-2167.doi: 10.1016/j.cub.2004.11.001.

[16] Lund, E., Guttinger, S., Calado, A., Dahlberg, J. E. and Kutay, U. (2004). Nuclear export ofmicroRNA precursors.Science, 303(5654), 95–98.
doi: 10.1126/science.1090599.

[17] Liu J, Carmell MA, Rivas FV, Marsden CG, Thomson JM, Song JJ, Hammond SM, Joshua-Tor L, Hannon GJ 2004. Argonaute2 is the catalytic engine of mammalian RNAi. Science 305: 1437–1441. doi: 10.1126/science.1102513.

[18] Meister G, Landthaler M, Patkaniowska A, Dorsett Y, Teng G, Tuschl T 2004. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. Mol Cell 15: 185–197 doi10.1016/j.molcel.2004.07.007.

[19] Ai-Ming Yu, Ye Tian, Mei-Juan Tu, Pui Yan Ho and Joseph L. Jilek (2016). MicroRNA Pharmacoepigenetics: Posttranscriptional Regulation Mechanisms behind Variable Drug Disposition and Strategy to Develop More Effective Therapy. Drug Metabolism and Disposition March 2016, 44 (3) 308-319; doi:https://doi.org/10.1124/dmd.115.067470.

[20] Lytle JR, Yario TA, Steitz JA (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. 2007;104(23):9667–72.doi:10.1073/pnas.0703820104.

[21] Voinnet O (2009). Origin, biogenesis, and activity of plant microRNAs.Cell. 2009;136(4):669–87.doi: 10.1016/j.cell.2009.01.046.

[22] Hon LS, Zhang Z (2007). The roles of binding site arrangement and combinatorial targeting in microRNA repression of gene expression. Genome Biol. 2007;8(8):R166.ndoi:10.1186/gb-2007-8-8-r166.

[23] Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 2007, 27:91-105. doi: 10.1016/j.molcel.2007.06.017.

[24] Yekta S, Shih IH, Bartel DP (2004). MicroRNA-directed cleavage of HOXB8 mRNA.Science. 2004;304(5670):594–6. doi: 10.1126/science.1097434.

[25] Kim VN (2005). Small RNAs: classification, biogenesis, and function. Mol Cells.2005; 19: 1–15.

[26] Nair VS, Maeda LS, Ioannidis JP (2012). Clinical outcome prediction by micrornas in human cancer: a systematic review. J Natl Cancer Inst 2012;104:528–40. doi: 10.1093/jnci/djs027.

[27] Zen K, Zhang CY (2012). Circulating microRNAs: a novel class of biomarkers to diagnose and monitor human cancers. Med Res Rev 2012;32:326–48. doi: 10.1002/med.20215.

[28] Du L, Pertsemlidis A (2012). MicroRNA regulation of cell viability and drug sensitivity in lung cancer. Expert Opin Biol Ther 2012;12:1221–39. doi: 10.1517/14712598.2012.697149.

[29] Wang H, Tan G, Dong L, Cheng L, Li K, Wang Z, et al (2012). Circulating mir-125b as a marker predicting chemoresistance in breast cancer. PLoS One 2012;7:e34210. doi: 10.1371/journal.pone.0034210.

[30] Ajay Francis Christopher, Raman Preet Kaur, Gunpreet Kaur, Amandeep Kaur, Vikas Gupta, Parveen Bansal (2016). MicroRNA therapeutics: Discovering novel targets and developing specific therapy. Perspect Clin Res. 2016 Apr-Jun; 7(2): 68–74. doi: 10.4103/2229-3485.179431.

[31] Fazli Wahid, Adeeb Shehzad, TaousKhan, You Young Kim (2010). MicroRNAs: Synthesis, mechanism, function, and recent clinical trials. BBA - Molecular Cell Research. Volume 1803, Issue 11, November 2010, Pages 1231-1243. doi: 10.1016/j.bbamcr.2010.06.013.

[32] Chiranjib Chakraborty, Ashish Ranjan Sharma, Garima Sharma, C. George Priya Doss, and Sang-Soo Lee. Therapeutic miRNA and siRNA: Moving from Bench to Clinic

as Next Generation Medicine. Mol Ther Nucleic Acids.2017 Sep 15; 8: 132–143.doi:10.1016/j.omtn.2017.06.005.

[32a] Eva van Rooij, Angela L. Purcell, Arthur A. Levin (2012). Developing MicroRNA Therapeutics. Circ Res. 2012;110:496-507. doi: 10.1161/CIRCRESAHA.111.247916.

[33] Betel D., Koppa A., Agius P., Sander C., Leslie C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biology 11:R90. doi: 10.1186/gb-2010-11-8-r90.

[34]Kevin C. Miranda, Tien Huynh, Yvonne Tay,Yen-Sin Ang, Wai-Leong Tam, Andrew M. Thomson, Bing Lim and Isidore Rigoutsos (2006). A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes. Cell 126, 1203-1217. doi: 10.1016/j.cell.2006.07.031.

[35] Witkos, T. M., Koscianska, E., and Krzyzosiak, W. J. (2011). Practical aspects of microRNA target prediction. Curr. Mol. Med. 11, 93–109. doi: 10.2174/156652411794859250.

[36] Garcia, D. M., Baek, D., Shin, C., Bell, G. W., Grimson, A., and Bartel, D. P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat. Struct. Mol. Biol. 18, 1139–1146. doi: 10.1038/nsmb.2115.

[37] 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. doi: 10.1038/ng2135.

[38] Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 2007, 27:91-105. doi: 10.1016/j.molcel.2007.06.017.

[39] Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., and Burge, C. B. (2008). Proliferating cells express mRNAs with shortened 3′UTRs and fewer microRNA target sites. Science 320, 1643–1647. doi: 10.1126/science.1155390.

[40] Lewis BP1, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB (2003).Prediction of mammalian microRNA targets.Cell. 2003 Dec 26;115(7): 787-798.doi: https://doi.org/10.1016/S0092-8674(03)01018-3.

[41] Peterson SM, Thompson JA, Ufkin ML, Sathyanarayana P, Liaw L, Congdon CB (2014). Common features of microRNA target prediction tools.Front Genet. 2014 Feb 18;5:23. doi: 10.3389/fgene.2014.00023.

[42] Website of TargetScan :http://www.targetscan.org/vert_72/.

[43] Brennecke J., Stark A., Russell R. B. & Cohen S. M (2005). Principles of microRNA-target recognition.PLoS Biol. 3, e85 (2005).doi: 10.1371/journal.pbio.0030085.

[43a] Heeyoung Seok, Juyoung Ham, Eun-Sook Jang, Sung Wook Chi (2016). MicroRNA Target Recognition: Insights from Transcriptome-Wide Non-Canonical Interactions. Molecules and Cells 2016; 39(5): 375-381. doi: https://doi.org/10.14348/molcells.2016.0013.

[44] Friedman R. C., Farh K. K., Burge C. B., Bartel D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 19, 92–105. doi: 10.1101/gr.082701.108.

[45] Fujiwara T., Yada T. (2013). miRNA-target prediction based on transcriptional regulation. BMC Genomics 2:S3. doi: 10.1186/1471-2164-14-S2-S3.

[46] Ohler U., Yekta S., Lim L. P., Bartel D. P., Burge C. B. (2004). Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. RNA 10, 1309–1322. doi: 10.1261/rna.5206304.

[47] Yue D, Liu H, Huang Y (2009). Survey of Computational Algorithms for MicroRNA Target Prediction.Curr Genomics. 2009 Nov; 10(7):478-92. doi: 10.2174/138920209789208219.

[48] Mahen E. M., Watson P. Y., Cottrell J. W., Fedor M. J. (2010). mRNA secondary structures fold sequentially but exchange rapidly in vivo. PLoS Biol. 8:e1000307. doi: 10.1371/journal.pbio.1000307.

[49] Long D., Lee R., Williams P., Chan C. Y., Ambros V., Ding Y. (2007). Potent effect of target structure on microRNA function. Nat. Struct. Mol. Biol. 14, 287–294. doi: 10.1038/nsmb1226.

[50] Riffo-Campos ÁL, Riquelme I, Brebi-Mieville P (2016). Tools for Sequence-Based miRNA Target Prediction: What to Choose? UI-TEI K, Pichler M, eds. International Journal of Molecular Sciences. 2016;17(12):1987. doi: 10.3390/ijms17121987.

[51] Oliveira AC, Bovolenta LA, Nachtigall PG, Herkenhoff ME, Lemke N, Pinhal D. Combining Results from Distinct MicroRNA Target Prediction Tools Enhances the Performance of Analyses. Frontiers in Genetics. 2017;8:59. doi: 10.3389/fgene.2017.00059.

[52] L.P. Lim, N.C. Lau, P. Garrett-Engele, A. Grimson, J.M. Schelter, J. Castle, et al. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature, 433 (2005), pp. 769-773. doi: 10.1038/nature03315.

[53] R. Ji, Y. Cheng, J. Yue, J. Yang, X. Liu, H. Chen, et al. (2007). RNA expression signature and antisense-mediated depletion reveal an essential role of MicroRNA in vascular neointimal lesion formation. Circ Res, 100 (2007), pp. 1579-1588. doi:10.1161/CIRCRESAHA.106.141986.

[54] Thomson D.W., Bracken C.P., Goodall G.J (2011) . Experimental strategies for microRNA target identification. Nucleic Acids Res. 2011;39:6845–6853. doi: 10.1093/nar/gkr330.

[55] Chi S.W., Zang J.B., Mele A., Darnell R.B (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature. 2009;460:479–486. doi: 10.1038/nature08170.

[56] Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N(2008). Widespread changes in protein synthesis induced by microRNAs. Nature 2008, 455:58-63. doi: 10.1038/nature07228.

[57] German M.A., Luo S., Schroth G., Meyers B.C., Green P.J (2009). Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. Nat Protoc. 2009;4:356–362. doi: 10.1038/nprot.2009.8.

[58] D. Baek, J. Villén, C. Shin, F.D. Camargo, S.P. Gygi, D.P. Bartel (2008). The impact of microRNAs on protein output. Nature, 455 (2008), pp. 64-71.doi: 10.1038/nature07242.

[59] M. Selbach, B. Schwanhausser, N. Thierfelder, Z. Fang, R. Khanin, N (2008). RajewskyWidespread changes in protein synthesis induced by microRNAs. Nature, 455 (2008), pp. 58-63.doi: 10.1038/nature07228.

[60] Karagkouni, D., M. D. Paraskevopoulou, S. Chatzopoulos, I. S. Vlachos, S. Tastsoglou, I. Kanellos, D. Papadimitriou, et al. 2018). DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. Nucleic Acids Research 46 (Database issue): D239-D245. doi: 10.1093/nar/gkx1141.

[61] Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan (2011).A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat Methods. 2011 May 15; 8(7):559-64. doi: 10.1038/nmeth.1608.

[62] Moore MJ, Scheel TK, Luna JM, Park CY, Fak JJ, Nishiuchi E, Rice CM, Darnell RB (2015). miRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. Nat Commun. 2015 Nov 25; 6():8864. doi: 10.1038/ncomms9864.

[63] Helwak A, Kudla G, Dudnakova T, Tollervey D (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell. 2013 Apr 25; 153(3):654-65. doi: 10.1016/j.cell.2013.03.043.

[64] Vlachos I.S., Georgakilas G., Tastsoglou S., Paraskevopoulou M.D., Karagkouni D., Hatzigeorgiou A.G. De Pietri Tonelli D., editor (2017). Computational challenges and -omics approaches for the identification of miRNAs and targets. Essentials of Noncoding RNA in Neuroscience. 2017; Boston: Academic Press; 39–60.

[65] Vlachos I.S., Paraskevopoulou M.D., Karagkouni D., Georgakilas G., Vergoulis T., Kanellos I., Anastasopoulos I.L., Maniou S., Karathanou K., Kalfakakou D. et al. (2015). DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Res. 2015; 43:D153–D159. doi: 10.1093/nar/gku1215.

[66] Antonia Vlahou, Fulvio Magni, Harald Mischak, Jerome Zoidakis (2018). Integration of Omics Approaches and Systems Biology for Clinical Applications. 2018 John Wiley & Sons, Inc. Published : 27 January 2018. 2018. pages: 67-92. doi: 10.1002/9781119183952.

[67] Tan S.M., Kirchner R., Jin J., Hofmann O., McReynolds L., Hide W., Lieberman J. (2014). Sequencing of Captive Target Transcripts Identifies the Network of Regulated Genes and Functions of Primate-Specific miR-522. Cell Rep. 2014;8:1225–1239. doi: 10.1016/j.celrep.2014.07.023.

[68] Wolter J.M., Kotagama K., Pierre-Bez A.C., Firago M., Mangone M. (2014). 3′LIFE: a functional assay to detect miRNA targets in high-throughput. Nucleic Acids Res. 42:2015, e132.doi: 10.1093/nar/gku626.

[69] Barrett T., Wilhite S.E., Ledoux P., Evangelista C., Kim I.F., Tomashevsky M., Marshall K.A., Phillippy K.H., Sherman P.M., Holko M.(2012). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2012; 41:D991–D995. doi: 10.1093/nar/gks1193.

[70] Kodama Y., Shumway M., Leinonen R. (2011). The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res. 2011; 40:D54–D56. doi: 10.1093/nar/gkr854.

[71] Chih-Hung Chou, Sirjana Shrestha, Chi-Dung Yang, Nai-Wen Chang, Yu-Ling Lin, Kuang-Wen Liao, Wei-Chi Huang, Ting-Hsuan Sun, Siang-Jyun Tu, Wei-Hsiang Lee, Men-Yee Chiew, Chun-San Tai, Ting-Yen Wei, Tzi-Ren Tsai, Hsin-Tzu Huang, Chung-Yu Wang, Hsin-Yi Wu, Shu-Yi Ho, Pin-Rong Chen, Cheng-Hsun Chuang, Pei-Jung Hsieh, Yi-Shin Wu, Wen-Liang Chen, Meng-Ju Li, Yu-Chun Wu, Xin-Yi Huang, Fung Ling Ng, Waradee Buddhakosai, Pei-Chun Huang, Kuan-Chun Lan, Chia-Yen Huang, Shun-Long Weng, Yeong-Nan Cheng, Chao Liang, Wen-Lian Hsu, Hsien-Da Huang (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-

target interactions, Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D296–D302, doi: 10.1093/nar/gkx1067.

[72] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna M. Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M. Stuar, , Cancer Genome Atlas Research Network (2013). Cancer Genome Atlas Pan-Cancer Analysis Project. Nat Genet. 2013 Oct; 45(10): 1113–1120. doi: 10.1038/ng.276.

[73] Feifei Xiao, Zhixiang Zuo, Guoshuai Cai, Shuli Kang, Xiaolian Gao, Tongbin Li (2009). miRecords: an integrated resource for microRNA–target interactions. Nucleic Acids Res. 2009 Jan; 37(Database issue): D105–D110. doi: 10.1093/nar/gkn851.

[74] Jun-Hao Li, Shun Liu, Hui Zhou, Liang-Hu Qu, Jian-Hua Yang (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Research, Volume 42, Issue D1, 1 January 2014, Pages D92–D97. doi: 10.1093/nar/gkt1248.

[75] Vikram Agarwal, George W Bell, Jin-Wu Nam, David P Bartel (2015).Predicting effective microRNA target sites in mammalian mRNAs.Elife.doi: 10.7554/eLife.05005.

[76] Bazzini AA, Lee MT, Giraldez AJ. (2012). Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science. 2012;336:233–237. doi: 10.1126/science.1215704.

[77] Loeb GB, Khan AA, Canner D, Hiatt JB, Shendure J, Darnell RB, Leslie CS, Rudensky AY (2012). Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting.Mol Cell. 2012 Dec 14; 48(5):760-70. doi: 10.1016/j.molcel.2012.10.002.

[78] Rodriguez A, Vigorito E, Clare S, Warren MV, Couttet P, Soond DR, van Dongen S, Grocock RJ, Das PP, Miska EA, Vetrie D, Okkenhaug K, Enright AJ, Dougan G, Turner M, Bradley A (2007). Requirement of bic/microRNA-155 for normal immune function.Science. 2007 Apr 27; 316(5824):608-11. doi: 10.1126/science.1139253.

[79] Hafner M., et al (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.Cell 2010, vol. 141 (pg.129-141).doi: 10.1016/j.cell.2010.03.009.

[80] Linsley PS, Schelter J, Burchard J, Kibukawa M, Martin MM, Bartz SR, Johnson JM, Cummins JM, Raymond CK, Dai H, Chau N, Cleary M, Jackson AL, Carleton M, Lim L (2007).Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. Mol Cell Biol 2007, 27:2240-52. doi: 10.1128/MCB.02005-06.

[81] Tan SM, Kirchner R, Jin J, Hofmann O, McReynolds L, Hide W, Lieberman J (2014). Sequencing of captive target transcripts identifies the network of regulated genes and functions of primate-specific miR-522.Cell Rep. 2014 Aug 21; 8(4):1225-39.doi: 10.1016/j.celrep.2014.07.023.

[82] Eichhorn SW, Guo H, McGeary SE, Rodriguez-Mias RA, Shin C, Baek D, Hsu SH, Ghoshal K, Villén J, Bartel DP (2014). mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. Mol Cell. 2014 Oct 2; 56(1):104-15. doi: 10.1016/j.molcel.2014.08.028.

[83] Birmingham A, Anderson EM, Reynolds A, Ilsley-Tyree D, Leake D, Fedorov Y, Baskerville S, Maksimova E, Robinson K, Karpilow J, Marshall WS, Khvorova A (2006). 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. Nature Methods. 2006;3:199–204. doi: 10.1038/nmeth854.

[84] Schwarz DS, Ding HL, Kennington L, Moore JT, Schelter J, Burchard J, Linsley PS, Aronin N, Xu ZS, Zamore PD (2006). Designing siRNA that distinguish between genes that differ by a single nucleotide.PLOS Genetics. 2006;2:1307–1318. doi: 10.1371/journal.pgen.0020140.

[85] Jackson AL, Burchard J, Leake D, Reynolds A, Schelter J, Guo J, Johnson JM, Lim L, Karpilow J, Nichols K, Marshall W, Khvorova A, Linsley PS (2006). Position-specific chemical modification of siRNAs reduces "off-target" transcript silencing. RNA. 2006 Jul; 12(7):1197-205. doi: 10.1261/rna.30706.

[86] Jackson AL, Burchard J, Schelter J, Chau BN, Cleary M, Lim L, Linsley PS (2006). Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity.RNA. 2006 Jul; 12(7):1179-87. doi: 10.1261/rna.25706.

[87] Anderson EM, Birmingham A, Baskerville S, Reynolds A, Maksimova E, Leake D, Fedorov Y, Karpilow J, Khvorova A (2008). Experimental validation of the importance of seed complement frequency to siRNA specificity. RNA. 2008 May; 14(5):853-61. doi: 10.1261/rna.704708.

[88] Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, Enright AJ, Schier AF (2006). Zebrafish miR-430 promotes deadenylation and clearance of maternal mRNAs. Science. 2006;312:75–79. doi: 10.1126/science.1122689.

[89] Nam JW, Rissland OS, Koppstein D, Abreu-Goodger C, Jan CH, Agarwal V, Yildirim MA, Rodriguez A, Bartel DP (2014). Global analyses of the effect of different cellular contexts on microRNA targeting. Mol Cell. 2014 Mar 20; 53(6):1031-1043. doi: https://doi.org/10.1016/j.molcel.2014.02.013.

[90] Lipchina I, Elkabetz Y, Hafner M, Sheridan R, Mihailovic A, Tuschl T, Sander C, Studer L, Betel D (2011). Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response. Genes

Dev. 2011 Oct 15; 25(20):2173-86. doi: 10.1101/gad.17221311.

[91] Mayr C, Bartel DP (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell. 2009 Aug 21; 138(4):673-84. doi: 10.1016/j.cell.2009.06.016.

[92] Hausser J, Landthaler M, Jaskiewicz L, Gaidatzis D, Zavolan M (2009). Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. Genome Res. 2009 Nov; 19(11):2009-20. doi: 10.1101/gr.091181.109.

[93] Miroslav Šulc Ray M. Marín Harlan S. Robins Jiří Vaníček. (2015). PACCMIT/PACCMIT-CDS: identifying microRNA targets in 3′ UTRs and coding sequences. Nucleic Acids Res. 2015 Jul 1;43(W1):W474-9. doi: https://doi.org/10.1093/nar/gkv457.

[94] Marín RM, Vaníček J (2012). Optimal Use of Conservation and Accessibility Filters in MicroRNA Target Prediction. PLoS ONE 7(2): e32208.doi: 10.1371/journal.pone.0032208.

[95] M. Šulc, R. M. Marín, H. S. Robins, and J. Vaníček (2013). Searching the coding region for microRNA targets.RNA. 2013 Apr; 19(4): 467–474. doi: 10.1261/rna.035634.112.

[96] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34: D140–D144. doi: 10.1093/nar/gkj112.

[97] Knuth DE (1997). The art of computer programming, volume 2: Seminumerical algorithms. Addison-Wesley Professional.

[98]Rafal Gumienny and Mihaela Zavolan (2015). Accurate transcriptome-wide prediction of microRNAtargets and small interfering RNA off-targets with MIRZA-G. Nucleic Acids Res. 2015 Feb 18;43(3):1380-91. doi: 10.1093/nar/gkv050.

[99] Khorshid M., Hausser J., Zavolan M., Nimwegen E. (2013). A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. Nature Methods. 2013, vol. 10 (pg. 253-255). doi: https://www.nature.com/articles/nmeth.2341.

[100] Van Dongen S., Abreu-Goodger C., Enright A.J.(2008). Detecting microRNA binding and siRNA off-target effects from expression data.Nature Methods. 2008. vol. 5 (pg. 1023-1025). doi: 10.1038/nmeth.1267.

[101] Dahiya N., Sherman-Baust C.A., Wang T.-L., Davidson B., Shih I.-M., Zhang Y., Wood W.3rd, Becker K.G., Morin P.J. (2008). MicroRNA expression and identification of putative miRNA targets in ovarian cancer.PLoS One.2008., vol. 3 pg. e2436.

doi:http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0002436.

[102] Frankel L.B., Wen J., Lees M., Høyer-Hansen M., Farkas T., Krogh A., Jäättelä M., Lund A.H. (2011).microRNA-101 is a potent inhibitor of autophagy. EMBO J. 2011, vol. 30 (pg. 4628-4641).doi: 10.1038/emboj.2011.331.

[103] Gennarino V.A., et al.(2009). MicroRNA target prediction by expression analysis of host genes.Genome Res. 2009, vol. 19 (pg. 481 -490).doi: 10.1101/gr.084129.108.

[104] Hudson R.S., Yi M., Esposito D., Glynn S.A., Starks A.M., Yang Y., Schetter A.J., Watkins S.K., Hurwitz A.A., Dorsey T.H., et al. (2013). MicroRNA-106b-25 cluster expression is associated with early disease recurrence and targets caspase-7 and focal adhesion in human prostate cancer. Oncogene.2013, vol. 32 (pg. 4139-4147).doi: 10.1038/onc.2012.424.

[105] Leivonen S.-K., Mäkelä R., Ostling P., Kohonen P., Haapa-Paananen S., Kleivi K., Enerly E., Aakula A., Hellström K., Sahlberg N., et al. (2009). Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines.Oncogene.2009, vol. 28 (pg. 3926-3936). doi: 10.1038/onc.2009.241.

[106] Gaidatzis D., van Nimwegen E., Hausser J., Zavolan M. (2007). Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC Bioinformatics. 2007, vol. 8 (pg. 69-90). doi: https://doi.org/10.1186/1471-2105-8-69.

[107] R.F. Doolittle, M.W. Hunkapiller, L.E. Hood, S.G. Devare, K.C. Robbins, S.A. Aaronson, H.N. Antoniades (1983). Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. Science, 221 (1983), pp. 275-277. doi: 10.1126/science.6304883.

[108] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R (2003). Rfam: an RNA family database.Nucleic Acids Res. 2003; 31: 439–441.

[109] Rigoutsos, I., Huynh, T., Miranda, K., Tsirigos, A., McHardy, A., and Platt, D.Proc (2006). Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. Natl. Acad. Sci. USA. 2006; 103: 6605–6610.doi: 10.1073/pnas.0601688103.

[110] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J (1990).Basic local alignment search tool.Altschul, Mol. Biol. 1990; 215: 403–410.doi: 10.1016/S0022-2836(05)80360-2.

[111] Rigoutsos, I. and Floratos (1998). Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. A.Bioinformatics. 1998; 14: 55–

67.

[112] Brazma, A., Jonassen, I., Eidhammer, I., and Gilbert, D.J (1998).Approaches to the automatic discovery of patterns in biosequences.Comput.Biol. 1998; 5: 279–305.doi:10.1089/cmb.1998.5.279.

[113] Kawai, K., Yokota, C., Ohashi, S., Watanabe, Y., and Yamashita, K. (1995).Evidence that glucagon stimulates insulin secretion through its own receptor in rats. Diabetologia. 1995; 38: 274–276.

[114] MacDonald, P.E., Ha, X.F., Wang, J., Smukler, S.R., Sun, A.M., Gaisano, H.Y., Salapatek, A.M., Backx, P.H., and Wheeler, M.B. Members of the Kv1 and Kv2 voltage-dependent K(+) channel families regulate insulin secretion.Mol. Endocrinol. 2001; 15: 1423–1435.doi: 10.1210/mend.15.8.0685.

[115] Cydney B. Nielsen, Noam Shomron, Rickard Sandberg, Eran Hornstein, Jacob Kitzman and Christopher B. Burge (2007).Determinants of targeting by endogenous and exogenous microRNAs and siRNAs.RNA 2007. 13: 1894-1910. doi: 10.1261/rna.768207.

[116] Jackson, A.L., Bartz, S.R., Schelter, J., Kobayashi, S.V., Burchard, J., Mao, M., Li, B., Cavet, G., and Linsley, P.S. (2003). Expression profiling reveals off-target gene regulation by RNAi. Nat. Biotechnol. 21: 635–637. doi:10.1038/nbt831.

[117] Jackson, A.L., Burchard, J., Leake, D., Reynolds, A., Schelter, J., Guo, J., Johnson, J.M., Lim, L., Karpilow, J., Nichols, K., et al. (2006). Position-specific chemical modification of siRNAs reduces ‘‘off-target’’ transcript silencing. RNA 12: 1197–1205.doi: 10.1261/rna.30706.

[118] Hayashi, S. and McMahon, A.P. (2002). Efficient recombination in diverse tissues by a tamoxifen-inducible form of Cre: A tool for temporally regulated gene activation/inactivation in the mouse. Dev. Biol. 244: 305–318.doi:10.1006/dbio.2002.0597.

[119] Harfe, B.D., McManus, M.T., Mansfield, J.H., Hornstein, E., and Tabin, C.J. (2005). The RNaseIII enzyme Dicer is required for morphogenesis but not patterning of the vertebrate limb. Proc. Natl. Acad. Sci. 102: 10898–10903.doi: 10.1073/pnas.0504834102.

[120] Soriano, P. (1999).Generalized lacZ expression with the ROSA26 Cre reporter strain. Nat. Genet. 21: 70–71. doi: 10.1038/5007.

[121] Abbondanzo, S.J., Gadi, I., and Stewart, C.L. (1993).Derivation of embryonic stem cell lines.Methods Enzymol. 225: 803–823.

[122] Tafer H, Ameres SL, Obernosterer G, Gebeshuber CA, Schroeder R, Martinez

J, Hofacker IL (2008). The impact of target site accessibility on the design of effective siRNAs. Nat Biotechnol 2008, 26:578-83. doi: 10.1038/nbt1404.

[123] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005). Evolutionarily conservedelements in vertebrate, insect, worm, and yeast genomes (2005). Genome Res 2005, 15:1034-50. doi: 10.1101/gr.3715005.

[124] Gabriely G, Wurdinger T, Kesari S, Esau CC, Burchard J, Linsley PS, Krichevsky AM (2008). MicroRNA 21 promotes glioma invasion by targeting matrix metalloproteinase regulators. Mol Cell Biol 2008, 28:5369-80. doi: 10.1128/MCB.00479-08.

[125] Elmén J, Lindow M, Silahtaroglu A, Bak M, Christensen M, Lind-Thomsen A, Hedtjärn M, Hansen JB, Hansen HF, Straarup EM, McCullagh K, Kearney P, Kauppinen S (2008). Antagonism of microRNA-122 in mice by systemically administered LNA-antimiR leads to up-regulation of a large set of predicted target mRNAs in the liver. Nucleic Acids Res 2008, 36:1153-62. doi: 10.1093/nar/gkm1113.

[126] Landthaler M, Gaidatzis D, Rothballer A, Chen PY, Soll SJ, Dinic L, Ojo T, Hafner M, Zavolan M, Tuschl T (2008). Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. RNA 2008, 14:2580-96. doi: 10.1261/rna.1351608.

[127] RNA regulatory networks, Zavolan Lab. [http://www.http://www.mirz.unibas.ch/]

[128] Bandyopadhyay S., Ghosh D., Mitra R. & Zhao Z. (2015). MBSTAR: multiple instance learning for predicting specific functional binding sites in microRNA targets. Sci. Rep. 5, 8004.doi:10.1038/srep08004.

[129] Bandyopadhyay, S. & Mitra, R. (2009). TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. Bioinformatics. 25, 2625–2631 .doi: 10.1093/bioinformatics/btp503.

[130] University of California Santa Cruz. UCSC Genome Browser.Available at: http://genome.ucsc.edu/.

[131] Vergoulis, T. et al. (2012). TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic Acids Res.40, D222–D229 (2012).doi: 10.1093/nar/gkr1161.

[132] Jianhua, Y. starBase. (2010). Available at:http://starbase.sysu.edu.cn/download.php.

[133] Liu H., Yue D., Chen Y., Gao S. J., Huang Y. (2010). Improving performance of mammalian microRNA target prediction.BMC Bioinformatics.doi: 10.1186/1471-2105-11-476.

[134] Mark Menor, Travers Ching, Xun Zhu, David Garmire and Lana X. Garmire (2014).mirMark: a site-level and UTR-level classifier for miRNA target prediction. Genome Biol. 2014; 15(10): 500. doi: 10.1186/s13059-014-0500-5.

[135] John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS (2004). Human microRNA targets. PLoS Biol. 2004;2:e363. doi: 10.1371/journal.pbio.0020363.

[136] Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T (2009). miRecords: an integrated resource for microRNA–target interactions. Nucleic Acids Res. 2009;37:D105–D110. doi: 10.1093/nar/gkn851.

[137] Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH, Chu CF, Huang HY, Lin CM, Ho SY, Jian TY, Lin FM, Chang TH, Weng SL, Liao KW, Liao IE, Liu CC, Huang HD (2014). miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res. 2014;42:D78–D85. doi: 10.1093/nar/gkt1266.

[138] Lorenz R, Bernhart SH, Zu Siederdissen CH, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011). ViennaRNA Package 20. Algorithm Mol Biol. 2011;6:26. doi: 10.1186/1748-7188-6-26.

[139] Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H. (2002). The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 2002;12:1611–1618. doi: 10.1101/gr.361602.

[140] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009). The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter. 2009;11:10–18. doi: 10.1145/1656274.1656278.

[141] Hausser J, Strimmer K (2009).Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks.JMLR. 2009;10:1469–1484. doi: https://arxiv.org/abs/0811.3579.

[142] Lu Y, Leslie CS (2016). Learning to PredictmiRNA-mRNA Interactions from AGO CLIP Sequencing and CLASH Data. PLoS Comput Biol 12 (7): e1005026. doi:10.1371/journal.pcbi.1005026.

[143] Chi SW, Hannon GJ, Darnell RB (2012).An alternative mode of microRNA target recognition. Nat Struct Mol Biol. 2012 Feb 12; 19(3):321-7. doi: 10.1038/nsmb.2230.

[144] Grosswendt S, Filipchyk A, Manzano M, Klironomos F, Schilling M, Herzog

M, Gottwein E, Rajewsky N (2014.) Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. Mol Cell. 2014 Jun 19; 54(6):1042-1054. doi: 10.1016/j.molcel.2014.03.049.

[145] Sonnenburg S, Zien A, Philips P, Rätsch G (2008). POIMs: positional oligomer importance matrices--understanding support vector machine-based signal detectors. Bioinformatics. 2008 Jul 1; 24(13):i6-14. doi: 10.1093/bioinformatics/btn170.

[146] Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG.(2012). Functional microRNA targets in protein coding sequences.Bioinformatics.2012 Jan 27.doi: 10.1093/bioinformatics/bts043.

[147] Wang X., Wang X.(2006). Systematic identification of microRNA functions by combining target prediction and expression profiling. Nucleic Acids Res. 2006, vol. 34 (pg. 1646-1652). doi: 10.1093/nar/gkl068.

[148] Bronwen L. Aken, Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Friederike Bernsdorff, Jyothish Bhai, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Thomas Juettemann, Stephen Keenan, Matthew R. Laird, Ilias Lavidas, Thomas Maurel, William McLaren, Benjamin Moore, Daniel N. Murphy, Rishi Nag, Victoria Newman, Michael Nuhn, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Daniel Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Steven P. Wilder, Amonida Zadissa, Myrto Kostadima, Fergal J. Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Daniel M. Staines, Stephen J. Trevanion, Fiona Cunningham, Andrew Yates, Daniel R. Zerbino, Paul Flicek; Ensembl 2017, Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D635–D642, doi: https://doi.org/10.1093/nar/gkw1104.

[148a] Schultz N., Marenstein D.R., De Angelis D.A., Wang W.-Q., Nelander S., Jacobsen A., Marks D.S., Massagué J., Sander C. (2011). Off-target effects dominate a large-scale RNAi screen for modulators of the TGF-β pathway and reveal microRNA regulation of TGFBR2. Silence. 2011, vol. 2 pg. 3. doi: 10.1186/1758-907X-2-3.

[149] RNA splicing: *https://en.wikipedia.org/wiki/RNA_splicing#/.*